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 and Law, University of Calabria, Italy*

Abstract

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.

Keywords: Social media, Sentiment Analysis, Rating, Knowledge Management

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

*Corresponding author. Tel.: +39 0984 49 2450
Email address: michelangelo.misuraca@unical.it (Michelangelo Misuraca)

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.

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

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

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, Karpowicz, Fürschuß, Auinger, Stříteský & Holzinger, 2013).

Analysing opinions written in natural language is a very interesting research domain, known as opinion mining (OM) or sentiment analysis (SA) (Petz, Karpowicz, Fürschuß, Auinger, Stříteský & Holzinger, 2014). According to Pang & 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 & Ashraf (2014). These authors share the idea of Tang, Tan & 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; Ravi & Ravi, 2015; Qazi, Raj, Hardaker & Standing, 2017).

A large number of papers mention SA in the context of the so-called *polarity classification* (e.g. Taboada, Brooke, Tofiloski, Voll & Stede, 2011; Cambria, Schuller, Xia & Havasi, 2013). 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 & Hui (2011), for example, proposed a supervised term weighting scheme based on the Kullback-Leibler divergence. Lin, Zhang, Wang & Zhou (2012) and Khan, Qamar & Bashir (2016) proposed the use of mutual information. Gann, Day & Zhou (2014) introduced a *total sentiment index* to score the polarity of the different terms.

As suggested by Pang & Lee (2005), it is helpful to have more than the binary distinction between positive and negative opinions. This classification has less information with respect to the differences highlighted by the polarity degree, because the polarity of an opinion can be measured on a continuous scale. This task is known as *rating inference* (Leung, Chan & Chung, 2011; Serrano-Guerrero, Olivas, Romero & Herrera-Viedma, 2015; Cosma & Acampora, 2016; Xue, Li & Rishe, 2017). Given positive and negative opinions, rating inference seeks to determine the overall sentiment implied by the user in the review, and map said sentiment onto a rating scale. As an example, a machine learning approach to predict the sentiment-polarity scores of reviews was developed by Okanohara & Tsujii (2005). The authors proposed a new sentiment polarity score based on a 1-to-5 star scale.

In common practice, there is a lack of algorithms able to combine judgments provided with both comments and scores. We propose a SA-based approach that seeks to quantify the textual content of each review in a numerical value, and then combine this value with the related score assigned by the user. In this way the poor informative power of the common rating scales is enriched.

3. The proposed method

In review and ratings social media, the rating scale used to assign a score to the reviewed product or service can be assumed as a global and comparable measure of the user experience. The reviews are textual descriptions highlighting which aspects of the product or service are personally considered positive or negative. Given the different evaluations expressed by the two systems, rating scores and textual reviews, we propose a strategy to calculate a *polarity-driven rating*. The new rating scale combines the rating assigned by the reviewer and the polarity score of the review in a unique measure.

Table 1: Meanings of the notations used in the following

Symbol	Definition	Symbol	Definition
H	number of rating categories	$r_{w_{ijk}}$	polarity score of a term k
c_h	a generic rating category	$r_{s_{ij}}$	polarity score of a sentence j
n	number of reviews	r_{d_i}	polarity score of a review i
d_i	a generic review i		
q_i	number of sentences in d_i		$h = 1,\ldots, H$
s_{ij}	a generic sentence in d_i		$i = 1,\ldots, n$
p_j	number of terms in s_{ij}		$j = 1,\ldots, q_i$
w_{ijk}	a generic term in s_{ij}		$k = 1,\ldots, p_j$

3.1. Text pre-processing

Let us consider a set of n reviews categorised with a 1-to-H rating scale, where c_h is a generic rating category. Each review d_i (with $i = 1,\ldots,n$) can be seen as a document written in natural language. It is possible to apply on the *corpus* of reviews the pre-treatment procedures usually carried out in a text mining framework. Because of the particular nature of the sentiment-based approach used in the following, we adopt a soft pre-treatment process. Only a normalisation of punctuation, blanks, tabulations and “not printable” characters is performed. Moreover, the stop-words are preserved in order to save the syntactical structures for the polarity calculation. After the pre-treatment, each review is segmented into the set of its q_i sentences $\{s_{i1},\ldots,s_{ij},\ldots,s_{iq_i}\}$, by considering only strong punctuation like full stops, question marks and exclamation marks as separators. We decided to not distinguish the sentences in terms of subjectivity/objectivity (Wilson, Wiebe & Hoffmann, 2005). Subjectivity/objectivity detection decides if a text expresses an opinion on its subjective matter or it has a factual nature. In the following all the sentences in a review are considered at the same time for the polarity calculation, due to the shortness of the text.

3.2. Computing the review polarities

After pre-processing the reviews, a sentiment-based approach is used to calculate the polarity. The polarity of the reviews is first calculated at sentence-level, then summarised at document-level. This approach seems to be more effective, because in the reviews, each sentence can express an opinion about a different aspect of the reviewed product or service. Each sentence j is represented as a sequence of its p_j terms $\{w_{ij1}, \ldots, w_{ijk}, \ldots, w_{ijp_j}\}$, preserving the order of the terms into the sentence. Each term w_{ijk} in the sentence s_{ij} of the review i is compared with a term-sentiment association lexicon, assigning a $r_{w_{ijk}}$ score of -1 for negative terms, and a score of $+1$ for positive terms, respectively. The terms not included into the lexicon are assumed to be neutral, with a score equal to 0. The polarity of each term is then properly weighted by taking into account negators (e.g. "never", "none"), amplifiers and de-amplifiers (e.g. "very", "few"), adversative and contrasting conjunctions (e.g. "but", "however"). This weighting scheme $-$ based on the effect of *shifters* onto polarised terms $-$ allows the positivity and negativity of each term to be emphasised or dampened, and leads to a more effective measure of the sentence polarity (Polanyi & Zaenen, 2004). The logic is to capture the polarity by considering the context of use of the different terms (see Saif, He, Fernandez & Alani, 2016; Xia, Xu, Yu, Qi & Cambria, 2016; Vechtomova, 2017).

The $r_{s_{ij}}$ total polarity score of each sentence is computed as the sum of its weighted term scores $r^*_{w_{ijk}}$, 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 semantic orientation. The logic of down-weighting neutral sentences is that they have less emotional impact in the review with respect to the polarised ones.

3.3. Computing the polarity-driven ratings

Because of the unboundedness of the scores calculated in Eq. 2, we bring all the r_{d_i} into a [0,1] range, where 0 represents the maximum negativity and 1 represents the maximum positivity. For each category c_h belonging to the rating system, the polarity values are computed according to a unity-based normalisation (also known as *feature scaling*):

$$\hat{r}_{d_i} = \frac{r_{d_i} - \min\limits_{d_i \in c_h} r_{d_i}}{\max\limits_{d_i \in c_h} r_{d_i} - \min\limits_{d_i \in c_h} r_{d_i}} \tag{3}$$

The rate assigned to each review is obtained by the algebraic sum of the original rate c_h together with the polarity score $\hat{r}_{d_i}$. The transformation induced by the proposed strategy leads to rating values on a continuous scale. The resulting new rating system has a [1,H+1] range, where 1 expresses the strongest criticism about the reviewed product or service, and H+1 expresses instead the strongest appreciation. The polarisation of the rating scale introduces also a useful finer-grained scale of the reviews. These means that it is possible to read the reviewer's opinions also in terms of ranking, from the worst to the best review. Fig. 1 graphically illustrates how the proposed strategy works.

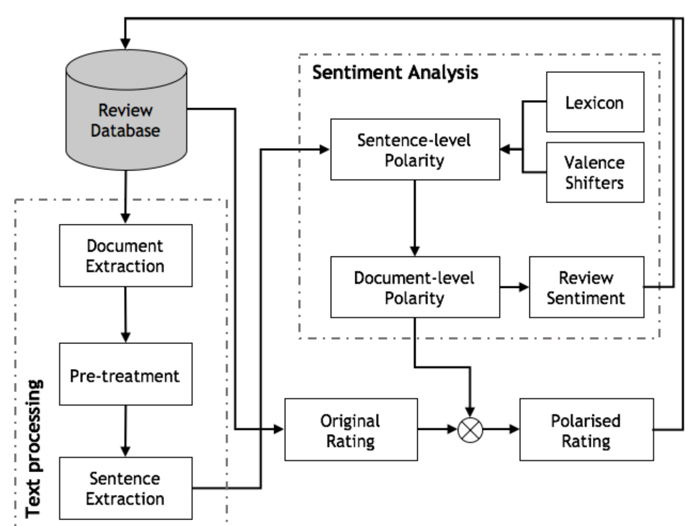

Figure 1: Flowchart of the proposed method

4. Experimental evaluation

4.1. Case study description

In the last few decades, several private and public institutions operating in the field of cultural heritage have considered the visitors in a customer satisfaction perspective. *Audi-*

ence analysis is becoming strategically central, because it frequently has a direct link with the sustainability of the institutions (Sheng & Chen, 2012; Jones, 2015). In this framework, it is increasingly important to measure satisfaction by means of different tools. As stated by Padilla-Meléndez & del Águila-Obra (2013), Web and social media usage has to be considered to explain online value creation by museums. 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.

TripAdvisor is one of the most popular website of travel reviews, and is becoming a fundamental source of information about preferences and trends in tourism. It was founded in the U.S. in February 2000. Since mid-2010, it is both an online service on the Web and a mobile application on portable devices. At present, it operates in 49 markets and is available in 28 languages. According to the TripAdvisor Fact Sheet, it contains 475 million reviews and opinions from travellers concerning 7 million businesses in more than 137,000 destinations, including about 1.1 million accommodations, 4.3 million restaurants and 760,000 touristic attractions. TripAdvisor uses a 1-to-5 rating scale, where the rating categories are associated with the terms *terrible*, *poor*, *average*, *very good* and *excellent*, respectively. Each rate is graphically represented with a corresponding number of bullets.

In the following, we evaluate the audience of the Uffizi Gallery in Florence (Italy), by analysing a set of reviews published on TripAdvisor. The Uffizi Gallery is one of the most important Italian museums, and it is also one of the largest and best-known museums in the world. According to the Italian Ministry of Cultural Heritage and Activities and Tourism, in 2016, 2 million people visited the Uffizi Gallery, and it is one of the preferred attractions of both Italian and International tourists[2].

4.2. Data collection and pre-processing

We used a scraping approach by launching a custom crawler on February 11th 2017. The Web crawler (see Fig. 2 for system architecture) uses a list of Uniform Resource Locators (URLs) to visit, namely the seed URLs, as input. Along with these URLs, some keywords are also provided to check the content relevance. When the Web crawler is initialised for the first time, the queue is built and populated with the seed URLs.

In each iteration, the crawling process checks the status of the queue. If it is empty, the crawling process terminates, otherwise the scheduler module − which defines the policies on how to manage the queue and the pool of downloader threads − selects the next URL.

[2] `http://www.statistica.beniculturali.it` (available only in Italian)

The downloader thread fetches the web pages from the Web indicated by the URL, and downloads it. In the data processing module, the HTML page is analysed to retrieve the reviews and other useful URLs. After this process, the reviews with their metadata are saved in a repository, while the other URLs are put in the queue.

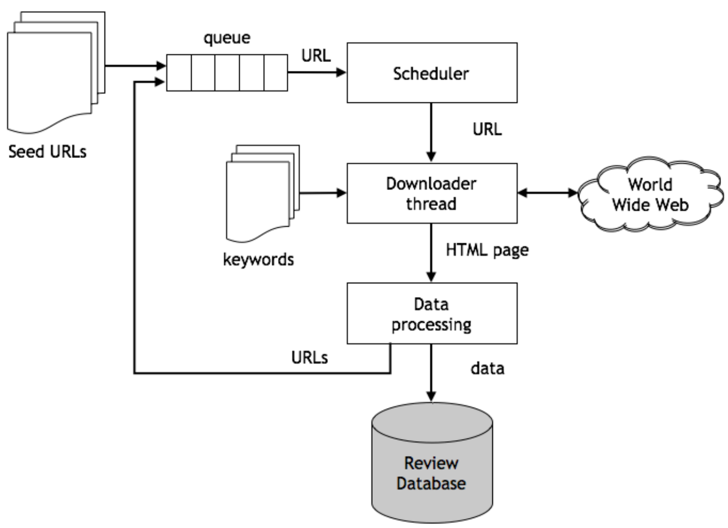

Figure 2: System architecture of the crawler

We retrieved 9,639 reviews written in English and posted on TripAdvisor between February 27^{th} 2003 and February 10^{th} 2017. The crawler also provided some metadata 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 decided to not perform any lexical pre-treatment on the reviews. Only the parts not written in English have been deleted, because reviews sometimes also contain sentences in the mother-tongue language of the authors.

Table 2: Rating distribution of the reviews about the Uffizi Gallery

Rating	Number of reviews	%	Average length (terms)
●○○○○	100	0.23%	137.47
●●○○○	236	1.10%	116.75
●●●○○	918	6.42%	95.42
●●●●○	2,322	21.64%	83.14
●●●●●	6,063	70.62%	71.48
Total	9,639		78.36

Tab. 2 shows the rating distribution of the reviews about the Uffizi Gallery written in English. The average rating is 4.45 bullets. The number of terrible and poor reviews is quite low with respect to very good and excellent reviews. It is also interesting to note that, on average, the reviews with a low rating are longer than the reviews with a high rating.

4.3. Experimental set-up

We decided to implement our strategy by using R. The text-preprocessing was performed with the packages *tm* and *korpus*, while the polarity calculation was performed with the package *sentimentr*. The polarity score of each sentence depends on the lexicon of polarised terms used in the analysis, while the polarity of the whole document depends on the polarities of its sentences. Both lexicon created manually (e.g. Tong, 2001) and lexicon created automatically or semi-automatically (e.g. Turney & Littman, 2003) can be considered. There are many papers in literature dealing with the problem of choosing a proper lexicon (Bravo-Marquez, Mendoza & Poblete, 2014). In order to assign the polarity to each term − and evaluate more effectively the polarity at a sentence level and at a review level − we decided to test different resources. These resources was originally developed for specific purposes, but widely used in the literature for several applicative domains. Tab. 3 reports the size, the polarity distribution and the main reference for each lexicon.

Table 3: Size, polarity distribution and reference of the tested lexicons

Lexicon	Terms	Negative	Neutral	Positive	Reference
afinn	2,477	64.51%	0.04%	35.45%	Nielsen (2011)
hu-liu	6,874	70.37%	0.19%	29.44%	Hu & Liu (2004)
jockers	10,738	66.65%	0.00%	33.35%	Jockers (2017)
nrc	5,468	59.27%	0.00%	40.73%	Mohammad & Turney (2010)
sentiword	20,094	54.89%	0.83%	44.29%	Baccianella, Esuli & Sebastiani (2010)
slang	48,277	76.31%	0.00%	23.69%	Wu, Morstatter & Liu (2016)
so-cal	3,290	50.06%	0.00%	49.94%	Taboada et al. (2011)
vadar	7,236	55.85%	0.00%	44.15%	Hutto & Gilbert (2014)

The criteria for determining if a term is neutral varies from one lexicon to another. Looking at the composition of the 8 lexicons, it is possible to see that the number of neutral terms is mostly equal to 0. This means that all the terms not included in a given lexicon will be considered neutral, even if they should have a negative/positive orientation.

We intersected the vocabulary of 15,524 types extracted from the reviews' collection with the different lexicons. Since *hu-liu* and *nrc* lexicons take into account only the polarity orientation, while the other ones assign also a strength value to each term, we considered only the polarity sign for comparing the resources. Tab. 4 shows the polarity distribution of the terms belonging to the collection's vocabulary.

Table 4: Polarity distribution of the vocabulary according to the different lexicons

Lexicon	Negative	Neutral	Positive
afinn	4.23%	92.19%	3.58%
hu-liu	8.02%	85.86%	6.12%
jockers	11.20%	78.09%	10.71%
nrc	6.40%	86.92%	6.69%
sentiword	8.91%	81.54%	9.55%
slang	2.78%	95.99%	1.22%
so-cal	3.27%	92.50%	4.23%
vadar	5.58%	88.38%	6.04%

As we can see, the neutrality level obtained by using the different resources – i.e. the fraction of terms marked as neutral and not providing relevant sentiment information – is quite high. This means that only few terms can be considered as polarised terms in the evaluation of the semantic orientation of the sentences, and hence, of the reviews. We decided to use the *jockers* lexicon in the following, since it shows a lower neutrality level with respect to the other lexicons. For calculating the polarity of each sentence a list of about 100 shifters (negators, amplifiers, de-amplifiers and adversative conjunctions) was also considered, according to the approach shown in Subsec. 3.2.

4.4. Main results

After pre-processing the 9,639 reviews, we obtained 48,684 different sentences. According to our proposal, we computed the polarity of each review. Tab. 5 shows the main statistics about the sentences, classified with respect to their semantic orientation.

Table 5: Statistics on sentences by semantic orientation

Semantic orientation	Negative	Neutral	Positive	Pooled
sentences	7,653	10,072	30,959	48,684
tokens	131,307	113,384	517,841	762,482
types	6,975	5,597	9,719	15,524
hapax	3,318	2,827	4,543	7,228
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 (30,959) is much greater than the number of neutral (10,072) and negative (7,653) ones. The type/token ratios of the negative, neutral and positive sentences – 5.31%, 4.94% and 1.88%, respectively – suggest that the language used by TripAdvisor users is quite repetitive, and with a low lexical complexity. Neutral sentences have a higher hapax/type ratio. This result is not surprising if we consider that the terms not included into the lexicon are considered neutral.

To visualise the peculiar language associated with positivity and negativity, the *sub-corpora* of positive and negative sentences obtained from the reviews can be analysed. After constructing the *terms* $\times$ *terms* co-occurrence matrices, the relations between the different terms are visualised as textual networks (through the R package *igraph* and the software IRAMUTEQ[3]). It is possible to highlight the main topics associated with positivity and negativity by using the so called *community detection* (Girvan & Newman, 2002). Communities are groups of vertices which probably share common properties and/or play similar roles within the network. In our analysis, each community represents a different topic related to the Uffizi experience of the visitor. These results show the richness of the information embedded into the textual content of the reviews.

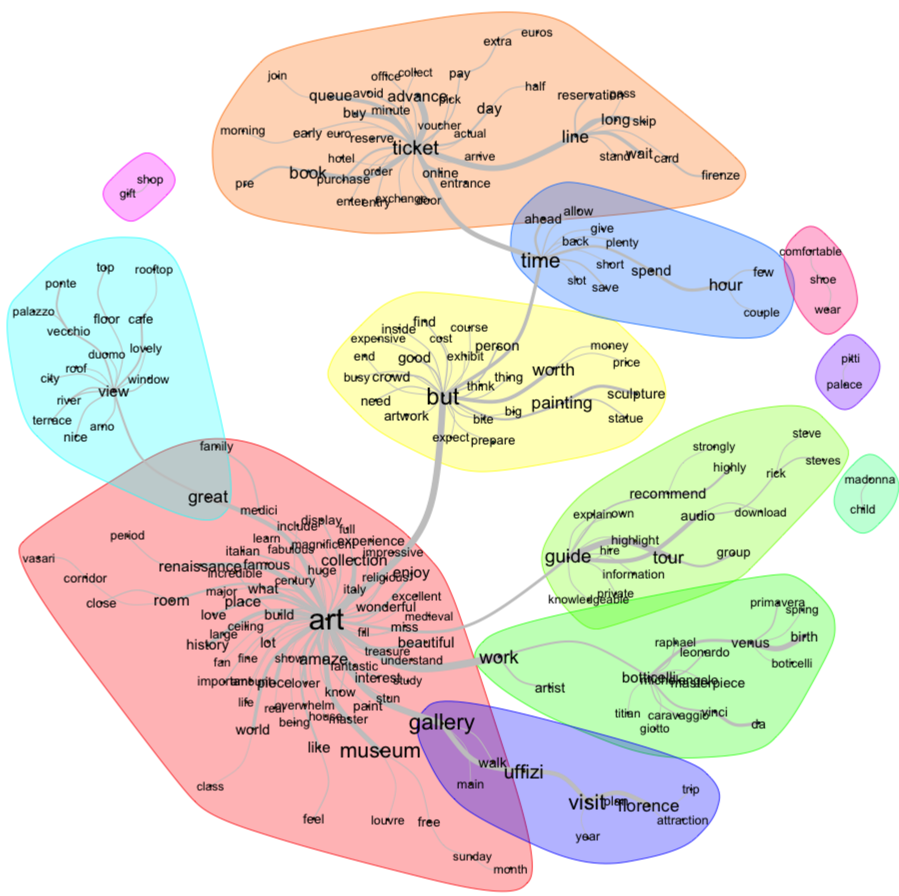

Figure 3: Community detection on co-occurrence network of terms: positive sentences

In Fig. 3, the communities of terms related to positivity are highlighted in different colours.

[3]`http://www.iramuteq.org/documentation`

The main aspects considered by visitors relate to how the tickets were bought, the option of reserving a guided tour, the different aspects related to the concept of Art, and the most important Masters in the gallery. We note the term “but” in the middle (in terms of edge-betweenness) of the network. Its adversative role gives, as seen above, a different weight to the sentence polarities. This means some aspects with a different sentiment orientation are also included in the positive sentences/reviews.

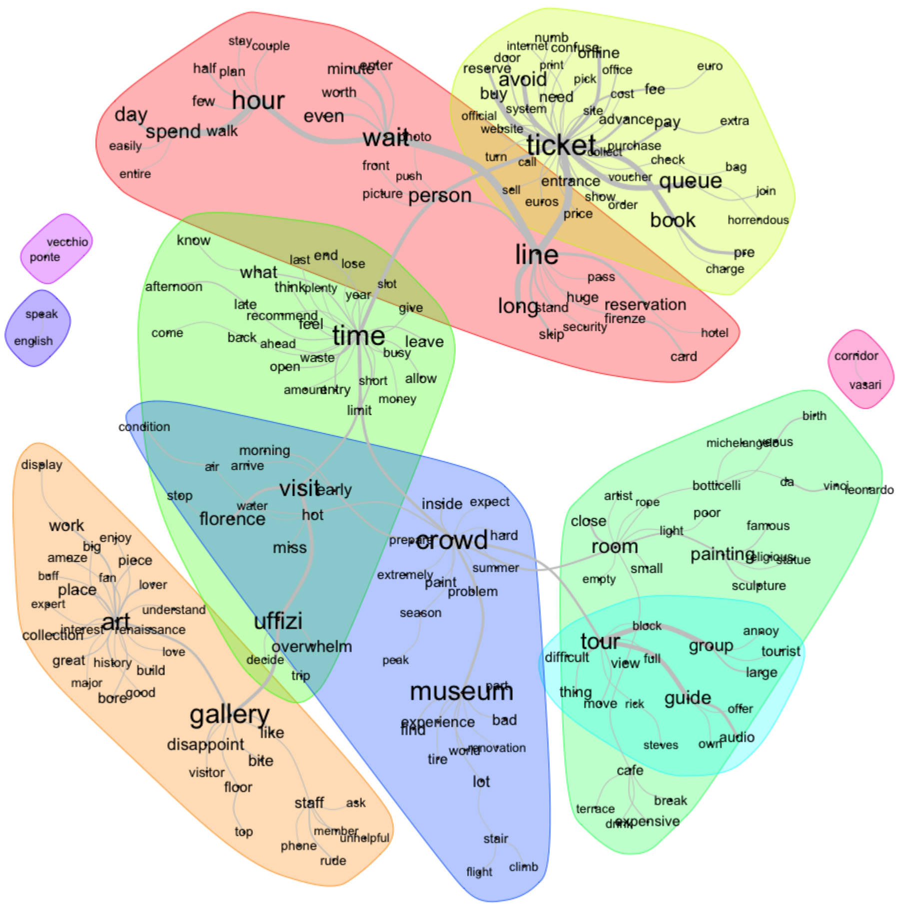

Figure 4: Community detection on co-occurrence network of terms: negative sentences

Analogously, Fig. 4 highlights the communities related to the negative sentences. It is interesting to note that, although we find some topics in common in the two networks, there are different paths. For example, the terms “art” and “gallery” in the network of negative sentences are related to the inefficiency of the “staff”, while in the network of positive sentences the same terms are used to describe the visitor experience (see Fig. 3).

It is interesting to note that high-rated reviews have a greater dispersion of polarity with respect to low-rated reviews. Fig. 5 shows the variability of the document-level polarity for each rating category. The range for the *1bullet* reviews is [-0.36,0.51], with an average polarity of 0.012. The range for the *2bullets* reviews is instead [-0.51,0.95], with an average polarity of 0.060. The range for the *3bullets* reviews is [-0.67,1.29], with an average polarity of 0.144. The range for the *4bullets* reviews is [-0.85,1.89], with an average polarity of 0.226. The range for the *5bullets* reviews is [-0.53,1.89], with an average polarity of 0.257.

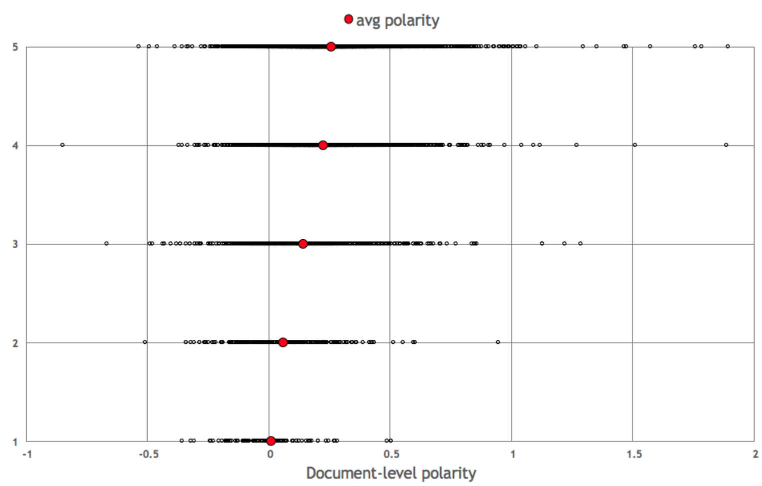

Figure 5: Scatterplot of the document-level polarity by rating category

As we can see from these values, there are negative and positive reviews in each category. Nevertheless, the negativity/positivity have a different impact in the user narration. On the other hand, looking only at the number of bullets does not enable the different levels of satisfaction to be identified. The polarity-driven rating copes with these limitations.

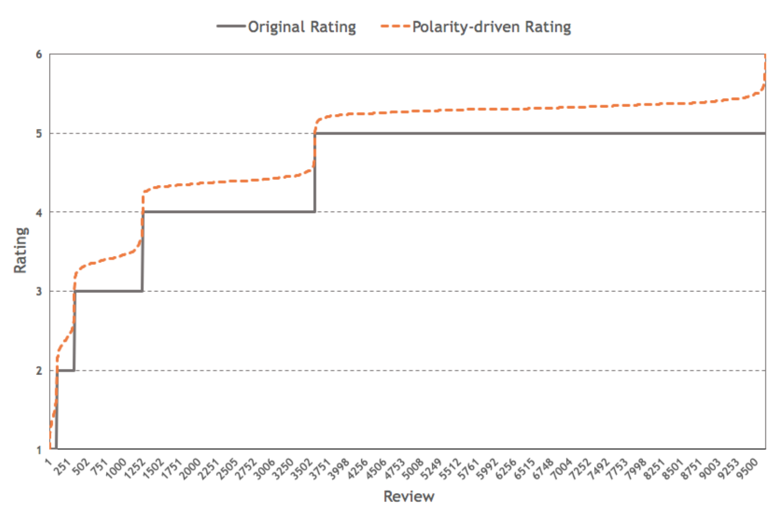

Figure 6: Distributions of the original ratings and the polarity-driven ratings

According to our strategy, the polarity of each review is combined with the rating to obtain the new measure. Fig. 6 shows the distribution of the original ratings and the distribution of the polarity-driven ratings, respectively.

The rating scores lie in a continuous interval [1,6] instead of a discrete interval [1,5]. Using this new rating system leads to a more informative scale than the original bullet scale, or the rating scale inferred from the textual content of the reviews. It is possible to discriminate the different grade of negativity/positivity of the rating categories, taking into account the sentiment of the reviews.

Two examples of reviews about the Uffizi Gallery, both rated with *1bullet* by the contributors, are shown below:

> **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.*

As we can see in these reviews, the sentiment associated with user feelings has a different impact on the overall evaluation. The polarity values computed as in Eq. 3 are 0.0 and 0.9, respectively. The resulting polarity-driven rating is 1.0 for the first review and 1.9 for the second review. This result confirms the effectiveness of the proposed strategy.

5. Conclusions and future developments

In this paper we present a new strategy for measuring user satisfaction in rating and review social media. Our proposal takes into account both the overall evaluation given by the rating scale and the sentiment underlying the written review, in terms of polarity score, obtaining what we called a "polarity-driven rating". The main advantage of using a polarity-driven rating is that we have a finer-grade continuous scale, which is more informative with respect to an ordinal scale. Usually, the ordinal value expressed in terms of bullets or stars is used by social media to evaluate a product or a service. The texts are commonly read only by the users to better understanding the positive and negative aspects related to the reviewed items. The proposed strategy combines the two sources of information in an addictive way. The sum allows to give in each category a similar importance to the review polarity, discriminating between harsh and lenient judgments. Other alternatives can be evaluated. It would be interesting to consider a prior to posterior approach, by introducing the polarity in terms of likelihood function. A study on the distribution of the new rating − and the corresponding conjugate prior − will be conducted in the development of this research.

One of the current assumptions of the proposed measure is that the original rating, usually known *a-priori*, has a strong influence on the final rating. We assumed that this user evaluation is consistent with the sentiment of the review, even if empirical evidences showed that in some cases this is not completely true. Moreover, the polarity scores depend on the lexicon used to identify the positive and the negative terms. The use of valence shifters allow to consider the context of each term and increases the effectiveness of the polarity calculation. An open issue in sentiment analysis is that is not possible to capture the peculiarities of the figurative language, e.g. sarcasm. Some meta-information about the style should be included in order to improve sentiment detection. In the future, we want to conduct an in-depth study relating to the consistency of the evaluations obtained by the ratings and the evaluations obtained by the reviews. We also want to consider the combined use of different lexicons in the polarity calculation step.

Furthermore, with our proposal it is possible to rank the reviews, sorting user experiences from the lowest to the highest appreciation of the product/service. This means that the review sentiment can also be included in a ranking algorithm, making more profitable textual information in recommender systems. We want to project an integrated system that automatically retrieves and scores the reviews. This system could be very useful for businesses and institutions that wish to monitor user satisfaction and consider their position with respect to the other competitors.

Acknowledgment

This work was supported by the 2016-17 University of Calabria ex60% grant funded by the Italian Ministry for Education, University and Research (MIUR). The authors acknowledge Antonio D'Ambrosio (University of Naples Federico II) and Bruno Scarpa (University of Padua) for some constructive suggestions. Thanks to two anonymous reviewers for their careful reading of this manuscript and the insightful comments. *This paper is dedicated to the memory of our beloved colleague, Simona Balbi, who recently passed away.*

References

Baccianella, S., Esuli, A., & Sebastiani, F. (2010). Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation* (pp. 19–21).

Balahur, A., & Jacquet, G. (2015). Sentiment analysis meets social media – challenges and solutions of the field in view of the current information sharing context. *Information Processing & Management*, *51*, 428–432.

Baumgartner, H., & Steenkamp, J.-B. E. M. (2001). Response styles in marketing research: a cross-national investigation. *Journal of Marketing Research*, *38*, 143–156.

Bravo-Marquez, F., Mendoza, M., & Poblete, B. (2014). Meta-level sentiment models for big social data analysis. *Knowledge-Based Systems*, *69*, 86–99.

Cambria, E., Schuller, B., Xia, Y., & Havasi, C. (2013). New avenues in opinion mining and sentiment analysis. *IEEE Intelligent Systems*, *28*, 15–21.

Cosma, G., & Acampora, G. (2016). A computational intelligence approach to efficiently predicting review ratings in e-commerce. *Applied Soft Computing*, *44*, 153–162.

Deng, Z.-H., Luo, K.-H., & Yu, H.-L. (2014). A study of supervised term weighting scheme for sentiment analysis. *Expert Systems with Applications*, *41*, 3506–3513.

Doaa, M. E.-D. M. (2016). A survey on sentiment analysis challenges. Journal of King Saud University – Engineering Sciences. doi:`https://doi.org/10.1016/j.jksues.2016.04.002`.

Gann, W.-J., Day, J., & Zhou, S. (2014). Twitter analytics for insider trading fraud detection system. In *Proceedings of the second ASE international conference on Big Data*.

Girvan, M., & Newman, M. E. J. (2002). Community structure in social and biological networks. *Proceedings of the National Academy of Sciences of the United States of America*, *99*, 7821–7826.

He, W., Zha, S., & Li, L. (2013). Social media competitive analysis and text mining: A case study in the pizza industry. *International Journal of Information Management*, *33*, 464–472.

Hennig-Thurau, T., & Walsh, G. (2004). Electronic word-of-mouth: Motives for and consequences of reading customer articulations on the internet. *International Journal of Electronic Commerce*, *8*, 51–74.

Hu, M., & Liu, B. (2004). Mining and summarizing customer reviews. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 168–177).

Hutto, C. J., & Gilbert, E. (2014). Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media* (pp. 216–225).

Jockers, M. L. (2017). An r package for the extraction of sentiment and sentiment-based plot arcs from text. URL: `https://github.com/mjockers/syuzhet`.

Jones, C. (2015). Enhancing our understanding of museum audiences: visitor studies in the twenty-first century. *Museum & Society*, *13*, 539–544.

Khairullah, K., Baharum, B., Aurnagzeb, K., & Ashraf, U. (2014). Mining opinion components from unstructured reviews: A review. *Journal of King Saud University – Computer and Information Sciences*, *26*, 258–275.

Khan, F. H., Qamar, U., & Bashir, S. (2016). Sentimi: Introducing point-wise mutual information with sentiwordnet to improve sentiment polarity detection. *Applied Soft Computing*, *39*, 140–153.

Leung, C. W.-K., Chan, S. C.-F., & Chung, F.-L. (2011). A probabilistic rating inference framework for mining user preferences from reviews. *World Wide Web*, *14*, 187–215.

Lin, Y., Zhang, J., Wang, X., & Zhou, A. (2012). An information theoretic approach to sentiment polarity classification. In *Proceedings of the 2nd Joint WICOW/AIRWeb Workshop on Web Quality* (pp. 35–40).

Liu, B. (2015). *Sentiment Analysis: mining opinions, sentiments, and emotions*. Cambridge University Press.

Liu, B., Hu, M., & Cheng, J. (2005). Opinion observer: Analyzing and comparing opinions on the web. In *Proceedings of the 14th IW3C2 conference* (pp. 342–352).

Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal*, *5*, 1093–1113.

Mohammad, S. M., & Turney, P. D. (2010). Emotions evoked by common words and phrases: using mechanical turk to create an emotion lexicon. In *Proceeding of Workshop on Computational Approaches to Analysis and Generation of Emotion in Text* (pp. 26–34).

Nguyen, T., Chang, K., & Hui, S. (2011). Supervised term weighting for sentiment analysis. In *Proceedings of 2011 IEEE International Conference on Intelligence and Security Informatics* (pp. 89–94).

Nielsen, F. Å. (2011). A new anew: Evaluation of a word list for sentiment analysis in microblogs. In M. Rowe, M. Stankovic, A.-S. Dadzie, & M. Hardey (Eds.), *Proceedings of the Workshop on Making Sense of Microposts: Big things come in small packages* (pp. 93–98). volume 718.

Okanohara, D., & Tsujii, J. (2005). Assigning polarity scores to reviews using machine learning techniques. In R. Dale, K. Wong, J. Su, & O. Kwong (Eds.), *Natural Language Processing – IJCNLP 2005* (pp. 314–325). volume 3651 of *Lecture Notes in Computer Science*.

Padilla-Meléndez, A., & del Águila-Obra, A. R. (2013). Web and social media usage by museums: Online value creation. *International Journal of Information Management*, *33*, 892–898.

Pang, B., & Lee, L. (2005). Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics* (pp. 115–124).

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, *2*, 1–135.

Petz, G., Karpowicz, M., Fürschuß, H., Auinger, A., Stříteský, V., & Holzinger, A. (2013). Opinion mining on the web 2.0 – characteristics of user generated content and their impacts. In A. Holzinger, & G. Pasi (Eds.), *Human-Computer Interaction and Knowledge Discovery in Complex, Unstructured, Big Data. Third International Workshop, HCI-KDD 2013* (pp. 35–46). Springer Berlin Heidelberg volume 7947 of *Lecture Notes in Computer Science*.

Petz, G., Karpowicz, M., Fürschuß, H., Auinger, A., Stříteský, V., & Holzinger, A. (2014). Computational approaches for mining user's opinions on the web 2.0. *Information Processing & Management*, *50*, 899–908.

Polanyi, L., & Zaenen, A. (2004). Contextual valence shifters. In J. G. Shanahan, Y. Qu, & J. Wiebe (Eds.), *Computing Attitude and Affect in Text: Theory and Applications* (pp. 1–10). Springer Netherlands. volume 20 of *The Information Retrieval Series*.

Qazi, A., Raj, R., Hardaker, G., & Standing, C. (2017). A systematic literature review on opinion types and sentiment analysis techniques: Tasks and challenges. *Internet Research*, *27*, 608–630.

Ravi, K., & Ravi, V. (2015). A survey on opinion mining and sentiment analysis: tasks, approaches and applications. *Knowledge-Based Systems*, *89*, 14–46.

Saif, H., He, Y., Fernandez, M., & Alani, H. (2016). Contextual semantics for sentiment analysis of twitter. *Information Processing & Management*, *52*, 5–19.

Sandes, F. S., & Urdan, A. T. (2013). Electronic word-of-mouth impacts on consumer behavior: Exploratory and experimental studies. *Journal of International Consumer Marketing*, *25*, 181–197.

Serrano-Guerrero, J., Olivas, J., Romero, F., & Herrera-Viedma, E. (2015). Sentiment analysis: A review and comparative analysis of web services. *Information Sciences*, *311*, 18–38.

Sheng, C.-W., & Chen, M.-C. (2012). A study of experience expectations of museum visitors. *Tourism Management*, *33*, 53–60.

Shimada, K., & Endo, T. (2008). Seeing several stars: A rating inference task for a document containing several evaluation criteria. In T. Washio, E. Suzuki, K. Ting, & A. Inokuchi (Eds.), *Advances in Knowledge Discovery and Data Mining: 12th Pacific-Asia Conference* Lecture Notes in Artificial Intelligence (pp. 1006–1014).

Taboada, M., Brooke, J., Tofiloski, M., Voll, K., & Stede, M. (2011). Lexicon-based methods for sentiment analysis. *Computational Linguistics*, *37*, 267–307.

Tan, L.-W., Na, J.-C., Theng, Y.-L., & Chang, K. (2011). Sentence-level sentiment polarity classification using a linguistic approach. In C. Xing, F. Crestani, & A. Rauber (Eds.), *Digital Libraries: For Cultural Heritage, Knowledge Dissemination, and Future Creation* (pp. 77–87). volume 7008 of *Lecture Notes in Computer Science*.

Tang, H., Tan, S., & Cheng, X. (2009). A survey on sentiment detection of reviews. *Expert Systems with Applications*, *36*, 10760–10773.

Tong, R. (2001). An operational system for detecting and tracking opinions in on-line discussions. In *Working Notes of the SIGIR Workshop on Operational Text Classification* (pp. 1–6).

Turney, P. D., & Littman, M. L. (2003). Measuring praise and criticism: Inference of semantic orientation from association. *ACM Transactions on Information Systems*, *21*, 315–346.

Vechtomova, O. (2017). Disambiguating context-dependent polarity of words: An information retrieval approach. *Information Processing & Management*, *53*, 1062–1079.

Wilson, T., Wiebe, J., & Hoffmann, P. (2005). Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing* (pp. 347–354).

Wu, L., Morstatter, F., & Liu, H. (2016). Slangsd: Building and using a sentiment dictionary of slang words for short-text sentiment classification. *CoRR*, *abs/1168.1058*, 1–15.

Xia, R., Xu, F., Yu, J., Qi, Y., & Cambria, E. (2016). Polarity shift detection, elimination and ensemble: A three-stage model for document-level sentiment analysis. *Information Processing & Management*, *52*, 36–45.

Xue, W., Li, T., & Rishe, N. (2017). Aspect identification and ratings inference for hotel reviews. *World Wide Web*, *20*, 23–37.