Noname manuscript No.
(will be inserted by the editor)

Fusing hotel ratings and reviews with hesitant terms and consensus measures

Jennifer Nguyen · Jordi Montserrat-Adell · Núria Agell* · Monica Sánchez · Francisco J. Ruiz

Received: date / Accepted: date

Abstract People have come to refer to reviews for valuable information on products before making a purchase. Digesting relevant opinions regarding a product by reading all the reviews is challenging. An automated methodology which aggregates opinions across all the reviews for a single product to help differentiate any two products having the same overall rating is defined. In order to facilitate this process, rating values, which capture the overall satisfaction, and written reviews, which contain the sentiment of the experience with a product, are fused together. In this manner, each reviewer's opinion is expressed as an interval rating by means of hesitant fuzzy linguistic term sets. These new expressions of opinion are then aggregated and expressed in terms of a central opinion and degree of consensus representing the agreement among the sentiment of the reviewers for an individual product. A real case example based on 2,506 TripAdvisor reviews of hotels in Rome during 2017 is provided. The efficiency of the proposed methodology when discriminating between two hotels is compared with the TripAdvisor rating and median of reviews. The proposed methodology obtains significant differentiation between product rankings.

Keywords hesitant fuzzy linguistic term sets · linguistic decision making · consensus models · tourism · reviews

1 Introduction

Marketing research has found that consumers influence each other in their decision making process [4]. On internet platforms, this influence is derived from ratings and

* Corresponding author: nuria.agell@esade.edu (Núria Agell)

J. Nguyen, M. Sánchez, and F.J. Ruiz
Universitat Politècnica de Catalunya - BarcelonaTech, Edifici Omega, Despatx 342, C. Jordi Girona, 1-3, 08034 Barcelona, Spain

J. Montserrat-Adell and Núria Agell
ESADE Business School, Ramon Llull University, 59 Av. Torreblanca, Sant Cugat 08172 Spain

reviews [2]. These reviews facilitate the decision-making process between people using the same platform; particularly, in the case of experiential good. Consumer reviews are important for products such as destinations, hotels, and restaurants because it is difficult for people to assess their quality before consuming them [14]. Hence, online ratings and reviews serve as a word-of-mouth providing indirect experiences to interested consumers [36]. According to Nielsen[1], 70% of social media users go online to read about other people's experiences with an item at least once a month.

Several online communities such as Tripadvisor [2], Yelp[3], and Booking[4] have become a preferred source of information in tourism and hospitality. However, while these communities facilitate consumers' search for information, it is difficult for them to process and judge it [14]. When online consumers view a product's website, they do not necessarily know the *true* quality of the product because they cannot touch it. Therefore, the consumer may not be able to judge a product's quality precisely [7], so he relies on the reviews and ratings posted to these online communities for information.

One study by Gavilan et al. [8], found that users trust low ratings more than high ratings. In addition, they identified a moderating effect between the relationship of the number of ratings and their trustworthiness. The trustworthiness of low ratings was not impacted by the number of ratings. In contrast, high ratings were found to be trustworthy only when expressed by a high number of reviews.

Several studies have shown that people are more comfortable expressing their preferences in an abstract manner based on linguistic models rather than purely in a quantitative manner [1, 9, 31]. Decision-support systems that consider linguistic values to describe alternatives have been developed to facilitate Group Decision-Making (GDM). These linguistic descriptors enable systems to handle the imprecision involved in decision processes as intervals or fuzzy values [1, 16]. Rodríguez et al. [25] introduced Hesitant Fuzzy Linguistic Term Sets (HFLTSs) over a well-ordered set of linguistic labels to reflect the hesitancy inherent in human reasoning. Several decision making approaches and applications based on HFLTSs have been developed [6, 10, 18, 26, 27, 28]. In addition, some contributions have analyzed the quantification of the level of agreement or consensus among reviewers by means of HFLTSs [5, 24, 34, 33].

The recommendation task has two distinct components. The first part considers the preferences of the user while the second part ranks the alternatives in order of relevance for the user and recommends an item [29]. In spite of a recommender system's ability to narrow information specific to a user's interest, users are still left with the task of differentiating between the suggested items. Therefore, the focus of this work is on the second part of the recommendation task. Our objective is to ease this differentiating process by introducing a measure of consensus representing the agreement among reviewers of an individual item. As previously mentioned, users trust high ratings only when they were expressed by a high number of reviews and

[1] http://www.nielsen.com/content/dam/corporate/us/en/reports-downloads/2012-Reports/The-Social-Media-Report-2012.pdf

[2] tripadvisor.com

[3] yelp.com

[4] booking.com

low ratings regardless the number of reviews [8]. Therefore, a measure of consensus which takes into account the number of reviews, and the agreement and disagreement among the opinions can expedite the decision making process.

This paper moves in two directions: first, building hesitant terms from reviewer ratings and written reviews and second, to measure the consensus of these reviews for a single item to discriminate between items in the same rating category. In this paper we consider HFLTSs to jointly represent the ratings and the text for each review. This new rating is obtained by incorporating the values obtained from sentiment analysis of the written reviews with the reviewer's rating. Then, a measure of consensus, defined by Montserrat et al. [19], that takes into account both the agreement and disagreement among reviewers, is taken. In this way, the methodology allows us to distinguish differences among ratings that were initially equally informed.

The main contributions of this paper are threefold. First, hesitancy is introduced into each review by combining a reviewer's product rating with sentiment from his written opinion. It provides a compact yet expressive product assessment. A classic approach to sentiment analysis has been chosen in this methodology in order to focus the attention on the development of hesitancy in the reviews. More sophisticated methods such as machine learning and deep learning [3] may have been implemented in this methodology. However, the objective of the paper is to showcase the benefits of expressing reviews as HFLTS. Second, reviews for a single product are represented by their centroid enabling users to grasp the range of opinions without having to read further as it fuses information from text written opinions and ratings. Third, the centroid is accompanied by a corresponding consensus measure which quantifies the level of agreement of the reviews. The proposed methodology is implemented on a real case example. The results show that a list of products having the same initial rating can now be distinguishable from one another by using the proposed methodology. It provides an order to sort the products in a way that can be easily understood by the user.

The rest of this paper is structured as follows: firstly, Section 2 introduces a method to define HFLTS from text reviews and ratings, and summarizes the basic concepts already presented in a previous study [19]. Section 3 introduces our proposed methodology and a real case application of it. Finally, Section 4 contains the main conclusions and lines of future research.

2 Preliminaries

This section briefly explains the concepts of the methodology. Specifically, the extension of sentiment analysis to HFLTSs and the measurement of their consensus.

2.1 Defining HFLTSs from text reviews and ratings

Affective computing and sentiment analysis have the potential to enhance the capabilities of recommender systems by excluding items which received negative feedback from a set of recommendations [3]. While affective computing is focused on detecting emotions [22], sentiment analysis classifies text according to polarity (eg. positive

and negative sentiment) [21]. Cambria [3] identified a list of common tasks between the two new interdisciplinary fields including identifying pro and con expressions in reviews which may influence the reviewer's overall judgement, agreement detection, subjectivity detection, and multimodal fusion. The article concludes with a call for next-generation sentiment-mining systems to better understand natural language opinions.

Recommender systems rely on a set of ratings for any particular item in order to provide users with a ranked list of items. Reviewers may provide ratings in different formats such as numerical ratings, number of stars, or written reviews. Linguistic ratings may be associated with the number of stars. For example, on TripAdvisor, an "average" rating is equivalent to three stars on a scale from one to five. We propose that written reviews along with ratings can be used to determine more representative linguistic expressions of human assessments of an item.

Words mean different things for different people and different contexts [15]. This has been an important issue in linguistic decision making in recent times. To tackle this concern, previous studies proposed personalized individual semantics with respect to HFLTS in large-scale groups of decision makers [12, 13]. In this paper, we address this difference in the interpretation of words by applying some sentiment analysis to the written reviews to capture the entire essence of a review.

In order to define this representative linguistic expression, the sentiment of each review needs to be determined. Some sentiment analysis methods consider degrees of positivity [3], while others focus on binary classification of positive and negative sentiment [32]. Since neutral sentiment is between positive and negative sentiment, it may be viewed as potential noise [32, 30]. Therefore, Valdivia et al. [32] proposed to detect and filter out neutral opinions to improve sentiment classification. Although the technique improves classification performance, our methodology does not require precise classification but rather the essence of positive and/or negative sentiment inherent in the text reviews to be reflected in the linguistic expression. Furthermore, they apply consensus voting between multiple sentiment analysis methods to compensate for their lack of agreement in neutrality detection [32]. Although this technique could potentially be applied to the context of the presented methodology, we have chosen to develop it with a classic approach and highlight the potential use of hesitancy in reviews.

Sentiment analysis with the AFINN lexicon [20] is proposed to evaluate the opinions in the text. AFINN is a list of 2477 English words and phrases rated for valence with an integer between minus five and plus five. Here minus is an indicator of *badness* and plus of *goodness* of a review. In a study of twenty-four off-the-shelf methods of sentiment analysis performed on eighteen labeled datasets, Ribeiro et al. [23] found that no single method achieved the best prediction performance across all the datasets tested. The methods were tested on tweet, comment, and review text. AFINN was second in mean rank for three-class classification (positive, negative, and neutral) across the comments datasets behind the VADER method [11]. VADER is a lexicon and rule-based sentiment analysis tool trained on sentiments in social media that is based on a gold standard. However, AFINN was eighth for two-class classification (positive and negative) for the reviews datasets ahead of VADER. There was no three-class classification for reviews as the datasets did not contain a considerable

number of neutral messages. As our method is interested in modeling all three types of sentiment, we selected AFINN.

Words for each review r are matched to the words in the AFINN dictionary. The output is a set of matching words $\{w_{r1}, \ldots, w_{rp}\}$. For each review, r, each word w_{rj}, $j \in \{1, \ldots, p\}$ is associated with a valence $v(w_{rj}) \in \{-5, \ldots, 5\}$ and a frequency of occurrence $f(w_{rj}) \in \{1, \ldots, q_r\}$ in the review.

Example 1 Let us consider a hotel with five reviews with each review consisting of text and a numerical rating. The results of the sentiment analysis output after applying AFINN is given in Table 1. A set of words identified from the AFINN dictionary along with their associated valence for the five reviews are shown.

Review	Word	Freq	Valence
#1	comfortable	1	2
	friendly	1	2
	lovely	1	3
	perfect	1	3
	wonderful	1	4
	worth	1	2
#2	comfortable	1	2
	easy	1	1
	friendly	1	2
	helpful	1	2
	lovely	1	3
#3	recommend	2	2
	superior	2	2
	free	1	1
	lied	1	-2
	mistake	1	-2
#4	clean	1	2
	confusing	1	-2
	fantastic	1	4
	happy	1	3
	helpful	1	2
	laughing	1	1
	love	1	3
	reached	1	1
	safe	1	1
	secured	1	2
	significant	1	1
	stop	1	-1
#5	comfortable	2	2
	cancelled	1	-1
	clean	1	2
	funny	1	4
	overlooked	1	-1

Table 1: Output of sentiment analysis for hotel reviews

Definition 1 Based on the valence $v(w_{rj})$ and frequency $f(w_{rj})$ of each word in a review, a HFLTS is assigned to each rating. The positive and negative sentiment contributions, S_r^+ and S_r^- are computed according to Equations 1 and 2, respectively.

$$S_r^+ = \frac{\sum_{j=1}^{p} \substack{v(w_{rj})\geq 0} v(w_{rj}) \cdot f(w_{rj})}{\sum_{j=1}^{p} \mid v(w_{rj}) \mid \cdot f(w_{rj})} \tag{1}$$

$$S_r^- = \frac{\sum_{j=1}^{p} \substack{v(w_{rj})<0} \mid v(w_{rj}) \mid \cdot f(w_{rj})}{\sum_{j=1}^{p} \mid v(w_{rj}) \mid \cdot f(w_{rj})} \tag{2}$$

Example 2 Continuing with Example 1, the positive and negative sentiment contributions are computed according to Equations 1 and 2, respectively. The results are provided in Table 2.

Review	S_r^+	S_r^-
#1	1.000	0.000
#2	1.000	0.000
#3	0.692	0.308
#4	0.870	0.130
#5	0.833	0.167

Table 2: Review sentiment contributions for a hotel

Definition 2 Once the positive and the negative sentiment contributions have been calculated for each review r, we can define a HFLTS $[a_{H_r^-}, a_{H_r^+}]$ following Equation 3. We consider a reviewer's hesitancy to be inclusive of the numerical rating R given for the same review r.

$$\begin{aligned} H_r^+ &= \min(5, \lfloor R + R \cdot S_r^+ \rfloor) \\ H_r^- &= \max(1, \lfloor R - R \cdot S_r^- \rfloor) \end{aligned} \tag{3}$$

Example 3 For each of the reviews r in Example 2, a reviewer has assigned a numerical rating R. The HFLTS computed from Equation 3 for each review is shown in Table 3.

Review	R	S_r^+	S_r^-	H
#1	5	1.000	0.000	$\{a_5\}$
#2	5	1.000	0.000	$\{a_5\}$
#3	1	0.692	0.308	$\{a_1\}$
#4	4	0.870	0.130	$[a_3, a_5]$
#5	4	0.833	0.167	$[a_3, a_5]$

Table 3: HFLTS from rating and text review

2.2 Measuring consensus among HFLTSs

This section presents a summary of basic concepts of HFLTSs, including the distance between HFLTSs and the measure of consensus that will be used in the proposed methodology.

From here on, let $\mathcal{S}$ denote a finite totally ordered set of linguistic terms, $\mathcal{S} = \{a_1, \ldots, a_n\}$ with $a_1 < \cdots < a_n$. For the rest of this article, a HFLTS is defined as a set $\{x \in \mathcal{S} | a_i \leq x \leq a_j\}$ that is denoted as $[a_i, a_j]$ if $i < j$ or $\{a_i\}$ if $j = i$. According to [17], $\mathcal{H}_\mathcal{S}$ is defined as the set of all possible HFLTS over $\mathcal{S}$ including the empty HFLTS, $\{0\}$, being $\mathcal{H}_\mathcal{S}^* = \mathcal{H}_\mathcal{S} - \{0\}$.

The set $\mathcal{H}_\mathcal{S}$ is extended to $\overline{\mathcal{H}_\mathcal{S}}$, to include the concepts of *positive HFLTSs*, *negative HFLTSs* and *zero HFLTSs*. The *positive HFLTSs* are the result of an intersection of two HFLTSs with some linguistic terms in common, the *zero HFLTSs* are the result of an intersection of two consecutive HFLTSs, while the *negative HFLTSs* are the result of an intersection of two HFLTSs with no common or consecutive linguistic terms. Hence, *negative HFLTs* are used to capture how big the gap is between non-overlapping assessments.

In addition, in the frame of $\overline{\mathcal{H}_\mathcal{S}}$, an *extended inclusion relation* is presented in [17], and, in this context, the extended connected union and extended intersection operators are considered.

1. The *extended intersection* of H_1 and H_2, $H_1 \sqcap H_2$, is the largest element in $\overline{\mathcal{H}_\mathcal{S}}$ that is contained in H_1 and H_2.
2. The *extended connected union* of H_1 and H_2, $H_1 \sqcup H_2$, is the smallest element in $\overline{\mathcal{H}_\mathcal{S}}$ that contains H_1 and H_2.

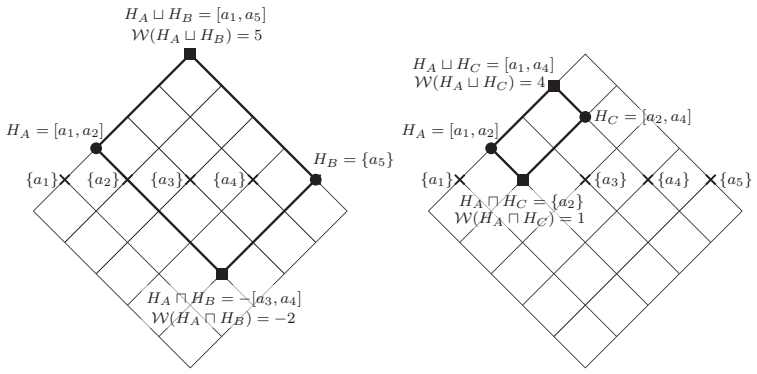
Finally, we consider the distance between two HFLTSs as defined in [17]. Given H_1 and $H_2 \in \overline{\mathcal{H}_\mathcal{S}}$, the *width* of H, $\mathcal{W}(H)$, is defined as the number of linguistic terms contained in H, or cardinality, $card(H)$, if $H \in \mathcal{H}_\mathcal{S}$ or $-card(-H)$ if H is a negative HFLTS. Then the distance between HFLTSs in $\overline{\mathcal{H}_\mathcal{S}}$ is computed between H_1 and H_2, as:

$$D(H_1, H_2) := \mathcal{W}(H_1 \sqcup H_2) - \mathcal{W}(H_1 \sqcap H_2). \tag{4}$$

The following example illustrates the previous concepts:

Example 4 Given a set of possible traveler ratings in linguistic terms: $\mathcal{S} = \{a_1, a_2, a_3, a_4, a_5\}$, being $a_1 =$ *terrible*, $a_2 =$ *poor*, $a_3 =$ *average*, $a_4 =$ *good* and $a_5 =$

excellent, three travelers provided the following linguistic assessments of a hotel: $A =$ *"below average"*, $B =$ *"excellent"* and $C =$ *"not excellent but not terrible"*, whose corresponding HFLTSs by means of $\mathcal{S}$ are $H_A = [a_1, a_2]$, $H_B = \{a_5\}$ and $H_C = [a_2, a_4]$ respectively. Figure 1 shows the extended connected union and extended intersection of H_A and H_B as well as H_A and H_C.



(a) H_A and H_B. (b) H_A and H_C.

Fig. 1: Representation of extended connected union and extended intersection of two HFLTSs.

If we focus on assessments A and B, given that *"below average"* and *"excellent"* are non-overlapping assessments, the extended intersection between them is a negative HFLTS containing the linguistic terms missing between H_A and H_B. Since there are two terms missing, its width is -2. In addition, the extended connected union between them give us all terms from $\mathcal{S}$, that is why its width is 5. If we have a look at assessments A and C, we can see that they share one linguistic term in common, so the width of the extended intersection is 1, while the width of the extended connected union is 4.

According to these results, now we can calculate the distances between H_A and H_B and between H_A and H_C using Equation 4 as follows: $D(H_A, H_B) = 5 - (-2) = 7$, $D(H_A, H_C) = 4 - 1 = 3$.

The distance D is used to obtain a central opinion (or centroid) of a group of reviewers about an item λ as follows:

Definition 3 ([17]) Let λ be an item, G a group of k reviewers and $H_1, \ldots, H_k$ the HFLTS of λ provided by the reviewers in G. Then, the *centroid of the group* is:

$$C_o = \arg \min_{H \in \mathcal{H}_S^*} \sum_{i=1}^{k} D(H, H_i). \tag{5}$$

The centroid is similar to the median of a group. It is a central measure for ordinal scales with hesitancy. In order to ease the calculation of the centroid, [17] proved that, for each specific alternative λ, if $F_H^p(\lambda) = [a_{i_p}, a_{j_p}]$ is the HFLTS used by Decision Maker (DM) p to assess λ, then the set of all the HFLTSs associated with the centroid of the group for λ is:

$$\{[a_i, a_j] \in \mathcal{H}_{\mathcal{S}}^* \mid i \in \mathcal{M}(i_1, \ldots, i_k), j \in \mathcal{M}(j_1, \ldots, j_k)\}, \tag{6}$$

where $\mathcal{M}(x_1, \ldots, x_k)$ is the set that contains just the median of the values if k is odd or any integer number between the two central values sorted from smallest to largest if k is even.

Hence, in order to find the centroid of a group of assessments, it is enough to find the median of the worst linguistic term of each assessment and the same for the best ones. These two terms will be worst and best linguistic terms, respectively, of the centroid.

Example 5 Let G be a group of 5 reviewers who are assessing a hotel λ by means of HFLTSs over the set $\mathcal{S}$ from Example 4, and let H_1, H_2, H_3, H_4, H_5 be the HFLTSs describing their corresponding assessments shown in Table 4. Then, the centroid of the group, C_o, can be calculated through the medians by Equation 6. Since the worst linguistic term from each assessment are a_2, a_2, a_4, a_1, and a_1, we need to find the median of $2, 2, 4, 1$, and 1, which is 2. Therefore, the worst term of the centroid is a_2. Doing the same with the best linguistic terms, we find that the centroid is $[a_2, a_3]$. Figure 2 shows a representation of the centroid, C_o with respect to the HFLTSs.

	H_1	H_2	H_3	H_4	H_5	$\boldsymbol{C_o}$
λ	$[a_2, a_3]$	$\{a_2\}$	$[a_4, a_5]$	$[a_1, a_2]$	$[a_1, a_4]$	$\boldsymbol{[a_2, a_3]}$

Table 4: Centroid of the group G for λ.

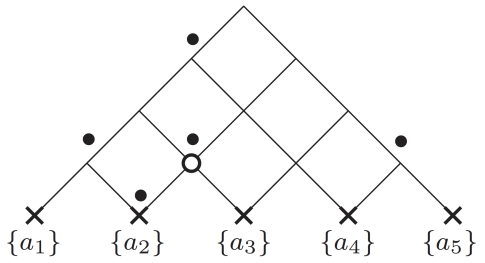


Fig. 2: H_1, H_2, H_3, H_4, H_5 and C_o from Example 5.

Next, in this section we present the consensus degree introduced in [19] that seeks to quantify the agreement between a group of reviewers when rating an item.

Definition 4 ([19]) Let G be a group of k reviewers of an item λ, and $H_1, \ldots, H_k$ be their respective ratings by means of HFLTSs. Let C_o be the centroid review of the group. Then, the *degree of consensus of G on λ* is defined as:

$$\delta_\lambda(G) = 1 - \frac{\sum_{i=1}^{k} D(C_o, H_i)}{k \cdot (n-1)}. \tag{7}$$

Note that $0 \leq \delta_\lambda(G) \leq 1$ due to $k \cdot (n-1)$ is an upper bound of the addition of distances between the centroid and the HFLTSs reviewers [19].

Example 6 Following Example 5, G is a group of 5 reviewers who are assessing a hotel λ by means of HFLTSs over the set $\mathcal{S}$. In Table 5, D_i are the distances from each assessment to the central opinion and $\delta_\lambda(G)$ the degree of consensus of G.

	D_1	D_2	D_3	D_4	D_5	$\sum_{i=1}^{5} D_i$	$\delta_\lambda(G)$
λ	0	1	4	2	2	9	0.45

Table 5: Consensus on the evaluation of a hotel

3 Fusing reviews: A real case example

3.1 Methodology

The concepts introduced in the preliminaries are used to define a methodology which fuse each reviewer's rating and text review for an individual item into a single representation of his opinion. Each reviewer's opinion is expressed in terms of its hesitancy as a HFLTS. The opinions for each item are then aggregated according to this new rating based on the degree of disagreement and agreement among all the reviewers of an item. This process computes a new rating for the item in terms of the centroid of the opinions and the consensus among them and facilitates the ranking between similar items.

The new methodology follows the steps of Figure 3. It begins with several inputs, an item, such as a hotel, a set of reviewer ratings with which it is associated, and the corresponding text reviews.

- **Step 1: Perform Sentiment Analysis on Reviews**
  This step applies the AFINN lexicon explained in Section 2.1 to each review to identify words, their frequency of occurrence, and respective valence as shown in Example 1.

– **Step 2: Express ratings and review sentiment in hesitant terms**
  Once the valence has been obtained for words in each review text, the positive and negative contributions are determined taking into consideration the word sentiment and frequency per review. The sentiment contributions are computed following Equations 1 and 2. This part is demonstrated in Example 2. Next, a HFLTS is assigned to each rating incorporating the sentiment contributions according to Equation 3. This part is shown in Example 3.
– **Step 3: Find the centroid**
  Next, the centroid of the group of reviews which assessed the item is computed with Equation 5. The centroid provides the central opinion of all of the reviews for the item. An example of this calculation is given in Example 5. Once the centroid has been identified, the consensus among the reviewers of the item can be found.
– **Step 4: Measure the consensus** The distance D between each of the HFLTS and the centroid is calculated for the item from Equation 4. These distances enable us to measure the degree of consensus for a hotel according to Equation 7. This step is demonstrated in Example 6. The centroid and consensus represent the new rating for the item.

The process is applied to each item to fuse all of the individual ratings and reviews for each item into an aggregated rating represented by the centroid and consensus of the collective reviews and ratings. Given this new rating, we are able to "totally" order the hotel reviews for each category of rating. In the following subsections, the methodology is applied and evaluated on a real case example using TripAdvisor reviews of hotels in Rome.

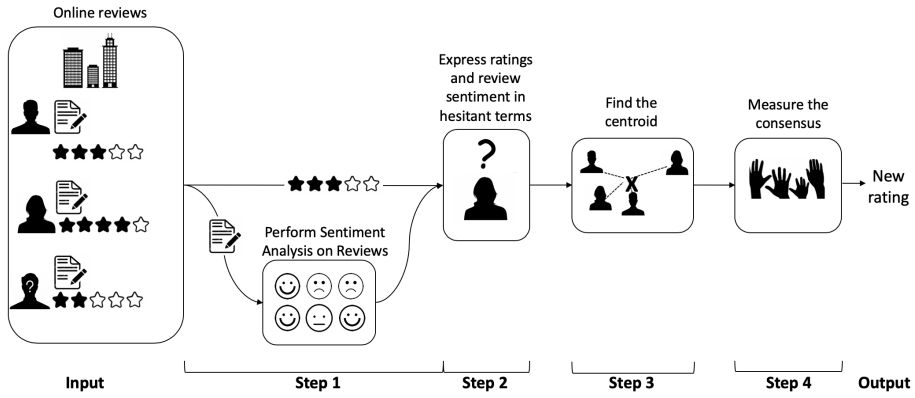


Fig. 3: Methodology to combine ratings and reviews into HFLTS and express their consensus.

3.2 Data Set

In this subsection, we present the data set from the TripAdvisor platform that is used in the experiments conducted to test the viability of our methodology. Xiang et al. [35]

found TripAdvisor reviews to have higher overall quality when compared to other online sites. In addition, the authors concluded that the connections between ratings, helpfulness, and review topics are stronger in TripAdvisor reviews in comparison to other sites suggesting some consistency between written reviews and ratings.

We used a data set of TripAdvisor bed and breakfast reviews in Italy during 2017 provided on Kaggle[5]. The initial set of data contained 31,622 reviews for 2,716 hotels. For each review the title, date, rating, text, language, reviewer id, and property id were provided. In addition, each review contained a rating value, R, associated with a linguistic term a_R, from an ordinal scale $\{a_1, \ldots, a_5\}$. For each hotel the hotel id, name, total number of reviews, average displayed on TripAdvisor, address, and coordinates were provided. We began by narrowing down our data set. Duplicate rows were removed and complete cases were selected resulting in a data set of 30,748 reviews for 2,715 hotels. Table 6 is a summary of the data set. We processed the data and reviewed descriptive statistics including average text length, number of reviews per hotel, and percentage of English reviews following [35]. All non-English reviews were eliminated. The data set could have been filtered for other languages as the AFINN dictionary can accommodate them, as well. Thirty-six percent of the reviews were written in English. As the methodology considers the consensus of a set of reviews for each hotel, we chose hotels that had at least thirty reviews. The final data set consisted of 2,506 reviews for 52 hotels. Figure 4 shows the distribution of the lengths of the reviews across the data set.

	No. of Hotels	**No. of Reviews**	**Avg. No. Reviews / Hotel**
2017	2715	30748	5.98
English only	1846	11047	5.98
Min Rev=30	52	2506	48.19

Table 6: Summary of data set

[5] https://www.kaggle.com/nicodds/rome-wasn-t-built-in-a-day-spotting-fake-reviews

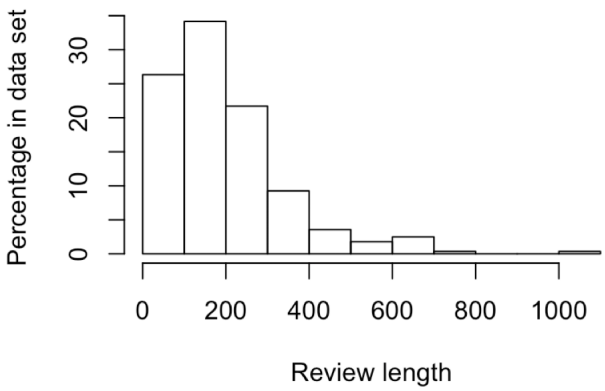


Fig. 4: Distribution of review length in data set

3.3 Experimental Approach

All reviews in English were pre-processed in preparation for semantic analysis. Transformations were applied to convert all reviews to lower case letters, remove numbers and non-alphabetical characters, and shrink white space. Then, stop words that did not contribute to the review meanings were removed. The stop words table included in the R tidytext package was applied and customized to include the term "Rome". The subsequent data set is described in Table 7 and the distribution of the ratings is provided in Table 8.

Avg. age of reviews in days	**Avg. length of reviews (N tokens)**	**Avg. rating**
703.73	29.39	4.73

Table 7: Summary of reviews after text pre-processing

	1	**2**	**3**	**4**	**5**
Rating	21	23	80	373	2009

Table 8: Distribution of review ratings

After performing sentiment analysis on the data set, only 2,439 valid reviews remained in the data set. The difference in the number of reviews can be due to the AFINN dictionary not having a valence associated with the words contained within the "Express ratings and review sentiment in hesitant terms" step. Based on the valence $v(w_{rj})$ and frequency $f(w_{rj})$ of each word in the associated review, positive and negative sentiment contribution, S_r^+ and S_r^-, are computed according to Equations 1 and 2, respectively. A HFLTS $[a_{H_r^-}, a_{H_r^+}]$ is defined for each review following

Equation 3. In order to determine the centroid C_o of the k reviews for a hotel λ, we apply Definition 3 with H_r being the HFLTSs of the reviews determined according to Equation 3. Then the distance D between each of the HFLTSs and the centroid of the reviews for each hotel is computed from Equation 4. The degree of consensus for each hotel is measured via the distance between the reviews and the centroid as defined in Equation 7. In the next section, we present and comment the results that we have obtained.

3.4 Results and Discussion

To analyze the results obtained with the proposed methodology and evaluate its potential applicability, we summarized the top and bottom five ranked hotels in Rome. Using our methodology based on the reviews and ratings, the hotels have been ordered first by their centroid, then by their degree of consensus. The top five ranked hotels are listed in Table 9 and the bottom five ranked hotels are listed in Table 10. With our methodology we can observe that the ranking of hotels is more precise and informative than the TripAdvisor ratings allowing the methodology to rank all of the hotels without concern for ties. Each hotel rating can be discerned from the others.

We can compare the results of the proposed HFLTSs rating with the actual TripAdvisor ratings for the hotels at the moment the data set was extracted. Four out of five hotels in Table 9 have the same TripAdvisor rating. Three out of five of the hotels in Table 10 have the same TripAdvisor rating. In contrast, our methodology allows us to sort all of them. As the centroid of the new ratings encompass those of TripAdvisor, they offer additional information. For example, hotel #49 has the centroid $\{a_5\}$ indicating that central opinion among reviewers is *excellent* which is consistent with the TripAdvisor rating. However, hotel #22 has the centroid $[a_4, a_5]$ suggesting that the central opinion is *good to excellent*, which is more informative than the TripAdvisor rating of 4.5. The new rating implies to the user that customers did not give the hotel a rating of 4.5 definitively or on average but *good to excellent*, drawing the user's attention to the existence of sub 4.5 reviews for the hotel. When combined with the information in the consensus, the user can see from the low degree of consensus that there is wide disagreement among reviewers.

Hotel	**Centroid**	**Consensus**	**TA Rating**	**TA median**
49	$\{a_5\}$	0.928	5	5
6	$\{a_5\}$	0.923	5	5
24	$\{a_5\}$	0.917	4.5	5
48	$\{a_5\}$	0.907	5	5
4	$\{a_5\}$	0.905	5	5

Table 9: Comparison of results of top 5 ranked hotels in Rome

Hotel	Centroid	Consensus	TA Rating	TA median
22	$[a_4, a_5]$	0.663	4.5	4
44	$[a_4, a_5]$	0.639	4.5	5
33	$[a_4, a_5]$	0.554	4.5	5
26	$[a_3, a_5]$	0.564	4	4
2	$[a_3, a_5]$	0.493	4	4

Table 10: Comparison of results of bottom 5 ranked hotels in Rome

In addition, the consensus can facilitate rating interpretation when the centroids for each of the hotels are the same, as can be seen from the rankings in Table 9. This case can be expected when hotels have all been rated highly. Hotels with a low degree of consensus could be seen as debatable in terms of the opinion of customers. For example, both hotels #49 and #4 have their centroids at $\{a_5\}$ indicating that customers have rated both hotels highly and equally. Their respective consensuses provide additional information regarding the aggregated customer ratings. Specifically, there was greater agreement among the reviews and ratings of hotel #49 than hotel #4. Therefore, a user can quickly infer that amidst the highly rated customer reviews, there were some less than perfect.

Note that the degree of consensus can be helpful when comparing two hotels that seem different at first glance. Consider hotels #24 and #6. Hotel #24 was rated 4.5 by TripAdvisor and $\{a_5\}$ by the new rating while #6 was rated 5 by TripAdvisor and $\{a_5\}$ by the new rating. Based on the TripAdvisor rating, a user might think that hotel #6 is considerable better than hotel #24. However, the degree of consensus can assist the user with understanding that the difference is not huge, a difference in agreement of 0.923 and 0.917.

In the event of a tie, where the centroid and consensus of two or more hotels coincide, a tie breaker is computed. In the experiment, the tie breaker is the percentage of reviewers who assessed the centroid.

Finally, to evaluate the performance of the introduced methodology, we compare our results with a ranking of TripAdvisor Ratings, and a ranking of the usual median of all the reviewers' ratings for each hotel. For this comparison, we find the percentage of differentiable pairs of hotels from our 52 hotels by using all the possible combinations. The results are presented in Table 11.

	Proposed ranking	TA rating ranking	TA median ranking
% of differentiation	100%	53.47%	11.09%

Table 11: Performance evaluation of the proposed methodology.

As can be seen from the table, any pair of hotels is distinguishable with the proposed methodology, while only slightly more than half of the possible pairs of hotels are distinguishable using the TripAdvisor rating to rank the hotels. Finally, using the usual median of the reviewers' rating to rank the hotels, almost all of them have the same median, so very few are distinguishable.

4 Conclusions and future work

This paper presents a new methodology to associate an interval rating (hesitant term) together with a measure of consensus to the hotel ratings derived from a group of reviewers. Specifically, it gives readers the ability to extend reviewer opinions from ratings to hesitant fuzzy linguistic term sets by combining the opinion of ratings and written reviews. From each set of extended reviewer opinions it considers the centroid to be the global opinion of each hotel. In this way, group consensus can be measured for each hotel and used to differentiate hotels having the same ratings.

The contributions of this paper are threefold. First, it introduces hesitancy in the assessment of each review by means of sentiment analysis. Second the centroid allows us to fuse the information introduced in the text reviews and ratings. Third, the consensus measure allows us to better understand previous ratings allowing readers to immediately identify which of the hotels will have more variability in their reviews. Given a ranked list of hotels, the centroid and consensus measures help readers to distinguish between hotels which previously had the same rating. For example, it may not be readily understandable why a recommended list of hotels all rated with five stars has been ranked in a specific order. The proposed methodology provides an explanation to the reader with regards to the range of sentiment and the disparity of the opinions. We have presented a tool to help users make decisions with more information than just taking the average of the reviews. From a general perspective, the ability to distinguish between items having the same rating could be beneficial to intelligent personal assistants. Rather than offering a list of the top items based on ratings, an intelligent personal assistant may suggest a single alternative to the user. This scenario would be more reflective of a conversation between friends.

Future research will be focused in two main directions. First, a further study of the properties of the presented consensus degree in comparison with other similar measures will be carried out. Second, to analyze users' preferences and incorporate them into a profile that will help us in terms of recommendations. Hence, we can also apply the presented methodology into the field of recommender systems. Lastly, some experiments will be run to test the applicability of the methodology in real recommendation scenarios. This third direction will consider its applicability in different domains, model performance, and interpretation by real users.

Acknowledgement

This research has been partially supported by the Secretary of Universities and Research of the Department of Enterprise and Knowledge of the Generalitat de Catalunya (2017 DI 086) and by the INVITE Research Project (TIN2016-80049-C2-1-R and TIN2016-80049-C2-2-R (AEI/FEDER, UE)), funded by the Spanish Ministry of Science and Information Technology.

Conflict of Interest

The authors declare that they have no conflict of interest.

References

1. Alonso S, Herrera-Viedma E, Chiclana F, Herrera F (2010) A web based consensus support system for group decision making problems and incomplete preferences. Information Sciences 180(23):4477–4495
2. Amblee N, Bui T (2011) Harnessing the influence of social proof in online shopping: The effect of electronic word of mouth on sales of digital microproducts. International Journal of Electronic Commerce 16(2):91–114
3. Cambria E (2016) Affective computing and sentiment analysis. IEEE Intelligent Systems 31(2):102–107
4. Dichter E (1966) How word-of-mouth advertising works. Harvard business review 44(6):147–160
5. Dong Y, Chen X, Herrera F (2015) Minimizing adjusted simple terms in the consensus reaching process with hesitant linguistic assessments in group decision making. Information Sciences 297:95 – 117
6. Fahmi A, Kahraman C, Bilen U (2016) Electre i method using hesitant linguistic term sets: An application to supplier selection. International Journal of Computational Intelligence Systems 9(1):153–167
7. Fung R, Lee M (1999) Ec-trust (trust in electronic commerce): exploring the antecedent factors. AMCIS 1999 Proceedings p 179
8. Gavilan D, Avello M, Martinez-Navarro G (2018) The influence of online ratings and reviews on hotel booking consideration. Tourism Management 66:53–61
9. Herrera F, Herrera-Viedma E (2000) Linguistic decision analysis: steps for solving decision problems under linguistic information. Fuzzy Sets and systems 115(1):67–82
10. Huang HC, Yang X (2014) Pairwise comparison and distance measure of hesitant fuzzy linguistic term sets. Mathematical Problems in Engineering 2014
11. Hutto C, Gilbert E (2015) Vader: A parsimonious rule-based model for sentiment analysis of social media text. In: Proceedings of the 8th International Conference on Weblogs and Social Media, ICWSM 2014, pp 216–225
12. Li CC, Rodríguez RM, Martínez L, Dong Y, Herrera F (2018) Personalized individual semantics based on consistency in hesitant linguistic group decision making with comparative linguistic expressions. Knowledge-Based Systems 145:156–165
13. Li CC, Dong Y, Herrera F (2019) A consensus model for large-scale linguistic group decision making with a feedback recommendation based on clustered personalized individual semantics and opposing consensus groups. IEEE Transactions on Fuzzy Systems 27(2):221–233
14. Liu Z, Park S (2015) What makes a useful online review? implication for travel product websites. Tourism Management 47:140–151
15. Mendel JM, Zadeh LA, Trillas E, Yager R, Lawry J, Hagras H, Guadarrama S (2010) What computing with words means to me [discussion forum]. IEEE computational intelligence magazine 5(1):20–26
16. Montes R, Sánchez AM, Villar P, Herrera F (2015) A web tool to support decision making in the housing market using hesitant fuzzy linguistic term sets. Applied Soft Computing 35:949–957

17. Montserrat-Adell J, Agell N, Sánchez M, Ruiz FJ (2016) A Representative in Group Decision by Means of the Extended Set of Hesitant Fuzzy Linguistic Term Sets, Springer International Publishing, Cham, pp 56–67
18. Montserrat-Adell J, Agell N, Sánchez M, Prats F, Ruiz FJ (2017) Modeling group assessments by means of hesitant fuzzy linguistic term sets. Journal of Applied Logic 23:40–50
19. Montserrat-Adell J, Agell N, Sánchez M, Ruiz FJ (2018) Consensus, dissension and precision in group decision making by means of an algebraic extension of hesitant fuzzy linguistic term sets. Information Fusion 42:1–11
20. Nielsen FÅ (2011) A new anew: Evaluation of a word list for sentiment analysis in microblogs. Computing Research Repository (CoRR)
21. Pang B, Lee L, et al. (2008) Opinion mining and sentiment analysis. Foundations and Trends® in Information Retrieval 2(1–2):1–135
22. Picard RW (2000) Affective computing. MIT press
23. Ribeiro FN, Araújo M, Gonçalves P, Gonçalves MA, Benevenuto F (2016) Sentibench-a benchmark comparison of state-of-the-practice sentiment analysis methods. EPJ Data Science 5(1):23
24. Rodríguez RM, Martínez L (2015) A consensus model for group decision making with hesitant fuzzy linguistic information. In: 2015 10th International Conference on Intelligent Systems and Knowledge Engineering (ISKE), pp 540–545
25. Rodríguez RM, Martínez L, Herrera F (2012) Hesitant fuzzy linguistic terms sets for decision making. IEEE Transactions on Fuzzy Systems 20(1):109–119
26. Rodríguez RM, Martínez L, Herrera F (2013) A group decision making model dealing with comparative linguistic expressions based on hesitant fuzzy linguistic term sets. Information Sciences 241:28 – 42
27. Rodríguez RM, Liu H, Martínez L (2014) A Fuzzy Representation for the Semantics of Hesitant Fuzzy Linguistic Term Sets, Springer Berlin Heidelberg, Berlin, Heidelberg, pp 745–757
28. Rodríguez RM, Labella Á, Martínez L (2016) An overview on fuzzy modelling of complex linguistic preferences in decision making. International Journal of Computational Intelligence Systems 9(sup1):81–94
29. Rubens N, Elahi M, Sugiyama M, Kaplan D (2015) Active learning in recommender systems. In: Recommender systems handbook, Springer, pp 809–846
30. Sáez JA, Galar M, Luengo J, Herrera F (2016) Inffc: an iterative class noise filter based on the fusion of classifiers with noise sensitivity control. Information Fusion 27:19–32
31. Travé-Massuyès L, Prats F, Sánchez M, Agell N (2005) Relative and absolute order-of-magnitude models unified. Annals of Mathematics and Artificial Intelligence 45(3-4):323–341
32. Valdivia A, Luzón MV, Cambria E, Herrera F (2018) Consensus vote models for detecting and filtering neutrality in sentiment analysis. Information Fusion 44:126–135
33. Wu Z, Xu J (2016) An interactive consensus reaching model for decision making under hesitation linguistic environment. Journal of Intelligent and Fuzzy Systems 31:1635–1644

34. Wu Z, Xu J (2016) Possibility distribution-based approach for magdm with hesitant fuzzy linguistic information. IEEE Transactions on Cybernetics 46(3):694–705
35. Xiang Z, Du Q, Ma Y, Fan W (2017) A comparative analysis of major online review platforms: Implications for social media analytics in hospitality and tourism. Tourism Management 58:51–65
36. Ye Q, Law R, Gu B, Chen W (2011) The influence of user-generated content on traveler behavior: An empirical investigation on the effects of e-word-of-mouth to hotel online bookings. Computers in Human behavior 27(2):634–639