Visualizing Reviews Summaries as a Tool for Restaurants Recommendation

Yaakov Danone
The University of Haifa
Haifa, Israel
dnkobi@gmail.com

Tsvi Kuflik
The University of Haifa
Haifa, Israel
tsvikak@is.haifa.ac.il

Osnat Mokryn
The University of Haifa
Haifa, Israel
ossimo@gmail.com

ABSTRACT
Online customers' opinions about products and services, in the form of reviews, are a major part of today's web culture. However, customers, when looking for a product or service, do not have the time or the desire to read even a small part of the available product reviews (which themselves may be lengthy and not easy to read). Moreover, they often would like to examine reviews of similar products, and get a comprehensive picture of how different aspects of these products compare. In this work, by introducing a generic framework for analyzing and presenting a visual summary based on comparative sentences extracted from customer reviews, we offer the user an easy and intuitive understanding of the differences between a set of products.

The contribution of this study is twofold: First, it focuses on reviews of intangible services (using the restaurant domain as a case study), unlike most of the related studies that consider physical products. Second, it combines state-of-the-art text analysis techniques with an intuitive visualization into an easy to use prototype to visualize summarized service comparisons to the users.

The system's usefulness and intuitiveness were confirmed in multiple user studies.

Author Keywords
Visualizing comparisons; Reviews summarization; Information visualization.

ACM Classification Keywords
• Human-centered computing~Social recommendation • Human-centered computing~Information visualization • Human-centered computing~Empirical studies in HCI

INTRODUCTION
Customer reviews are based on the experience of using the product "in the field" and thus provide more information than the technical evaluation of experts, and obviously, more than the product description provided by the seller [45]. Studies show that customer reviews have a major impact on consumer purchase decisions [29, 13, 28, 46]. In many cases, to better explain and illustrate the differences between products they are reviewing, consumers compare them with similar products, identifying their relative strengths and weaknesses [42]. Comparisons are different from opinions and sentimental evaluations. While opinions and sentimental evaluations are subjective, comparisons can be either subjective or objective [1]. For example, an opinion sentence about a car may be "Car X is very nice". A subjective comparative sentence may be "Car X is much better than Car Y". An objective comparative sentence may be "Car X is 2 feet longer than Car Y". Each statement has different language constructs, and therefore, they should be treated differently than general opinion statements [15].

Today's information overload, where there are too many reviews and opinions available, make it hard to present a compact representation of a product's accumulated reviews to a customer [45]. Reviews can be lengthy, making it hard for the interested reader to understand and exploit them [12]. When focusing on the restaurants domain it is worth noting that differently from reviews of physical goods, in reviews of intangible services, customers tend to "tell stories" and report on their experiences using longer and more complex sentences [24].

One way of addressing this information overload is by generating a summary of reviews, which can be done in various ways. In our study, following past research, we used comparisons as a key factor in summarizing reviews, following [15, 42]. Comparative sentences have a similar structure and they appear in both physical goods and intangible services reviews, so it is reasonable to assume that the approach suggested by [42] may be applied also to restaurant services reviews. To present these summaries to end users, we apply information visualization techniques. Specifically, this work addresses the following research question: **How can visualization of comparison statements extracted from restaurant reviews help the customer in understanding the strengths and weaknesses of a restaurant, with respect to features in which she is interested?**

To answer the above question, the following challenges have been addressed, integrating state of the art research results with simple, intuitive visualizations:

- How to identify a comparison sentence in restaurant reviews?

IUI'18, March 7–11, 2018, Tokyo, Japan

ACM ISBN 978-1-4503-4945-1/18/03…$15.00
https://doi.org/10.1145/3172944.3172947

- How to extract a comparison relation from a comparison sentence in restaurant reviews?

- How to find the direction of the comparison relation?

- How to summarize and visualize an integrated set of comparison sentences about features of interest extracted from reviews for the customer to use?

As a first step in the study, a prototype system that addresses the above challenges using simple common visualizations was developed and evaluated.

RELATED WORK

Consumer products and subsequently their reviews may be classified into one of two categories: physical goods and intangible services [22]. Generally, in physical product reviews, customers write about the features they liked/disliked, without adding much extraneous information. This contrasts with reviews of intangible services. In reviews of the latter type, customers tend to report on their experiences using complex and longer sentences [24]. For example, "*The food is just Ok. Their Happy Hour menu is pretty good. However, if it wasn't for Lynette, the bartender, I probably wouldn't frequent this place as much as I do. She is awesome, friendly, makes SPECTACULAR drinks . . . overall she makes what would be an A-OK experience, simply marvelous!!*" The example taken from a restaurant review talks about the restaurant's physical goods such as food and its intangible services such as the staff. Different from the physical goods, there are multiple mentions of the service that is being reviewed (the bartender) and opinions are given via different and complex expressions referring to the features or subcomponents of the reviewed service. Cruz et al. [4] confirmed the importance of the domain being examined in the field of Opinion Mining (OM) for building accurate opinion extraction systems, while indicating that people express their opinions on a given domain differently. Restaurants provide physical goods (the food) as well as services in the form of ambience or settings. To the best of our knowledge, most existing works, which utilized comparative sentences, focused mainly on pure physical product reviews, as demonstrated by [16, 42, 11, 17, 40]. Few studies did address services, including Gu and Yoo [10] that dealt with restaurants reviews in Korean. They identified comparison sentences as those containing one of the words {than, compared to, far, more} in Korean, and used semantic rule labeling on the resulted sentences. We assume that there may be a few more that we missed, but in general, the focus was on products reviews. In this study we analyze services provided by restaurants, as a case study, focusing on their intangible features as opposed to their physical aspects.

Previous work also tried to eliminate the need to read every review. One basic application, suggested by Turney [38], presented a simple unsupervised learning algorithm for classifying reviews as recommending (thumb up) or not recommending (thumb down), where the customer can promote the review by selecting it as useful or not. Another approach is to summarize the reviews. For example, Hu and Liu [12] identified the features of a particular product in the reviews and classified reviews as positive or negative with respect to these features, producing a summary about the products. Their work was a pioneering effort in OM and was followed by other studies like [44], which created a product ranking system derived from customer reviews. Their study aimed to develop a tool that selected the highest rated products based on their past reviews. It used a weight-directed graph that showed the quality of the products based on their reviews and presented the results as a ranked table for the top-rated products. Another attempt to summaries reviews and opinions was that of Ying and Jiang [43], which used text quality and subjectivity to decide which sentences are good summary sentences.

A comprehensive survey of reviews and text summarizing with the different approaches can be found in Liu and Zhang [21]. The works of Hu and Liu [12] and Zhang et al. [44], as well as most of the similar works in the field, focused on digital cameras, TVs and other electronic products and did not consider intangible services which is the focus of our study.

Comparative Sentence Mining

In a typical customer review, there are two kinds of judgmental sentences [44]:

1. Subjective sentences: Sentences with opinions, positive or negative, about an entity.

2. Comparative sentences: Sentences that compare objects.

The latter sentence type, comparative sentences, is one of the most important ways of evaluating an entity or event [14, 15] and a rich information source for identifying the relative strengths and weaknesses of products compared with their competitors [42]. We believe that this type of sentences can potentially be used as a key factor in summarizing customer reviews; therefore, our work focused on visualization of comparative sentences.

Jindal and Liu [14,15] were pioneers in analyzing comparisons using natural language processing (NLP) techniques and were the first to try to identify and extract comparative sentences from customer reviews. In their work, they had two main objectives:

1. Identify comparative sentences in the text.

2. Identify and extract comparative relations from comparative sentences.

They applied Class Sequential Rules (CSR, see [20]) to identify comparative sentences. To build a CSR training dataset, they treated each sentence as a sequence and manually defined it as belonging to either a comparative or non-comparative class. They identified the comparative keyword in a sentence (for example, adjectives with morphemes such as er/est, or general comparative keywords such as more, less, same, prefer, lead, beat etc.) and used a

radius of words before and after the keyword to capture a sequence. Although their use of comparative keywords identified most of the corpus' comparative sentences, many sentences among these were not comparative (they achieved high recall 98% but with low precision 32%). Comparative keywords together with CSR and Naïve Bayes (NB) classifier gave an overall precision of 79% and an overall recall of 81% in distinguishing between comparative and non-comparative sentences. Originally, the actual words created too many patterns. To overcome this limitation, the authors changed each word (beside the comparative keyword) in the radius to its matched Part of Speech (POS) tag, and this way, the patterns emerged. In addition, Jindal and Liu [14] defined four types of comparative sentences: "non-equal gradable", "equative", "superlative" and "non-gradable". Their method, however, cannot identify the direction of the "non-equal gradable" (better/worse) sentence, which makes it inadequate to judge the sentiment polarities of customers, based on the comparative sentence. Xu et al. [42] improved Jindal and Liu's [14, 15] work by adding direction to the comparative relation and proposed a new form of the comparative relation:

<direction> (<product1>, <product2>, <attribute>, <sentiment>).

They recognized four possible directions: '>', which means P1 is better than P2, '<' worse, '=' equal and '~' no comparison. To extract a comparative relation, at first, Xu et al. [41] had to recognize the product names, the attribute and the sentiment phrase. Since the domain in their work was smartphones, the set of products and attributes is finite. They created a list that included most of the product names and attributes in the smartphone domain. For the sentiment phrases, the authors built a comparative keyword set from known lexical resources [27,6,7] (the same lexicon resources that was used by Jindal and Liu's [14,15] and also in this work). Next, they searched for the direction of a comparative relation. Finding the direction is not a trivial task; there are many comparative sentences that use the same comparative keyword but have opposite directions. For example, the following are restaurant reviews:

1. The Front Porch: "The food here is sooooo much better than Cha Cha Cha!"

2. Zazie: "I personally think the food is better at Citizen Cake".

The first comparative sentence in the example appears in a review about a restaurant, The Front Porch. The comparative keyword in the sentence is 'better' and the direction is in favor of the first restaurant over the second restaurant, Cha Cha Cha. The second comparative sentence contains the same comparative keyword 'better', even the attribute 'food' is the same, but the direction is opposite—in favor of the second restaurant.

To address this challenge, Xu et al. [42] found the connection between the words and the entities in the sentence as well as between the different comparative relations in the comparative sentence. For that purpose, they enhanced the one-level Conditional Random Field (CRF)[1], making it into a two-level CRF model. They used the different levels of the model and encoded the model to calculate the probability of finding the direction of each comparative relation. They used the word, entity and between-relations functions to consider the characteristics of the comparative sentence and decided the right direction of each comparative relation based on the characteristics. In this way, they managed to identify the directions for the relations with an overall accuracy of 66.17% and an average F-score result of 56.68 as shown in Table 2.

Table 2: Performance of the two-level CRF when identifying the direction of the relations ([Xu et al. 2011])

Direction	***Precision (%)***	***Recall (%)***	***F-score (%)***
>	76.63	81.90	78.56
<	59.72	48.39	51.77
=	61.11	23.33	31.19
~	54.29	83.33	65.20
Average	**62.94**	**59.24**	**56.68**

Xu et al. [42] visualized the results of the comparison of several items using a comparative relation map, which an undirected graph. The main node, in the center of the graph, represents the candidate product, to which other products are compared. It is connected to the other products via links, where each link represents comparisons based on one predefined feature. In the middle of each link, there is a red or a blue box. The red box grows in direct relation to the number of positive comparisons in favor of the main product and the blue box grows in direct relation to the number of positive comparisons in favor of competitors.

Still, the work of Xu et al. [42] was limited to the smartphone domain – a specific product with a well-defined set of features while we aim at services which has a wider set of features, not all of them are well defined. In addition, a consultation with a visualization expert confirmed that the comparative relation map visualization is not very intuitive, and necessitates the user getting instructions or training before using it.

Information Visualization

Information visualization is intended to provide an easily understandable interpretation of a huge and complex amount

[1] CRF is a framework for building probabilistic models to segment and label sequential data [19, 32].

of information, while creating visual representations of it that exploit the individual's perceptual and cognitive problem-solving capabilities [18,41]. Many studies in OM use visualization techniques to summarize customers' opinions of products from online reviews. Miao et al. [26] used a pie chart to present the proportion of positive and negative reviews of a product's specific feature. Gamon et al. [9] used a tree map to present common words used to describe features of vehicles. Shamim et al. [33] categorized visualization techniques used in OM into four categories:

1. Radial visualization: Opinion wheel and spider graph.

2. Graph visualization: Coordinated, comparative relation map, positioning map, line graph and pie chart.

3. Hierarchical visualization: Tree map and visual summary report.

4. Bar chart: Glowing bars, bars with different shapes and bar chart.

Usability, quality and expressiveness are important aspects of data usability and visualizations [8]. How can usability, however, be measured? Some approaches explain and define usability measurement aspects [25,34,2]. Bai et al. [2], for instance, defined seven visualization assessments: psychology of the observer, visual representation model, visual impact, overall performance, overall design style, information quality, and information presentation model. Another usability measurement is the System Usability Scale (SUS) [3], which is a standard questionnaire that is commonly used and contains 10 questions to measure the usability of a system.

TOOLS AND METODS

This study followed a "design study" approach [39]. Prototypes providing solutions for the individual research questions were developed, and then we developed research prototype that enabled us to demonstrate and evaluate these solutions using a publicly available dataset.

Dataset

For our experimentation, we used a dataset extracted from Yelp[2]. It contains approximately 270,000 Restaurant reviews taken from the San Francisco area (see [5] for details). In addition, for completion purposes, 50,000 more reviews were taken from the Yelp Dataset Challenge[3], 320,000 restaurant reviews overall. From this dataset, a set of 1,738 sentences that named another restaurant (hence they potentially contain comparisons) was extracted. They were manually analyzed by the first author and 1,380 of them were identified as comparative sentences.

Comparative Relations

Previous work [14,15] showed how to identify comparative sentences accurately. In the scope of this research, sentences appearing in a review of a restaurant that contained another restaurant's name together with a comparative keyword were considered to be comparative sentences. To find a second restaurant name, as a preliminary requirement, we needed a database of reviews where each review would be marked with the name of the restaurant the review was written about. We extracted all the names from the reviews and saved them into a list that used to identify a second restaurant name in a sentence. To extract comparison relations from comparison sentences, we needed to identify the entities in the sentence, namely the restaurants, the attributes and the comparative keywords. Following Xu et al. [42], for automatic identification of the attributes in comparative sentences, we manually collected the attributes from our dataset and saved them into a restaurants ontology that contained 242 attributes. In addition, we built a comparative keyword list, with comparative keywords collected from available lexical resources [27,6,7] that contained 143 comparative keywords and phrases. We do not claim that our lists are complete, but we do believe that we have a sizable number of indicative words and phrases.

To discover the direction of the comparative relations, we implemented the CRF technique used by Xu et al. [41]. It offers a natural formalism for exploiting the dependence structure among entities and directly model the conditional probability distribution of the output given the input, so they can exploit the rich and global features of the input without representing the dependencies in it [37]. Therefore, CRFs offer several advantages when modelling complicated and long-range dependencies in an intuitive way [42].

For POS tagging, we used the Stanford Parser [36, 23] that can automatically annotate with high accuracy.

The performance of the techniques we experimented with was evaluated using accuracy, precision, recall, and the F1 measure [35].

Visualization

To select the visualizations to be used in our framework, we examined a few known visualization techniques. We selected those that seemed to match our needs, consulted with a visualization expert about them and performed a user study where we used the SUS questionnaire with additional open and closed questions to gather information about user experience and preferences with respect to these visualization techniques. This enabled us to select one technique that was preferred by the users, for presenting the comparisons from the customer reviews.

There are many attributes in the restaurant domain and we need to visually present them to a customer while at the same time provide a comprehensive overview of the reviews. Hence, we reduced and indexed the attributes into categories. Saad and Conway [31] cataloged the factors involved in

[2] www.yelp.com

[3] http://www.yelp.com/dataset_challenge

customer satisfaction in the restaurant industry into the following five categories:

1. Responsiveness (In this research called Service): Are the employees attentive, helpful and professional in responding to the customer's needs?

2. Food quality/reliability: Is the food fresh, tasty and agrees with the customer, i.e., type of food (Italian, Indian, etc.), food properties (drinks, lunch, dessert, etc.)?

3. Design and appearance: This relates to the restaurant lighting, available parking, location, procedures, features, cleanliness, decoration and the general environment.

4. Price: How much has the customer paid for his food and items?

5. Satisfaction: What is the customer's overall satisfaction and experience?

We used the above categories to help our users focus on a limited number of attributes presented simultaneously.

VISUALIZING SUMMARISED COMPARISONS

We developed a framework for extracting, processing and integrating comparison sentences drawn from customer reviews and then presenting a summary of the comparisons in an intuitive, visual way.

Figure 5: The main stages of the abstracted framework

As presented in Figure 5, the framework has three main components: Comparison sentence elicitation, data aggregation and visualization. The first step is to read the data and extract the comparative relations out of it, then aggregate and summarize the comparisons and reach a conclusion, and finally, present the results in an intuitive visualization. The following sections describe the implementation of a prototype that follows the suggested framework by:

- Compiling training datasets and extracting patterns and lexicons from the datasets to be used by the methods for identifying comparative relations from the reviews.

- Implementing extraction methods to extract comparative relations from incoming sentences.

- Selecting a visualization technique and developing an interactive visual user interface.

Extracting Comparative Relations

For automatic identification of comparative relations, we used the set of patterns extracted from our dataset, as described above. Following Xu et al. [42] method, the restaurants names, the attributes that are compared in the relations, and the comparative keyword in every sentence were extracted and represented in the standard form as:

<direction> {<restaurant1>, <srestaurant2>, <attribute>, <comparative keyword>}

Each comparative sentence was manually tagged with one of the four directions: >, <, = and ~, when > indicates that the first restaurant is better than the second restaurant, < worse, = equal and ~ has no gradable comparison. We used the standard forms of patterns identified using our training set and the pre-defined lists of entities (restaurants names, attributes, relations) to identify comparison sentences, comparison relations and relations directions in our test set.

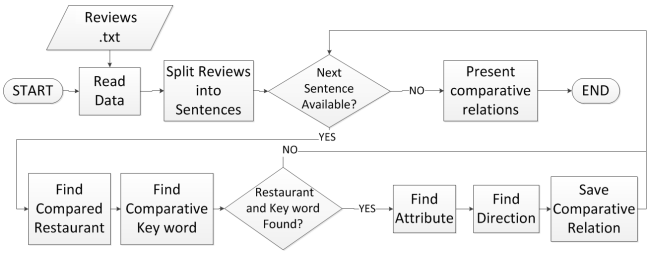

Figure 6: Workflow for implementing the framework

Figure 6 shows the workflow for the framework implementation. The analysis process starts by searching for the entities: the name of a second restaurant. Then finding a comparative keyword and the compared attribute from the predefined lists. In sentences that contain multiple entities, to verify which entity is right for the comparative relation we adopted the simple "closest first" approach, following Xu et al. [42]. The closest comparative keyword to the compared restaurant name in the sentence was selected, and following this, the closest attribute to the comparative keyword was selected. To find the direction of the comparative relation out of the four (>, <, =, ~) we first tried to use our own modified CSR method, hoping to receive better results than related works like the work of Xu et al. [42]. We used the POS tagging together with the comparative keywords to create CSR sequences: for every sentence from the dataset, we isolated the words around the comparative keywords by an optimal radius of words after evaluating the different options. As presented in Table 3 below, radius of three words for the new sentences combined with a radius of four words for the training patterns produced the best results in our experiment. To reduce the number of patterns, the POS tag of each word was used instead of the word itself (beside the keyword, which is left as is). Last, we matched a direction as a class to the generated sequence based on the direction manually defined earlier, we used four classes for the four directions (>, <, =, ~ as presented in [42]).

Unfortunately, in our experiment, the existing CRF method produced slightly better results than our modified CSR method (see table 4 for the CRF results). Accordingly, CRF as presented in [42], is the preferred method to identify the directions of comparative relations.

Table 3: The results using different radii to create patterns for the CSR method

		POS radii for the training patterns			
	F1	3	4	5	6
POS radii new sentences	2	60.21	59.17	59.55	58.77
	3	63.03	63.63	60.89	62.07

For selecting possible information visualization methods to present the comparative sentence to the customer, we consulted with a visualization expert on how to present the comparative relations that were extracted from the reviews as a preliminary stage in selecting the visualization techniques. As mentioned above, Shamim et al. [33] categorized the known visualization techniques in OM into four main categories: Radial visualization, graphs, hierarchical visualization and bar charts. From these categories, we selected six known visualization techniques that seemed to meet most of our needs. The six visualizations considered were: bar graph, spider graph, pie chart, tree map, table and comparative relation map.

The conclusions from the interview were to experiment with three visualization methods—bar graph, table and spider graph—and to compare only three restaurants in each visualization. In addition, for the table to be more precise and abundant, it was decided that each cell in the visualization table would be filled with the percentage and colors reflecting the restaurant ratings. In contrast to the expert's recommendation to consider testing time and performance measurements, we decided to focus only on usability since the visualizations were intuitive and the users were expected to be familiar with them. The visualizations were evaluated in a web-based user study, where we asked the participants a few informative questions about some information that was presented by the different visualizations (it was a simple comparative task) to allow them to reason about the visualizations. Then the participants used SUS to evaluate different visualizations. This process was repeated for all three visualizations. Last, the participants were asked which method he or she most preferred and why.

The experimental prototype

For experimental purposes, a Restaurant Comparison Presentation tool was developed, in order to enable us to assess the restaurant comparison visualization is a realistic setting. The tool contains three main sections: A Visualization window, a Data Grid (with the comparisons) and a Setup section that enables the user to choose how he or she wants to see the comparisons. The user can select attributes, restaurants for comparison, and a visualization method. Three visualization methods were used: bar graph, spider graph and table (see Figures 7, 8 and 9).

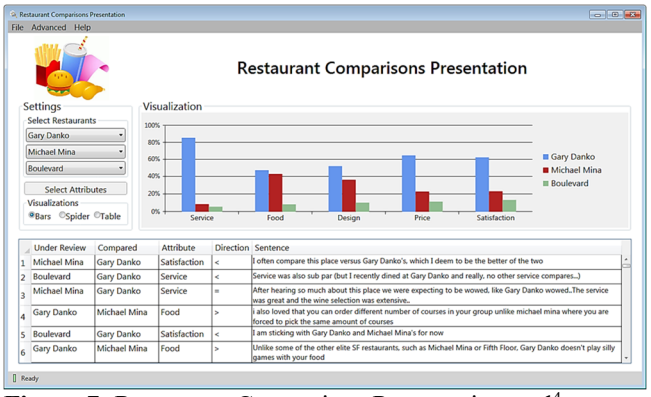

Figure 7: Restaurant Comparison Presentation tool[4]

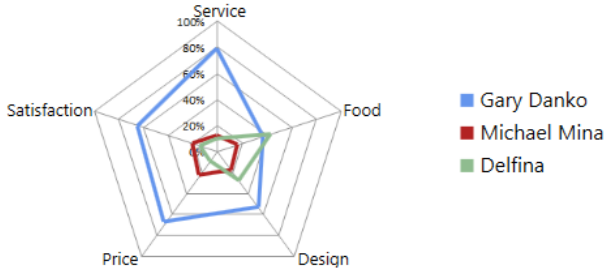

Figure 8: Spider graph visualization

	Service	Food	Design	Price	Satisfaction
Gary Danko	78%	39%	53%	67%	65%
Michael Mina	13%	17%	18%	24%	22%
Delfina	9%	44%	29%	9%	13%

Best
Middle
Worst

Figure 9: Table based visualization

EXPERIMENTAL SETUP AND EXPERIMENTAL RESULTS

To evaluate our framework, new comparative sentences were extracted from the dataset of reviews about restaurants in the San Francisco area. These were used as a test set and a gold standard. To build the test set, we randomly selected 10 restaurants and manually analyzed their reviews. The entities and the directions of the relations were manually extracted from each sentence. In this way, 456 comparative sentences were ready for the testing. Using the test set, we executed our implemented methods and compared the results achieved.

Table 4: The results of using CRF to find the direction of comparative sentences

Direction	Recall	Precision	% in Dataset	F1
>	40.00	40.51	17.58	40.25
<	75.57	73.88	57.58	74.72
=	68.57	72.73	23.08	70.59
~	50.00	44.44	1.76	47.06
Mean:	67.25	67.23	100.00	67.22

[4] available for downloading at: http://doi.org/10.5281/zenodo.1111755

Results of Comparative Relation Extraction

First, we evaluated the ability of CRF to identify comparative relations and the direction of the relations. It had an overall success of 67.22% (See table 4 for results). Next, given the fact that our list of attributes in comparative sentences is finite and most likely does not contain all the attributes available in the domain, we measured how well the attributes were found in the new dataset. We automatically extracted the attributes from the test set using our attributes list and compared it to the ones manually found it the new dataset. We used the accuracy parameter to evaluate the results. The right attribute was successfully found in the comparative sentences with an accuracy of 82%.

In addition, multiple attributes occasionally appear in the sentences while only one should match the comparative relation. By experimenting, we verified that it is better to use the attribute closest to the comparative keyword and not the one closest to the second restaurant name. We found that searching for the attributes closest to the comparative keyword in the sentence, correctly identified 82% of the attributes compared with 77% attributes that were closest to the second restaurant name.

Another experiment regarding the "inverted" comparative sentence (not, unlike, etc.) was also conducted. Such words invert the direction of the relation in the sentence. We tested all radii from zero (not searching) to six words before the comparative keyword to search for such inverted words. A distance of four words before the keywords while searching for "inverting" words worked best in our settings. Table 5 presents the final F-score results for finding comparative direction using the different distances while searching for "inverting" words:

Table 5: F-score results of finding the direction of the relation using the different radii to search for "inverted" comparisons (not, unlike, etc.)

Radius	**0**	**1**	**2**	**3**	**4**	**5**	**6**
F1	64.86	65.34	65.54	65.76	**67.22**	66.71	66.16

The Selected Visualization Technique

To evaluate and decide which visualization technique to use for presenting the comparisons of customer reviews, three visualization techniques were evaluated using an online questionnaire based on SUS. To date, 117 participants (recruited via social media) have responded to the questionnaire: 38% men and 62% women from different demographic layers. The final SUS results (from 0 to 100) of the questionnaire are presented in Table 6.

The bar graph is the most favored method, receiving a total SUS score of 82, followed by the table, a close second with 80, and then by the spider graph, which only received a score of 49.

Table 6: The results of the SUS questionnaire evaluating the visualization methods.

	Average SUS Score	**Median**	**STD**
Bar Graph	82	87	20
Table	80	87	19
Spider	49	47	24

Although the statistical ANOVA test between the table and the bar graph showed—with a confidence level of 95%--that there is no significant difference between the table and the bar graph, 65% of the volunteers declared that bar graph is their favorite visualization technique, 28% voted for the table and only 7% voted for the spider. Most volunteers declared the bar graph is the most "common", "clear", "simple", "eye pleasing", "intuitive", etc. technique.

Evaluating the Visualization of Restaurant Comparisons

To evaluate and examine the performance and usability of the implemented framework, we used the Restaurant Comparison Presentation prototype. 15 users tested the system functionality and answered the SUS questionnaire. The average SUS score was 86, the median was 90 and the standard deviation was 13, which show that overall, the users were highly satisfied with the system.

DISCUSSION

This study aimed to provide customers with a quick, visual overview of a large body of restaurant reviews. This goal was achieved by visualizing the results of uncovering and analyzing comparison sentences appearing in the reviews. As the suggested approach is generic, the same approach may be extended to additional domains besides the restaurant domain, which when applied, alleviates the information overload facing consumers today when looking at service and product reviews.

In this study, we followed Xu et al. [42] that, in their research, addressed sentences as either comparative or not comparative and assumed all comparative sentences have a clear direction, although they did not address "non-gradable" sentences such as "*It's different than Beretta*". In such sentences, though a comparison is being made, there is no clear direction. Therefore, in our research, we used the same directions as in Xu et al. [42] and considered the not comparative (~) direction as a non-gradable comparative. We found that 3.10% of the sentences in our data set were "non-gradable" sentences.

When comparing our results to the results obtained by Xu et al. [41], it can be noticed that by applying the same approach we achieved better results in general, but the results for the specific relations were quite different. This is an interesting issue for further analysis. In identifying the direction of the comparative relations, identifying "better" and "not equal" relations was less successful compared with "worse" and

"equal" relations. We suspect that this is due to reviewers' tendencies to use "worse" and "equal" relations more in their reviews, leading these to be selected more often as the right direction for the relation, thereby causing the skew among the results.

All in all, we found that the users appreciated the idea of using simple, visualized summaries of comparisons of features of interest. Hence it seems that this is a direction for future research - better automation and experimentation in additional domains.

Regarding visualization, it is worth noting that there are several good approaches by commercial systems that do summarize restaurant reviews and provide them to consumers with some visualization aspects. For example, Yelp highlights excerpts with bold keywords. We go a step further and in addition to focusing on comparison sentences, we also provide visual representations of the opinions regarding the selected features, which is much more informative than highlighting keywords.

This study had a few limitations. We found out that 9% of the comparison sentences in our dataset do not contain comparative keywords, unlike previous studies [14,15,42], which asserted that almost all their comparison sentences contained comparative keywords (98% but with low precision). Hence, without comparative keywords, we were unable to predict the direction of the relation; comparative sentences without a comparative keyword were discarded. For example: "*Sorta wish I went to their sister restaurant, SPQR.*"; "*It's no Gary Danko's but they know their fish.*" and more. Another limitation relates to the restaurants' names, which can contain an attribute as well as mentioned in the sentence with a nickname. For example, Lime, Americano is also a restaurant and also a food or attribute. Sometimes, restaurants have nicknames; for example, Anchor and Core are sometimes called A&C, or Foreign Cinema is FC, etc. Nicknames that were not recognized by the restaurant lists were ignored as well. Another limitation was found vis-à-vis synonyms. For example, the comparative keyword 'like' in some cases is not a comparative keyword and has a different meaning. It can be used as a synonym for 'love' and not for 'same'. The same is true for 'even', which can be similar to 'equal' or can be a 'contrast' word. One solution for such issues is using the comparative keyword with POS tagging and learning how to distinguish between meanings, but doing so exceeds the scope of this research and should be considered as future work.

Then, a limitation exists regarding the assumption that each sentence contained only one comparison relation in it. In a long comparative sentence, there can be multiple relations separated by delimiters such as commas (",") and conjunctions such as "and" and "but". Multi-comparisons and complex sentences were treated as if they contained only one comparison. For example, the comparison from the Credo restaurant review, "*Sorry Barbacco...Credo has much better food and wine!*" In this example, the reviewer describes two comparative relations in one sentence. One relation claims that Credo has better food than Barbacco and the other claims that Credo also has better wine. Alternatively, there are comparative sentences that do not contain any comparative keyword or attribute.

Regarding visualization, we selected a few well-known visualizations for experimentation, following an expert recommendation. It is possible that a more thorough evaluation of possible visualizations that can be used would have resulted in better results and better users' satisfaction. Related to that is the simplicity of the evaluation – we used SUS to elicit users' opinion about the usability of the system, while defining a specific task may have resulted in better understanding. We see these aspects as ideas for future work.

Finally, the evaluation was done with 15 subjects, a very small number of participant that limits the significance of the evaluation. Again, this is an issue for future research that will extend the current study and address the limitations noted above.

CONCLUSIONS AND FUTURE WORK

In this study, we suggested, demonstrated and evaluated a generic framework for visualizing comparisons of restaurants with respect to the features the user is interested in, in order to allow the user to better assess the quality of a given restaurant. We hypothesized that visualized summaries will be perceived informative and easy to use and this hypothesis was confirmed in a user study. The suggested framework is not limited to restaurants review, but can be applied in any domain where comparisons of features are used in reviews.

There are a few directions that could be examined in future work. First, the various vocabularies used in this research were pre-compiled and dedicated to the restaurant domain. To make the framework more general, and to use different domains, new attribute lists need to be compiled for each domain. This requirement makes the system quite difficult to adapt. It may be worthwhile considering Liu's [20] Label Sequence Rules (LSR) method to find the entities in a sentence. The method might be adapted to create a more global framework that does not require entering fixed attributes and does not depend on the domain.

Another idea for future work is to explore the sentiments in the sentences [30] and enhance the methods for finding the direction using these sentiments as another indicator to support the decision about the correct comparative direction.

Finally, to improve the visualization method, instead of using an existing one, a new method, dedicated specifically to presenting comparative sentences, should be developed. The classic methods are excellent, and together with a data grid, present comparisons in an intuitive way, but maybe there is another method, as yet unknown in the field, which might be even better. A more thorough evaluation (like performing a

search and evaluation task) may help understand better the pros and cons of the different visualizations. As part of this extension, a large-scale user study will be carried out.

REFERENCES

1. Arora, S., Joshi, M., & Rosé, C. P. (2009). Identifying types of claims in online customer reviews. In Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers, 37-40. ACL.
2. Bai, X., White, D., & Sundaram, D. (2011). Purposeful visualization. In System Sciences (HICSS), 2011 44th Hawaii International Conference on (pp. 1-10). IEEE.
3. Brooke, J. (1996). SUS-A quick and dirty usability scale. Usability Evaluation in Industry, 189(194), 4-7.
4. Cruz, F. L., Troyano, J. A., Enríquez, F., Ortega, F. J., & Vallejo, C. G. (2013). 'Long autonomy or long delay?' The importance of domain in opinion mining. Expert Systems with Applications, 40(8), 3174-3184.
5. Removed for annoymization.
6. Esuli, A., & Sebastiani, F. (2006). Sentiwordnet: A publicly available lexical resource for opinion mining. In Proceedings of Language Resources and Evaluation Conference, Vol. 6, 417-422.
7. Fei, Z., Huang, X., & Wu, L. (2006). Mining the relation between sentiment expression and target using dependency of words. In Proceedings of the 20th Pacific Asia Conference on Language, Information and Computation (PACLIC20), Wuhan, China, 257-264.
8. Freitas, C. M., Luzzardi, P. R., Cava, R. A., Winckler, M., Pimenta, M. S., & Nedel, L. P. (2002). On evaluating information visualization techniques. In Proceedings of the Working Conference on Advanced Visual Interfaces, 373-374. ACM.
9. Gamon, M., Aue, A., Corston-Oliver, S., & Ringger, E. (2005). Pulse: Mining customer opinions from free text. In Advances in Intelligent Data Analysis VI, 121-132. Springer Berlin Heidelberg.
10. Gu, Y. H., & Yoo, S. J. (2010). Searching a best product based on mining comparison sentences. In SCIS & ISIS SCIS & ISIS 2010, 929-933. Japan Society for Fuzzy Theory and Intelligent Informatics.
11. He, S., Yuan, F., & Wang, Y. (2012, April). Extracting the comparative relations for mobile reviews. In Consumer Electronics, Communications and Networks (CECNet), 2012 2nd International Conference on, 3247-3250). IEEE.
12. Hu, M., & Liu, B. (2004, August). Mining and summarizing customer reviews. In Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (168-177). ACM.
13. Hu, N., Liu, L., & Zhang, J. J. (2008). Do online reviews affect product sales? The role of reviewer characteristics and temporal effects. Information Technology and Management, 9(3), 201-214.
14. Jindal, N., & Liu, B. (2006b). Identifying comparative sentences in text documents. In Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 244-251). ACM.
15. Jindal, N., & Liu, B. (2006a). Mining comparative sentences and relations. In Association for the Advancement of Artificial Intelligence, Vol. 22, 1331-1336.
16. Kessler, J. S., Eckert, M., Clark, L., & Nicolov, N. (2010). The ICWSM 2010 JDPA sentiment corpus for the automotive domain. In 4th International AAAI Conf. on Weblogs and Social Media Data Workshop Challenge (ICWSM-DWC), Washington, DC.
17. Kessler, W., & Kuhn, J. (2013). Detection of product comparisons-how far does an out-of-the-box semantic role labeling system take you?. In EMNLP 1892-1897.
18. Khan, M., & Khan, S. S. (2011). Data and information visualization methods, and interactive mechanisms: A survey. International Journal of Computer Applications, 34(1), 1-14.
19. Lafferty, J., McCallum, A., & Pereira, F. C. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In Proceedings of the 18th International Conference on Machine Learning 2001 (ICML 2001) 31-36. ACM.
20. Liu, B. (2007). Web data mining: exploring hyperlinks, contents, and usage data. Springer Science & Business Media.
21. Liu, B., & Zhang, L. (2012). A survey of opinion mining and sentiment analysis. In Mining Text Data (pp. 415-463). Springer US.
22. Lovelock, C. H. (1983). Classifying services to gain strategic marketing insights. Journal of Marketing, 9-20.
23. Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J. R., Bethard, S., & McClosky, D. (2014). The Stanford CoreNLP Natural Language Processing Toolkit. In ACL (System Demonstrations) 55-60.
24. Marrese-Taylor, E., Velásquez, J. D., & Bravo-Marquez, F. (2014). A novel deterministic approach for aspect-based opinion mining in tourism products reviews. Expert Systems with Applications, 41(17), 7764-7775.
25. Mayhew, D. J. (1991). Principles and guidelines in software user interface design. Prentice-Hall, Inc.

26. Miao, Q., Li, Q., & Dai, R. (2009). AMAZING: A sentiment mining and retrieval system. Expert Systems with Applications, 36(3), 7192-7198.
27. Miller, G. A. (1995). WordNet: A lexical database for English. Communications of the ACM, 38(11), 39-41.
28. Mudambi, S. M., & Schuff, D. (2010). What makes a helpful review? A study of customer reviews on Amazon.com. MIS Quarterly, 34(1), 185-200.
29. Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval, 2(1-2), 1-135.
30. Ravi, K., & Ravi, V. (2015). A survey on opinion mining and sentiment analysis: Tasks, approaches and applications. Knowledge-Based Systems, 89, 14-46.
31. Saad Andaleeb, S., & Conway, C. (2006). Customer satisfaction in the restaurant industry: an examination of the transaction-specific model. Journal of Services Marketing, 20(1), 3-11.
32. Sha, F., & Pereira, F. (2003). Shallow parsing with conditional random fields. In Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Vol. 1, 134-141). ACL.
33. Shamim, A., Balakrishnan, V., & Tahir, M. (2014). Evaluation of opinion visualization techniques. Information Visualization, 147387/1614550537.
34. Smith, S. L., & Mosier, J. N. (1984). Design guidelines for user-system interface software (No. MTR-9420). MITRE CORP BEDFORD MA.
35. Sokolova, M., Japkowicz, N., & Szpakowicz, S. (2006). Beyond accuracy, F-score and ROC: A family of discriminant measures for performance evaluation. In Australasian Joint Conference on Artificial Intelligence, 1015-1021. Springer Berlin Heidelberg.
36. Stanford Natural Language Processing Group, (2009). The Stanford Parser: A statistical parser, http://nlp.stanford.edu/software/lex-parser.shtml.
37. Sutton, C., & McCallum, A. (2006). An introduction to conditional random fields for relational learning. Introduction to statistical relational learning, 93-128.
38. Turney, P. D. (2002). Thumbs up or thumbs down?: Semantic orientation applied to unsupervised classification of reviews. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, 417-424. ACL.
39. Von Alan, R. H., March, S. T., Park, J., & Ram, S. (2004) Design science in information systems research. MIS Quarterly, 28(1), 75-105.
40. Wang, W., Zhao, T., Xin, G., & Xu, Y. (2014). Recognizing Comparative Sentences from Chinese Review Texts. International Journal of Database Theory and Application, 7(5), 29-38.
41. Ware, C. (2012). Information visualization: Perception for design. Elsevier.
42. Xu, K., Liao, S. S., Li, J., & Song, Y. (2011). Mining comparative opinions from customer reviews for Competitive Intelligence. Decision support systems, 50(4), 743-754.
43. Ying, D., & Jiang, J. (2015). Towards opinion summarization from online forums. ACL.
44. Zhang, K., Narayanan, R., & Choudhary, A. (2009). Mining online customer reviews for ranking products. Technical Report, EECS Dept, Northwestern Univ.
45. Zhang, K., Narayanan, R., & Choudhary, A. (2010, June). Voice of the customers: Mining online customer reviews for product feature-based ranking. In Proceedings of the 3rd Conference on Online Social Networks, 11.
46. Zhang, L., & Liu, B. (2014). Aspect and entity extraction for opinion mining. In Data Mining and Knowledge Discovery for Big Data, 1-40. Springer Berlin Heidelberg.