OpinionZoom, a modular tool to explore tourism opinions on the Web

Edison Marrese-Taylor
Department of Industrial Engineering
University of Chile
Santiago, Chile
emarrese@wi.dii.uchile.cl

Juan D. Velásquez
Department of Industrial Engineering
University of Chile
Santiago, Chile
jvelasqu@dii.uchile.cl

Felipe Bravo-Marquez
Department of Computer Science
University of Chile
Santiago, Chile
fbravo@dcc.uchile.cl

Abstract—In this paper, we propose OpinionZoom, a modular software that helps users in an easy manner to understand the vast amount of tourism opinions disposed all over the Web. We also successfully implemented and tested OpinionZoom, encompassing the situation of the tourism industry in Los Lagos, also known as the Lake District, in Chile. Results showed the effectiveness of the designed proposal when applied to solving this specific industry's issues.

Keywords-opinion mining; tourism product reviews; sentiment analysis; aspect-based

I. Introduction

Nowadays, tourists check opinions and experiences published by other travelers on different web platforms when planning their own vacations. Travelers from all around the world share their experiences on websites like TripAdvisor, LonelyPlanet and Booking, which have become major sources of information for tourists.

When trying to analyze web opinions, tourists are often overwhelmed with information due to the amount of available opinionated text. Considering this problem, we propose OpinionZoom, a tool that offers a set of summarization methods to help users digest the vast availability of opinions in an easy manner. We also use our design to implement a prototype to analyze opinions from TripAdvisor in the context of the tourism industry in Los Lagos, a Chilean administrative region also known as the Lake District. Our system was intended to help users understand the attitude and the overall appreciation of web users in the tourism domain by easily finding and extracting relevant subjective information from customer reviews published in TripAdvisor.

II. Related Work

Opinion mining or sentiment analysis comprises an area of NLP, computational linguistics and text mining, and refers to a set of techniques that deals with data about opinions and tries to obtain valuable information from them. As stated in [1], literature offers two main approaches, aspect-based and non-aspect-based opinion mining. Aspect-based opinion mining techniques divide input texts into *aspects*, which are components or attributes of specific entities that are discussed in the opinionated documents. In the case of tourism product reviews, these entities correspond to the product that is being evaluated. The process usually includes the following steps: (1) Aspect identification, to find and extract important topics in the text that will then be used to summarize, (2) Sentiment Prediction, to determine the sentiment orientation on each aspect, and (3) Summary Generation, to present the processed results in a simple manner.

Opinion mining has attracted the attention of many research fields and many applications exist so far. A considerable number of of these applications consider Twitter as a source of opinionated documents, such as *Sentiment 140*[1] and *TweetFeel*[2]. On the other hand, *Socialmention*[3] offers a social media search and analysis platform that aggregates user-generated content from different social media sources. Our approach is different from all these applications since it is aspect-based and analyzes opinions at the sentence level.

In addition, there are a significant number of applications that mine sources that contain product reviews, such as the mentioned TripAdvisor and VirtualTourist (for tourism products) or Amazon and C|Net. Examples of these applications are Lexalytics Salience Engine[4] and Nebular [2]. These applications process opinionated documents and generally offer text summaries as output, lacking other visualization methods. In this area, OpinionZoom is different since it offers novel and intuitive graphic summaries of opinions. These summaries are intended to provide users a way of processing the vast amount of information available in social media about tourism products. Finally, we also found OpinionEQ[5], which offers an approach that seems very similar to ours. However, OpinionEQ is not proposed as an application but rather as a service.

III. System Architecture

OpinionZoom was designed using a modular programming paradigm. Figure 1 shows the proposed architecture.
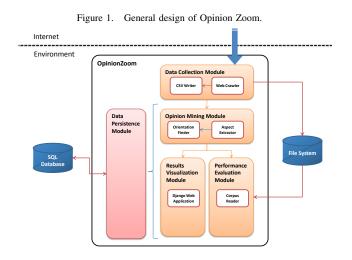
[1]http://twittersentiment.appspot.com
[2]http://www.tweetfeel.com
[3]http://www.socialmention.com
[4]http://library.lexalytics.com/content/opinion_mining
[5]http://www.opinioneq.com

IEEE
computer
society

The main functionalities are described in the following paragraphs.

Figure 1. General design of Opinion Zoom.

The Data Collection Module (DCM) is in charge of obtaining opinions from a set of given web sources. This module simply consists of a set of web crawlers, which must be source-specific. The crawlers parse HTML webpages containing opinions and pre-process the results, generating comma-separated CSV files containing the downloaded opinionated documents.

The Opinion Mining Module (OMM) implements aspect-based opinion mining algorithms on a given set of opinionated documents. Each opinionated document is separated into sentences, which are then split into tokens; POS tagging and syntactic chunking methods are then applied. Two different tasks need to be performed, *aspect* extraction and orientation determination, for which two sub-modules are included:

- Aspect Extraction Sub-Module: in charge of applying the *aspect* extraction algorithm to a set of POS-tagged sentences. This algorithm is based on [3], which uses the most frequent nouns and NP's to extract aspects.
- Orientation Determination Sub-Module: This sub-module applies an algorithm to determine the orientation of an opinion on a given aspect. This algorithm is taken from [4], an extension of Liu's proposals, consisting of a lexical and a rule-based approach. In the first place, the orientation of relevant words is obtained using an opinion lexicon, i.e. a list of words with known orientations or *opinion words*, and a set of linguistic rules, using common linguistic patterns such as negations. Then, we combine these orientations to determine the final orientation of each aspect in a sentence using algorithm 1. This algorithm computes the orientation of an aspect, represented as a set of nouns in a sentence, using the orientation of nearby opinion words; the nearer the word is, the more influence it has.

The sub-module also extracts the set of adjectives that appeared near each aspect.

Algorithm 1 Opinion Orientation

```
1:  if but_word is in sentence then
2:      orientation ← Opinion Orientation(aspect,marked_words,but_clause)
3:      if orientation ≠ 0 then
4:          return  orientation
5:      else
6:          orientation ← Opinion Orientation(aspect,marked_words,not but_clause)
7:          if orientation ≠ 0 then
8:              return  -1 × orientation
9:          else
10:             return  0
11:         end if
12:     end if
13: else
14:     for all aspect_position in aspect do
15:         for all aspect_word in aspect_position do
16:             for all word in marked_words do
17:                 suborientation += Word Orientation(word) / Word Distance(aspect_word,word)
18:             end for
19:             orientation += suborientation
20:         end for
21:         final_orientation += orientation
22:     end for
23:     if final_orientation > 0 then
24:         return  1
25:     else
26:         if final_orientation < 0 then
27:             return  -1
28:         else
29:             return  0
30:         end if
31:     end if
32: end if
```

The Results Visualization Module (RVM) is the visible portion of the application and interacts directly with the user. Users can give opinion data to the system which can then be used to apply the opinion mining process. Results include the following features:

- Aspect-Based summaries: Bar charts, in which each bar measures the number of positive and negative mentions of each attribute or component of one product. Bars are ordered according to the relative importance, which is computed supposing that an aspect that has a lot of positive and negative opinions is more likely to be relevant to the user. To capture this, for each aspect i, we calculate the standard deviation of the number of positive and negative opinions ($PScore_i$ and $NScore_i$ respectively), according to the following formulas. We apply min-max normalization to these values and obtain the *relative importance*.

$$AVScore_i = \frac{PScore_i + NScore_i}{2}$$

$$STDScore_i = \sqrt{\frac{(PScore_i - AvScore_i)^2 + (NScore_i - AvScore_i)^2}{2}}$$

- Adjective bubble charts: Nearby adjectives in all sentences where an *aspect* appears are shown in a bubble chart. The size of each bubble counts the times that each adjective is used to describe the aspect.
- Original opinions: A list of all original sentences is also displayed in an ad-hoc manner, separating them into positive or negative.

The system also provides a tagging interface that helps users to extract opinions from the opinionated documents and alter the algorithm's results. This functionality appears in a special menu that does not interfere with the rest of the specifications. In addition, after applying the opinion mining algorithms, OpinionZoom offers an interface that lets users see the list of the extracted aspect and select the ones that he really wishes to save. We included these two functionalities to receive relevance feedback from our users. Thus, choices and operations performed by users are stored and then used to improve the system performance.

The Performance Evaluation Module (PEM) is in charge of delivering a set of indexes that evaluates the performance of the opinion mining algorithms. In order to do this, the system allows users to elaborate and then provide specially annotated corpora, following the structure that appears in figure 2. To facilitate the annotation process, guidelines and examples are also offered. As a result, three tasks can be evaluated by comparing the extraction process results with the provided corpora: (1) Explicit aspect extraction, to measure the effectiveness of the explicit aspect extraction algorithm, (2) Subjectivity classification, to evaluate the effectiveness of opinion sentence extraction and (3) Sentiment classification, to measure the accuracy of the orientation prediction of each aspect in each sentence. We believe that the service provided by this module is crucial when trying to understand the usefulness of the system within a particular topic or domain. To the best of our knowledge, this represents an important difference between OpinionZoom and other existing tools.

Figure 2.   Example of sentences of an annotated corpora.

```
line
 1  [c1][s1] place[+], comfort[+][u], location[+][u] ### good place to stay at
    the end of a long flight in that it is very comfortable, with many facilities,
    and in the town , by the shore .
 2  [c1][s2] ### however, puerto montt is not the best of places to explore.
 3  [c1][s3] ### a better place is puerto varas which is just as near to the
    airport and has far more attractions.
 4  [c1][s4] ### could not fault this aparthotel.
 5  [c2][s1] hotel[+] ### my fiance and i spent three nights here in march
    2012 and it's a sweet, quaint hotel.
 6  [c2][s2] reservation[-] ### with that said, i called in early february 2012
    to make a reservation and it got lost/misplaced.
```

Finally, the Data Persistence Module or DPM manages all the database operations and constitutes a model layer for the whole system. The data layer is implemented using two relational models, which support all the data that needs to be stored.

IV. INDUSTRY APPLICATION

In this section, we show a real case application where the proposed design was implemented using Python. The application encompasses the situation in the Lake District, where tourism operators lack tools to understand what their

customers want or need. Our study used the *NLTK*[6] libraries for the NLP tasks in the OMM and the *Django Framework*[7] for the RVM, and included the stages that are explained below.

In the first place, we implemented a web crawler and downloaded all the reviews from hotels and restaurants originally written in English about the *Lake District* in TripAdvisor. We obtained a total of 1,435 reviews and saved them in two different CSV files, as defined in the design of the DCM. Then, we generated annotated corpora or datasets to evaluate the performance of the algorithms for the selected products by randomly selecting 100 restaurant and hotel reviews and following the provided guidelines for tagging. In both corpora, almost 80% of the sentences contained opinions, somewhat validating the use of TripAdvisor as a source of opinions for these tourism products. Table I gives details about the aspects that were manually extracted. Following our notation, aspects that appear in the manner of nouns or NPs in a sentence are called explicit aspects, while all other kinds of aspects are called implicit aspects. Results show that in both corpora, explicit aspects are the most common ones.

We implemented all the PEM specifications and then evaluated how the proposed opinion mining algorithms perform when applied to tourism product reviews using our corpora. Average results for the three tasks defined in the design are presented in table II.

Table I
DETAIL ON *aspects* FOUND IN CORPORA.

Aspect Type	Hotels Corpus		Restaurants Corpus	
	Number	Percentage	Number	Percentage
Explicit	229	73.87%	161	67.93%
Explicit and Implicit	30	9.68%	26	10.97%
Implicit	51	16.45%	50	21.1%
Total	310	100%	237	100%

Table II
OMM ALGORITHMS PERFORMANCE RESULTS.

Index Name	Average Performance		
	Precision	Recall	F-measure
Explicit Aspect Extraction	38%	33%	36%
Subjectivity Classification	80%	91%	85%
Sentiment Classification	90%	93%	92%

These results show that performance on the aspect extraction task is fairly poor, the algorithm only being capable of extracting nearly 35% of the total explicit expressions. Moreover, a high percentage of the extracted expressions do not correspond to real aspects. On the other hand, the system labels some non-opinionated sentences as opinion sentences because they contain both aspects and some *opinion words*, causing precision to decrease. Finally, sentiment classification shows fairly good results.
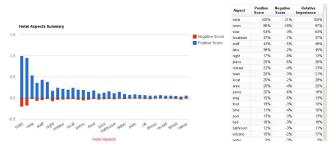
[6]http://nltk.org
[7]https://www.djangoproject.com

Since we intend to show the extracted aspects to users, we also consider the problem of aspect extraction from the perspective of Information Retrieval and measure *precision at k* of the extracted aspects, according to their relative importance. As shown in table III, results prove that this measure ensures that users see aspects with minimum noise. Since the task of sentiment classification has a fairly good performance, we then have empirical evidence that bar charts showing the top k aspects deliver accurate and true information to users.

Table III
PRECISION AT K FIRST ASPECTS, ACCORDING TO RELATIVE IMPORTANCE.

Precision at	Hotels Corpus	Restaurants Corpus	Average
10	100%	90%	95%
15	73,33%	93,3%	83%
20	75%	90%	83%

We implemented the RVM as a web-based application adding all the features mentioned in the design. As figure 3 shows, besides a bar chart a table shows the actual values of the $PScore$, $NScore$ and relative importance. By clicking the name of each column, the table and the bar chart are sorted according to the clicked column (each click alternates between an ascending or descending sort.) Below the chart and the table, a list with all the aspects is shown.
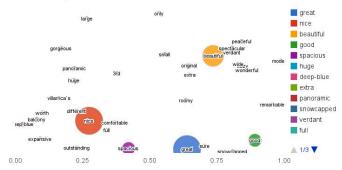
Figure 3. Bar chart for hotels in lake District. Aspects are ordered according to relative importance in a descending manner.

By clicking one item on the list, the user is redirected to a page showing the corresponding adjective bubble charts, which are built using the two nearest adjectives. In figure 4, the chart really offers valuable information, indicating that tourists in TripAdvisor tend to describe the lake view using strong positive adjectives, such as nice, great and beautiful. As mentioned before, the RVM also offers users an interface to select the aspects to be saved. In our case, considering the obtained performance of the aspect extraction task, this functionality became crucial.

Finally, since we wanted to know if the system is able to solve the proposed problem, we interviewed and surveyed a group of 27 tourism operators, who navigated through the charts as disposed on the website www.patagonialoslagos.cl. In relation to bar charts, 45% of the users completely understood the meaning of each bar without any additional explanation, while for bubble charts most of the users needed

Figure 4. Proposed bubble charts for the aspect *lake view* of hotels.

help understanding the meaning of the size of each bubble. In general, the charts were difficult to understand mostly for those users that were less familiar with technology or for those that had problems with English. However, results showed that most of the users (almost 80%) considered that the system adds valuable information to their business.

V. CONCLUSION

In this study, we presented a generic design of an tourism opinion mining system that aims to be useful in many industries. We also used our proposals to successfully implement the system and solve a specific problem in the *Lake District* tourism industry. The results obtained validate the proposed architecture and prove how useful and powerful our tool is. As for future work, we propose adding more opinion sources to the system and implementing different state-of-the-art aspect-based opinion extraction algorithms.

REFERENCES

[1] B. Liu, *Web data mining: exploring hyperlinks, contents, and usage data*. Springer Verlag, 2007.

[2] S. Palakvangsa-Na-Ayudhya, V. Sriarunrungreung, P. Thongprasan, and S. Porcharoen, "Nebular: A sentiment classification system for the tourism business," in *Computer Science and Software Engineering (JCSSE), 2011 Eighth International Joint Conference on*. IEEE, 2011, pp. 293–298.

[3] M. Hu and B. Liu, "Mining opinion features in customer reviews," in *Proceedings of the National Conference on Artificial Intelligence*. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2004, pp. 755–760.

[4] E. Marrese-Taylor, J. D. Velásquez, F. Bravo-Marquez, and Y. Matsuo, "Identifying customer preferences about tourism products using an aspect-based opinion mining approach," in *Knowledge-Based and Intelligent Information and Engineering Systems*. Procedia Computer Science, 2013.