Expert Systems with Applications 41 (2014) 7764–7775

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

A novel deterministic approach for aspect-based opinion mining in tourism products reviews

Edison Marrese-Taylor[a], Juan D. Velásquez[a,*], Felipe Bravo-Marquez[b]

[a] Department of Industrial Engineering, Universidad de Chile, Av. República 701, P.O. Box: 8370439, Santiago, Chile
[b] Department of Computer Science, The University of Waikato, Private Bag 3105, Hamilton 3240, New Zealand

ARTICLE INFO

Article history:
Available online 21 June 2014

Keywords:
Aspect-based
Opinion mining
Tourism
Product reviews

ABSTRACT

This work proposes an extension of Bing Liu's aspect-based opinion mining approach in order to apply it to the tourism domain. The extension concerns with the fact that users refer differently to different kinds of products when writing reviews on the Web. Since Liu's approach is focused on physical product reviews, it could not be directly applied to the tourism domain, which presents features that are not considered by the model. Through a detailed study of on-line tourism product reviews, we found these features and then model them in our extension, proposing the use of new and more complex NLP-based rules for the tasks of subjective and sentiment classification at the aspect-level. We also entail the task of opinion visualization and summarization and propose new methods to help users digest the vast availability of opinions in an easy manner. Our work also included the development of a generic architecture for an aspect-based opinion mining tool, which we then used to create a prototype and analyze opinions from TripAdvisor in the context of the tourism industry in Los Lagos, a Chilean administrative region also known as the Lake District. Results prove that our extension is able to perform better than Liu's model in the tourism domain, improving both Accuracy and Recall for the tasks of subjective and sentiment classification. Particularly, the approach is very effective in determining the sentiment orientation of opinions, achieving an *F*-measure of 92% for the task. However, on average, the algorithms were only capable of extracting 35% of the explicit *aspect expressions*, using a non-extended approach for this task. Finally, results also showed the effectiveness of our design when applied to solving the industry's specific issues in the Lake District, since almost 80% of the users that used our tool considered that our tool adds valuable information to their business.

1. Introduction

With the inception of the Web 2.0 and the explosive growth of social networks, enterprises and individuals are increasingly using the content in these media to make better decisions (Park & Kim, 2009; Zhu & Zhang, 2010). For instance, tourists check opinions and experiences published by other travelers on different Web platforms when planning their own vacations. On the other hand, for organizations, the vast amount of information available publicly on the Web could make polls, focus groups and some similar techniques an unnecessary requirement in market research.

However, due to the amount of available opinionated text, users are often overwhelmed with information when trying to analyze Web opinions. So far, many authors have tacked the problem of human limitation to process big amounts of information and extract consensus opinions (Velásquez & González, 2010) from a large number of sources relying on data-mining-based tools. Considering a similar problem, this work is an effort to create a tool that offers a set of summarization methods and help users digest in an easy manner the vast availability of opinions in the tourism domain. The core of our system is a novel extension of Bing Liu's aspect-based opinion mining methodology, which was developed by us in order to apply Liu's ideas to the tourism domain.

This extension is concerned with the fact that users refer differently to different kinds of products when writing reviews on the Web. Concretely, consider a *generic product*, which refers to the conceptual commodity produced by an industry (Smith, 1994). Most of the authors, including Kotler (2001), tend to classify these generic products using two categories, physical

* Corresponding author. Tel.: +56 2 2978 4834.
E-mail addresses: emarrese@wi.dii.uchile.cl (E. Marrese-Taylor), jvelasqu@dii.uchile.cl (J.D. Velásquez), fjb11@students.waikato.ac.nz (F. Bravo-Marquez).
URLs: http://wi.dii.uchile.cl/ (J.D. Velásquez), http://www.cs.waikato.ac.nz/~fjb11/ (F. Bravo-Marquez).

http://dx.doi.org/10.1016/j.eswa.2014.05.045
0957-4174/© 2014 Elsevier Ltd. All rights reserved.

goods and intangible services. To the best of our knowledge, most of the existing works in this topic, including Liu's, are focused only on physical product reviews. In these kinds of reviews, users generally go straight to the point and talk directly about the product features they liked or did not like. Furthermore, few people will care about issues like who has designed or manufactured the product. However, for other kinds of products, different phenomena occur.

Works like (Cruz, Troyano, Enríquez, Ortega, & Vallejo, 2013) have already discussed the importance of the domain in the field of opinion mining. For instance, (Zhuang, Jing, & Zhu, 2006) indicates that when a person writes a movie review, he probably comments not only movie elements, but also movie-related people. However, few authors have focused into the field of tourism products like restaurants, which provide a physical good (the food) but also services in the form of ambience and the setting. A detailed study of on-line tourism product reviews revealed the most prominent features appearing on this domain, which we then capture and model in our extension. In general terms, we realized that users tend to *tell stories* about their experiences when writing these reviews, using longer and more complex sentences. The following example, taken from a real review in TripAdvisor, is intended to introduce the features that we will later focus on.

> *"We had a lot of trouble finding the place, but after a while we finally made it. When we arrived to the hotel, it looked really good and only after trying several rooms we discovered the whole hotel was really mouldy in the interior. I barely had enough room to move around the 2 very small/short twin beds and the bathroom was smaller than most standard closets."*

In the first place, a lot of sentences include multiple mentions of the product that is being reviewed or also of any of its features and components. On the other hand, a lot of sentences contain no opinions, also mentioning objects that do not correspond to attributes or components of the reviewed product. These sentences are usually explanations of the writer's experience and help to elaborate the *story* is being told. Finally, we realized that tourists might use many different and complex expressions to refer to the features or subcomponents of the reviewed product.

Therefore, the contributions of this paper are mainly three. First, to the best of our knowledge existing approaches do not address the special issues detected in the tourism domain, so we developed a model for aspect-based opinion mining that specially considers these features. This extension also included the development of new summarization and visualization methods that give insights about the customer preferences of each reviewed product. Our idea is based on the well known proposals of Lancaster in Lancaster (1966), which state that customer preferences about a product are intrinsically related to its features. The proposal is that discovering what these features are and defining how customers feel about these features will undoubtedly lead to a better comprehension of preferences, conceived as an evaluative judgment in the sense of liking or disliking an object (Scherer, 2005).

Secondly, as a result of the analysis of the domain, we created special corpora or datasets that help portraying the features of the mentioned domain. We also use these datasets for the evaluation of the proposed models for opinion aspect-based mining. Finally, our work also included the development of a generic architecture for an aspect-based opinion mining tool, which we used to create 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.

The rest of this paper is structured in the following manner. In first place, we discuss related state-of-the-art techniques and applications in Section 2. Later, in Section 3, we do a complete revision of Bing Liu's ideas, which served as inspiration of this work. Then, we introduce our extension in Section 4 and our system architecture in Section 5. After, we present the results of our experiments and application, in Section 6. Finally, Section 7 details conclusions and proposed future work.

2. 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 Liu (2007), the literature offers two main approaches, aspect-based and non-aspect-based opinion mining. Aspect-based opinion mining techniques divide input texts into *aspects*, also called features or subtopics in literature, that usually correspond to arbitrary topics considered important or representative of the text that is being analyzed. The aspect-based approach is very popular and many authors have developed their own perspectives and models. Examples of them are (Archak, Ghose, & Ipeirotis, 2007; Decker & Trusov, 2010; Ku, Liang, & Chen, 2006; Lu, Zhai, & Sundaresan, 2009; Popescu & Etzioni, 2005; Titov & McDonald, 2008; Zhao & Li, 2009; Zhuang et al., 2006).

Based on an extensive revision of the state-of-the-art approaches and tools, we concluded that Bing Liu's ideas were probably the most comprehensive models on the topic of aspect-based opinion mining. For that reason, his ideas were used here by us as inspiration. In general, our work is based on the ideas summarized by Liu in Liu (2007), which includes a review of the state-of-the-art models, with special attention to his ideas. Most these ideas had already been discussed in the corresponding papers by Liu and his colleagues. Our approach is different from Liu's ideas since it is domain focused; intended to perform well with tourism product reviews. Other reviews of the state-of-the-art opinion mining techniques can be found in Kim, Ganesan, Sondhi, and Zhai (2011), Pang and Lee (2008) and Marrese-Taylor, Rodríguez, Velásquez, Ghosh, and Banerjee (2013).

Other related work includes (Xu, Cheng, Tan, Liu, & Shen, 2013), which proposes an approach for aspect-based opinion mining based on modified versions of Latent Dirichlet Allocation (LDA), similar to what is proposed in the pioneer paper (Titov & McDonald, 2008). These approaches are unsupervised topic-based document modeling techniques, which model an input document as a mixture of topics. A good example of this proposal can be found in Dueñas-Fenández, Velásquez, and LH́uillier (2014), where authors present a framework for trend modeling and detection on the Web, based on the fusion of freely available information. In this context, our work lies on a radically different paradigm, as the former consists in identifying the aspects reviewed in a piece of text based on a bag-of-words model of the document, rather than extracting individual feature mentions and their related opinions (Cruz et al., 2013). Therefore, our work is not directly comparable to these kind of works.

On the other hand, it's also possible to mention (Cruz et al., 2013), which analyzes the importance of the domain in opinion mining. On the paper, the authors show that different topics have completely different features and issues. They also developed a system that by the means of human intervention by generating annotated corpora for each domain, is capable of performing well