Available online at www.sciencedirect.com

ScienceDirect

Procedia Economics and Finance 23 (2015) 1666 – 1673

Procedia Economics and Finance

www.elsevier.com/locate/procedia

2nd GLOBAL CONFERENCE on BUSINESS, ECONOMICS, MANAGEMENT and TOURISM, 30-31 October 2014, Prague, Czech Republic

Using Opinion Mining Techniques in Tourism

Cristian Bucur[ab]*

[a]*Postdoctoral researcher, University of Economic Studies, Bucharest, Romania*
[b]*Lecturer PhD., Petroleum-Gas University of Ploiesti, Bd. Bucureşti 39, Ploiesti, Romania*

Abstract

This paper proposes a platform for extraction and summarizing of opinions expressed by users in tourism related online platforms. Extracting opinions from user generated reviews, regarding aspects specific to hotel services, are useful both to clients looking for accommodation, and also hotels trying to improve their services. The proposed system extracts hotel reviews from internet and classifies them, using an opinion mining technique. Platform is evaluated using a manually pre-classified dataset of user reviews. In the paper the efficiency of algorithms are analyzed using text mining domain specific measures, and are proposed methods for improving the results.

Selection and/ peer-review under responsibility of Academic World Research and Education Center

Keywords: opinion mining, sentiment analysis, web mining, hotel reviews, business intelligence

1. Introduction

Recent advances in web technologies and communications influenced the way people can access information. The web has become an enormous deposit of data, to which online users add new information every day. A part of that information is represented by online reviews. People now read these reviews and are influenced by them in the process of acquiring a product or service. But the enormous amount of data makes impossible for one to read it all. In this context is becoming important to have an automated system for collecting and processing data, capable of presenting to users relevant information.

* Cristian Bucur. Tel.: +40-720-065-651.
E-mail address: cr.bucur@gmail.com

Selection and/ peer-review under responsibility of Academic World Research and Education Center
doi:10.1016/S2212-5671(15)00471-2

Detection and extraction of opinions from online reviews is part of a new area of research developed in last decade. Opinion mining, also called in scientific literature as sentiment analysis, studies the determination and classification of opinions or feelings expressed in text, through the use of computing machines. The challenge of the research area is to extract knowledge from unstructured data. The reviews contains opinions expressed in natural language, common to people but uninterpretable by computers (Bucur, 2014).

The domain of tourism extended activity online in the last decade. There are a lot of people that book accommodation online because is less time consuming, cheaper and they have the possibility to get detailed information about facilities and location of hotels. Concomitantly to development of online booking platforms, sites dedicated to presenting reviews in tourism also evolved. Booking sites also include sections with reviews about presented hotels.

The advantage of having access to information and feedback, make users to prefer online booking. Studies about consumers online behavior revealed that the decision of acquiring a product is very much influenced by other buyers opinions (Bucur, 2014).

In the past one had trouble deciding to make a booking to a hotel not found in a guide or recommended by an agency, due to the lack of information. Now the problem is the excess of information. With so much sites providing rating and feedback, is impossible to read it all and, become extreme difficult to find the relevant information for one to get an overall image. Some sites only provide a rating system (by stars or numbers) or text reviews, others also provide a text review and a rating (Kasper & Vela, 2013).

A simple number on a rating system is not providing enough information, but neither a long review in which users express opinions about more than hotel features. There are a lot of reviews problems, which make them difficult to evaluate. Some of them are:

- Reviews are not concise
- Scalar reviews make difficult to compare hotels with different services offered
- Reviews refer to more than simple hotel accommodation
- Totally different opinions from one user to another
- Some aspects are more important so overall rating is not objective but more influenced on that aspects
- Some reviews contains answers of hotel stuff to customers complains

A system that could summarize the reviews, extracting the opinions from all this information, offering an overall perspective, would save a lot of time and ease the decision process for consumers. Such a system would also help hotel managers to find out how their hotel is seen by customers, what services they liked or disliked. A constructive feedback would help them on improving their services.

There are several methods used in research domain for extracting opinions from hotel reviews. The most used ones are approaches based on natural language processing techniques and lexical resources and approaches based on machine learning.

Research methods based on natural language processing and lexical resources are using part of speech identification and lexical databases like WordNet or other resources derived from it. Most methods based on machine learning are using Naïve Bayesian and Support Vector Machine (SVM) classification (Wilson, Wiebe and Hoffmann 2005). Naïve Bayesian method is using probability concepts and is based on Bayes theorem. Support Vector Machine is a supervised learning method used for classification by recognizing patters in data.

There are also opinion mining research methods that use multiple approaches combining supervised learning methods with lexical resources or ontologies, called hybrid approaches (Saggiona & Funk, 2010).

2. The proposed system architecture

The proposed framework has a modular architecture and uses an unsupervised method and a lexical resource to extract opinions from user reviews posted on TripAdvisor website. TripAdvisor is a travel web platform dedicated to publishing user generated content. On the website users are allowed to add reviews of travel related content. In the picture below we present an example of a hotel review:

Fig. 1. A review on TripAdvisor.com.

The system consist of two modules: a content acquisition module which collects the reviews from website and an analysis module, witch pre-process the extracted data and implements opinion mining process.

2.1. The acquisition module

The acquisition module consists of a web crawler that visit the tourism website starting from a given URL. The crawler collects all the links found in visited pages and register the visited ones. The content of visited pages that contain reviews, is sent to content extraction module that parse the html source of page and extract the review. (See figure 2). The extraction is done by using a predefined mask specific to visited website (Bucur & Tudorica, 2012). In this case the review is contained in a <p> tag inside a with class "entry".

```
<div id="review_233295839" class="reviewSelector ">
<div id="review_233164978" class="reviewSelector ">
  <div class="review basic_review inlineReviewUpdate provider0
  newFlag" style="display: block;">
    <div class="col1of2">
    <div class="col2of2">
      <div class="innerBubble">
        <div class="wrap">
          <div class="quote">
          <div class="rating reviewItemInline">
          <div class="entry">
            <p class="partial_entry">
              Just finished 4 excellent nights at Hotel
              Artemide. Certainly not the least expensive
              hotel but absolutely worth every penny. Great
              bar and restaurant on the 7th floor with
              wonderful views of the Rome. We had a Jr.
              Suite with sitting area with separate bedroom.
              Perfect set up. Bathroom was spotless as was
              the rest of the room. Wife and I...
              <span class="partnerRvw">
            </p>
          </div>
          <div class="wrap">
        </div>
      </div>
    </div>
  </div>
  <div class="review dyn_full_review inlineReviewUpdate provider0
  newFlag" style="display: none;">
</div>
<div id="review_233106797" class="reviewSelector ">
```

Fig. 2. Extracted HTML content.

The reviews extracted by the acquisition module are stored in ***Reviews Deposit***. The proposed solution uses a MySQL database as storage solution for reviews content.

2.2. The analysis module

The analysis module process the reviews from deposit and implement the opinion mining process. It includes the processing module, opinion mining module and SentiWordNet lexical database. Opinion mining is performed using an unsupervised approach at multiple level: word level, sentence level and document level. The processing module process the text for each review and split it into sentences. The review sentences are evaluated identifying parts of speech using a POS tagging algorithm. Proposed platform uses an implementation of Eric Brill algorithm in PHP and a Brown University lexicon corpus as training dataset.

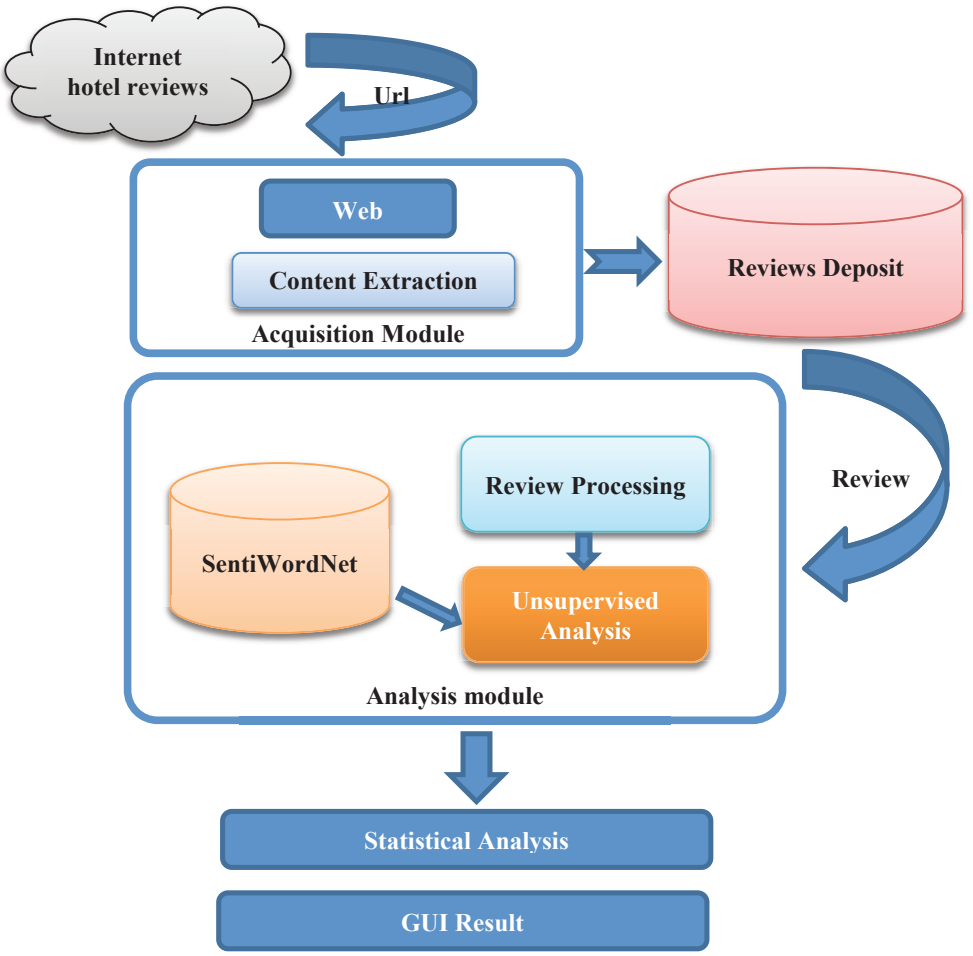

Fig. 3. Proposed platform for opinion mining in tourism.

For each sentence an opinion mining analysis is performed. Each sentence, through a tokenization process, is split into component words. The words polarity is evaluated using SentiWordNet.

SentiWordNet is a lexical resource derived from WordNet which assigns numerical values to each sysnet, representing the scores of positivity, negativity or objectivity (Esuli, Sebastiani, Baccianella 2010). Each score has a value between 0 and 1 and the sum of positivity, negativity or objectivity scores is 1. In proposed platform we use a

modified version of SentiWordNet transformed in a MySql data table. This resource was developed by Adam Westerski in GI2MO project (http://www.gi2mo.org/) and used for OPAL Drupal module. This version has only the highest absolute value of the three scores for each sysnet and the value is positive or negative depending of the predominant polarity.

According to each sysnet polarity value from SentiWordNet, it is calculated a sentence score, as a summary of component words scores. A threshold absolute value of 0.2 is selected for determining the classified class for sentence. So, if the obtained score is below -0.2 the sentence opinion is classified as negative, also if the score is greater than 0.2, is classified as positive. If the score is between -0.2 and 0.2 the sentence is considered as objective or neutral.

The document level evaluation of opinion is made by summarizing the score obtained for each sentence in review. The above classification rules with the threshold of 0.2 are also used at document level.

3. Platform Results

For estimating the performance of proposed system we use a dataset of reviews extracted from TripAdvisor and manually pre-classified. The opinion analysis corpus was collected by Enrique Vallés Balaguer and Paolo Rosso researchesr at the Natural Language Engineering (NLE) Lab, Universitat Politècnica de València (Technical University of Valencia), Spain (http://users.dsic.upv.es/grupos/nle/?file=kop4.php).

The corpus contains 3000 reviews posted by users on TripAdvisor.com about hotels in Rome. Reviews have been manually classified in positive and negative. Each review was extracted in a separate file tagged as one of two classes. In dataset there are 1500 files containing positive reviews and 1500 files containing reviews classified as negative (see table below).

Table 1. Evaluation corpus.

Review classes	No. of reviews	No. of sentences.
Positive reviews	1500	15377
Negative reviews	1500	16551

In total there are 31.928 sentences, with a medium of 10 sentences per review. The files have been processed and content was introduced into the platform Reviews Storage as a MySQL table.

System classification performance was evaluated using text mining specific measures. For current system are calculated: precision, accuracy, recall and F-measure. These values are calculated based on confusion matrix.

In the following table we present the confusion matrix for all classified reviews in the three classes (positive, negative and neutral) according to presented method:

Table 2. Confusion matrix for model.

Reviews	Classified as Positive	Classified as Negative	Classified as Neutral
Positive	1284	135	81
Negative	459	882	158

Accuracy is the proportion of correct classified instances in total classified instances. An accuracy of 1 means that all instances are correctly classified. Precision is the proportion of correctly classified instances from a class against all instances classified (predicted) as being part of that class. Recall is the proportion of correctly predicted instances of a class against all actual instances of that class (Padmaja & Fatima, 2013).

The following table presents the calculated values of accuracy, precision, recall and F-measure for 100, 1000, 2000, and all 3000 reviews.

Table 3. Evaluation of system performance.

No. of Reviews	Precision	Recall	Accuracy	F-measure
100	0.75510204081633	0.74	0.72	0.74747474747475
1000	0.78057553956835	0.868	0.765	0.8219696969697
2000	0.76232394366197	0.866	0.745	0.81086142322097
3000	0.73666092943201	0.856	0.722	0.79185938945421

The results obtained show that proposed system has an accuracy between 72% and 76.5%. The performance obtained is good for an unsupervised method, but lower than results obtained with supervised algorithms. From Table 2 it is noticed that prediction for positive reviews was better than for negative ones.

Approximately 30% of negative reviews were classified as positive and 13% of total negative and positive reviews were classified as objective. Analyzing the reviews it was observed that many users express different or mixed sentiments. They have a positive opinion regarding some aspects and negative opinion to other aspects, thus, in some cases a human can better distinguish the overall opinion of review. Also, in many reviews users express opinions about their travel experience, rather than just about hotel. All of these problems could cause error in classification process.

It was noticed that regardless of the number of reviews classified, the accuracy remain approximately constant, because in an unsupervised process, the volume of data does not influence the classification process. An advantage of the proposed platform is that it does not require the use of a training dataset, which is a resource consuming operation. Another advantage, due to proposed algorithm, is that the proposed platform is not domain dependent.

Table 4 Time spent performing the analysis

No. of Reviews	Number of sentences	Time spent	Accuracy
100	937	539.24	0.72
1000	10347	5985.11	0.765
2000	21361	12371.8	0.745
3000	31928	17987.58	0.722

The proposed platform performance is analyzed regarding the time needed for performing the classification. Tests were performed on a computer with an AMD Quad Core 3.4GHz processor and 8 GB RAM memory. As we can see in Table 4 the medium time to classify a review was around 6 seconds. It can be seen that, from the point of view of a real time application, the process is time consuming and cannot be used for real time processing but as support background system, if the task consist in analyzing a large number of reviews.

Fig. 4. Complexity of reviews vs time and accuracy.

From the chart in Figure 4 it can be seen that the execution time of classifying process has a linear dependency to the number of classified sentences. Also the number of sentences in a comment does not influence the overall accuracy of classification.

4. Conclusion

In this paper it was presented an opinion mining platform for extracting and classifying hotel reviews posted by users on tourism websites. The system visits web pages starting from a given URL, extracts the reviews from page content then uses an opinion mining module to process the content and classify reviews as positive, negative and neutral.

The proposed process has an acceptable accuracy and has the advantages that is domain independent and does not need expensive resources to operate. Analyzing the reviews, it can be concluded that, in the domain of tourism an aspect oriented analysis would improve the performance of the platform, due to the multitude of aspects users express opinions about, and the mixed sentiments that are present in reviews.

A future direction for improving the performance could be the use of an ontology, oriented to tourism domain. The proposed architecture could be a useful background tool for summarizing the opinions in tourism oriented web platforms.

Acknowledgements

This work was cofinanced from the European Social Fund through Sectorial Operational Programme Human Resources Development 2007-2013, project number POSDRU/159/1.5/S/134197 „Performance and excellence in doctoral and postdoctoral research in Romanian economics science domain”.

References

Amit Moran. (2012, Mar.) Sentiment Analysis: How does sentiment analysis work? [Online]. http://www.quora.com/Sentiment-Analysis/How-does-sentiment-analysis-work#ans606524

Andrea Esuli, Fabrizio Sebastiani Stefano Baccianella, (2010)."SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining," in Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10), Malta, Valetta.

Cristian Bucur, (2014). Opinion Mining Platform for Intelligence in Business, *Economic Insights-Trends and Challenges*, *Volume 3,* No 3/ p. 99-108, ISSN 2284-8576

Cristian Bucur, Bogdan George Tudorica, (2012).A research on retrieving and parsing of multiple web pages for storing then in large databases", *Revista Economica*, Supplement No. 5/, pag. 119-127, ISSN: 1582 – 6260

Cristian Bucur, Implications and Directions of Development of Web Business Intelligence Systems for Business Community, *Economic Insights-Trends and Challenges*, *64* (2), 96-108

Erik Cambria, Bjorn Schuller, Yunqing Xia, Catherine Havasi, (2013). "New Avenues in Opinion Mining and Sentiment Analysis," *IEEE Intelligent Systems*, *vol. 28*, no. 2, pp. 15-21, March-April, doi:10.1109/MIS.2013.30

Lillian Lee B. Pang, (2008). "Opinion mining and sentiment analysis," *Foundations and Trends in Information Retrieval, vol. 2*, no 1-2, pp. 1-135.

Opinion Mining from a Large Corpora of Natural Language Reviews, Beltrán Borja Fiz Pontiveros, (Master of Science Thesis), September 2012

Padmaja, S., Sameen S., Fatima, (2013)**.**Opinion Mining and Sentiment Analysis –An Assessment of Peoples' Belief: A Survey, *International Journal of Ad hoc, Sensor & Ubiquitous Computing (IJASUC) Vol.4,* No.1, February

Rao D. and Ravichandran, D. (2009)."Semi-Supervised Polarity Lexicon Induction," *Proc. 12th Conf. European Chapter of the Assoc. for Computational Linguistics, Assoc. for Computational Linguistics*, pp. 675–682.

Saggiona, H., Funk, A., (2010)."Interpreting SentiWordNet for Opinion Classification", *Proceedings of the Seventh conference on International Language Resources and Evaluation LREC1*.

Sebastiani F. & Esuli, A., "SENTIWORDNET: A Publicly Available Lexical Resource for Opinion Mining," in *Proceedings of the 5th Conference on Language Resources and Evaluation (LREC'06)*, 2006, pp. 417-422.

Wiebe, J., Wilson, T. and C. Cardie, "Annotating Expressions of Opinions and Emotions in Language," *Language Resources and Evaluation, vol. 39*, no. 2, 2005, pp. 165–210.