Using Big Data and Sentiment Analysis in Product Evaluation

Lada Banić, Ana Mihanović, Marko Brakus
Poslovna Inteligencija d.o.o., Zagreb, Croatia
{lada.banic, ana.mihanovic, marko.brakus}@inteligencija.com

Abstract - When purchasing a product for the first time one usually needs to choose among several products with similar characteristics. Companies use to promote their brands and products pointing out good characteristics avoiding to mention the poor ones. The best way to choose the most suitable product is to rely upon the opinions of others. The system to be described here collects opinions about hotels from the web, evaluates them, aggregates these evaluations and offers cumulative, easy-to-understand information. Generated information is intended for the possible prospective customer, but also for the hotel managers providing them with additional guidance in future business development.

I. INTRODUCTION

Rapid growth of online information concerning services and products has been accompanied by users' comments about these items. The amount of reviews has been increasing at a great speed making the web more and more subjective and opinionated. The Web visitors' comments cover almost all areas, as they are present not only on specialized review sites, but also on most of the published news and blog pages.

Independent unbiased consumer reviews are known to be the most credible sources of product or service information and people tend to rely primarily on them when making a decision about a purchase. According to [1] about 80 percent of users of TripAdvisor write travel reviews, 20 percent of visitors rely on other users' reviews when planning a trip; looking into other users' comments and travel blogs is the most popular online activity. Analysis of such texts is a more productive way of collecting user information than the traditional structured data collection by surveys where people are usually unwilling to take the time to answer presupposed questions. Conversely, sentiment analysis [2] "listens" to published opinions and answers "the unknown". However, due to its size, this online form of a word of mouth is difficult to grasp fully. Reading the reviews only partly might induce a biased opinion. This calls for application of algorithmic methods of analysis of a large number of reviews.

When working with the web content concerning online reviews, blogs, forums etc, one deals with huge amounts of unstructured data in the attempt to extract information. To be successful one needs those data to be structured so that the necessary information becomes available. When extracted, the information needs to be aggregated and presented to the interested party(s) in an understandable form.

During the process of unstructured data collection, information extraction or in this case sentiment extraction, aggregation of gathered information and presentation to the interested party(s) one is dealing with several innovative issues:

- BigData (storing and analyzing large amounts of unstructured data)
- text mining (deriving information from text)
- sentiment analysis (finding out opinions from text)

A. BigData

The term BigData [3] stands for the process of extracting actionable data from various, often nontraditional data sources. These sources may include structured data such as databases, and in this case unstructured data like HTML, as well as social data and images.

Unstructured data (or unstructured information) refers to the information that either does not have a pre-defined data model and/or does not fit well into relational tables. Unstructured information is typically text-heavy, but may contain data such as dates, numbers, and facts.

The unstructured source data are then structured up to a point and pushed down to a structured format, which is then stored in a database for further manipulation.

The amount of data being collected is more than traditional computer infrastructures can handle, exceeding the capacities of databases, storage, networks and everything in between. Extracting actionable intelligence from BigData requires handling large amounts of various data and processing them very quickly. Major issues in BigData processing is that data inputs must be consistent and clean.

IBM describes these new demands operating with BigData across four dimensions: volume, velocity, variety and veracity, all of them overwhelming

B. Text Mining and Text Analytics

Text mining is analysis of data contained in a natural language text. The application of text mining techniques to solve business problems is called text analytics. It is an interdisciplinary field which combines information collection, data mining, statistics and computational linguistics.

The goal of text mining [4] is to derive high-quality information from the text. This is typically done through

recognizing the patterns in data. In other words, the purpose of text mining is to process unstructured information and to extract meaningful numeric indices from it. Generally speaking, text mining 'turns text into numbers'.

The numeric indices make the information contained in the text accessible to further analysis or to further data mining (statistical and machine learning) algorithms.

By text mining, different analyses are possible: of words and of clusters of words within documents, of similarities between documents, of the relation of documents to other variables of interest to e.g. data mining projects etc.

Text analysis involves the following processes:

- information retrieval
- Natural language processing
- named entity recognition
- recognition of pattern identified entities (features such as telephone numbers, e-mail addresses, quantities (with units)
- coreference: identification of noun phrases and other terms that refer to the same object
- relationship, fact, and event extraction: identification of associations among entities and other information in text
- sentiment analysis
- Quantitative text analysis

C. Sentiment Analysis

Sentiment analysis [2] is one of the applications of the text mining techniques.

An important aspect of our information-gathering is to find out what other people think. Even before the World Wide Web has become widely spread, people asked friends for opinion on different subjects in order to make a better and wiser decision. The availability of the world-wide-web and text-mining techniques allows us, on a much wider scale, to find out opinions of other people who are neither our personal acquaintances nor well-known professional critics. In addition to individuals businesses also seek to identify and capture the substance of the "word of mouth", that is, the information consumers exchange with one another. Their aim is to manage the impact which this information, along with its consequent e-reputation, can have on their products and brands, and to take it into account when developing a strategy or improving business operation.

II. PRODUCT EVALUATION SYSTEM OVERVIEW

The system presented here has been being developed as part of the FAIR project, which is carried out by three partners: Testntrust from France, Beia from Romania and Poslovna inteligencija from Croatia.

The FAIR project encompasses customer satisfaction ratings, brand ratings, social networking, connecting brands with their customers, exchanging and sharing relevant information.

The first product chosen for sentiment analysis and evaluation was hotel.

The project is concerned with collecting hotel reviews, storing them, and analyzing their sentiment and aggregating analysis results into single hotel-based estimation. The system of review collection handles crawling, extracting the hotel reviews and storing them for analysis. The review dataset obtained is subjected to text mining and sentiment analysis resulting in evaluation of every single review. Review evaluations are aggregated on the hotel level in order to get a cumulative estimate for each hotel.

A. Data (Hotel Reviews) Collection

In gathering hotel review data [5], focus was set on the web sites specialized for travelers' reviews. The notable ones were tripadvisor.com, hotels.com, laterooms.com, booking.com, the tripadvisor.com being the most prolific. Review pages were obtained by crawling the web sites for hotels and reviews, using the sitemaps.org format and reading RSS feeds. The most efficient solution is offered by sitemaps.org, an XML file format specifying a map of the web site using an updatable list of URLs. Similar and more frequent solutions were RSS feeds published on the web site. Both sitemaps.org XML index files and the RSS feeds were checked periodically for newly added or updated URLs. However, in a large majority of cases, such indices of site's interesting and newly updated web pages were not available and one had to rely on the web site crawling.

A web crawler is a program that traverses the web site starting from a given set of initial URLs and follows the links matching a given pattern to a certain depth. An ideal crawler for the purpose of quickly downloading only the pages containing reviews and checking if they are updated, would be distributed or at least parallel, incremental and focused. To update a set of downloaded pages it is preferable to apply incremental crawling rather than to restart crawling. In focused crawling the space of crawled pages is narrowed by the use of a classifier, which decides whether a page is interesting or not. This can be a simple URL pattern to match. Parallel crawling is performed by running multiple processes simultaneously, to crawl web sites in a reasonable amount of time. For the aforementioned reasons and for the purpose of work planned in the future, the general-purpose web crawler offered by Apache Nutch [7], an open source web-search software project was used.

All hotel review sites have their own review page layout usually containing several reviews for a hotel. Although it is not very scalable to have a manually created scrapping template for each site, currently, we applied the DOM (Document Object Model) based scrapping method to extract review and hotel information from the HTML pages using the templates defined by sets of XPath and regular expressions. More automated methods of review scrapping appropriate for large-scale crawls are described in [6].

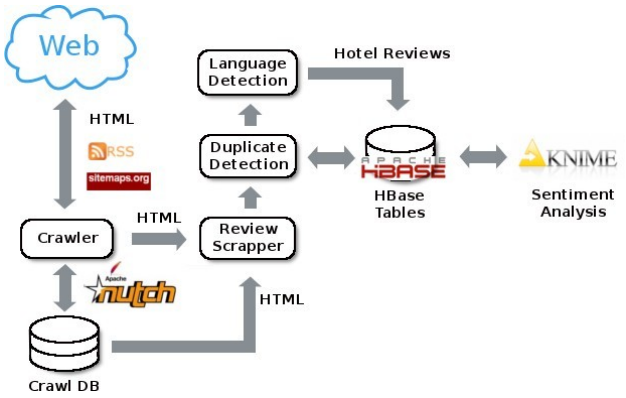

Figure 1. Product Evaluation System configuration

The extracted review information comprises the name (nickname) of the reviewer, the date of posting and the review text. The hotel and review information extracted from HTML was further processed to check the language of the text, as most of the sites contain multilingual reviews. Language detection of the review text was done by means of the Apache Tika [8] content analysis toolkit, which uses the N-gram technique. Special attention was paid to extracting and normalizing a unique hotel name out of the title of the hotel review, as the same hotel can be differently titled across the web sites.

Duplicate reviews were a result of repetition of the review on different pages inside the web site and of its copying across different sites. Duplicate detection was realized from coarse to fine, meaning that the URL name was first checked against the table of the pages visited earlier and the extracted review (its hash value) was then looked up in the table of all reviews.

To make the system extendable to products and services other than hotels and to scale to ever expanding unstructured data, Apache Hadoop [10], an open source implementation of the MapReduce framework was used as a distributed execution and storage environment. The Hadoop is popular for developing large-scale data-intensive applications. The Hadoop ecosystem comprises components such as HBase, Hive and Pig for storing, querying and analyzing data, respectively. Hotels and hotel reviews are stored in HBase tables. HBase [11] is a distributed NoSQL database, which efficiently holds unstructured data in large tables and can be concurrently and randomly accessed.

B. Dataset Description

A dataset of hotel reviews collected from tripadvisor.com. was used. Tripadvisor.com, a travel web site, offers information on a big diversity of touristic sites including hotels, restaurants, museums etc. Most of the information on this site is user generated. Users can log in and post a review for any touristic site. Also, hotel managers have a possibility to reply to any review.

Reviews for a given hotel are usually displayed in a list of five per page. Hotel information is displayed at the top of the page. When extracted from a web page, the web page content and the extracted information were structured in the following way:

1. For each extracted review a key indicator was created before the review was stored into a database

2. The hotel details extracted from the web site content were:
   - hotel name,
   - country,
   - city,
   - street.
3. Hotel reviews details included:
   - text of the review,
   - date of the review,
   - language of the text.

Hotel details were extracted in the mentioned way so that each hotel could be uniquely identified in the process of aggregation of evaluation results. Namely, if aggregations had been performed based on the hotel name only, evaluations of different hotels carrying the same name would have been aggregated, producing wrong evaluation results and supplying the end users with misleading information.

The review date allowed creation of aggregated evaluation for a defined period of time and follow up of the time related changes in evaluation results. The language attribute allowed the application of a proper dictionary in evaluation of downloaded review, for the English dictionary applied to a French review would fail to produce results.

In this phase of the process only the hotel reviews in English were downloaded.

C. Text Analytics and Sentiment Analysis

Text analytics and sentiment analysis were performed by means of the open-source software KNIME [9].

KNIME is a user-friendly graphical workbench for the entire analysis process: data access, data transformation, initial investigation, powerful predictive analytics, visualization and reporting. The open integration platform provides over 1000 modules (nodes), including those of the KNIME community and its extensive partner network.

By means of KNIME a sentiment analysis stream, consisting of the following major steps, was created:

- Retrieving data from the database
- Dictionary development and implementation
- Review scoring

1) Retrieving Data from the Database

Connectivity to the database was accomplished with the (in-house developed) „HBase reader“ node, which took a few simple parameters to connect to the database, to the host machine IP and to the port. Data were retrieved from the database in packages of 30 000 records. Only records that were not previously evaluated were retrieved, assuring that the reviews from the database were evaluated only once, speeding up the whole process.

2) Dictionary Development

Dictionary development and implementation A new dictionary containing words and phrases used in evaluation of hotel reviews was developed and implemented. It was developed based on the words and phrases found in a sample of downloaded Internet hotel reviews.

Composing a dictionary is precise and time-consuming task. It was necessary to include the terms and phrases from the current hotel reviews which could contribute to the evaluation, and to multiply them and modify them in a way in which they might occur in other users' reviews, following grammar rules but including also slang expressions and abbreviations used in everyday speech.

Hotel evaluation was planned to be carried out according to four different categories: tidiness, service, atmosphere and general category. Every term in the dictionary was associated with only one of the four categories and given a grade in the range from 1 to 5, 1 referring to bad, and 5 referring to excellent.

The category tidiness was meant to evaluate the cleanness of accommodation including room, bathroom and hotel in general, the category service dealt with the affability of the staff, the category atmosphere served for evaluation of hotel location, noise level and similar characteristics, whereas the category general included all terms suitable for hotel evaluation that could not be included in the previously mentioned categories.

Dictionary for KNIME is made with a simple text editor. It consisted of several columns: one column referred to the term of the phrase, the other columns referred to the above mentioned categories for evaluation, and only one of them contained the grade for the respective term or phrase.

The term or phrase within dictionary where word or the sequence of words to be recognized in an unstructured review text.

3) Review Scoring

When loaded from the HBase into KNIME reviews are treated as single documents.

Dictionary is applied to every document resulting in information about terms and phrases from the dictionary found within document.

Every term or phrase recognized in the document was assigned category and evaluation grade as specified in the dictionary.

In every document multiple terms or phrases belonging to the same evaluation category could be found, so that an average evaluation grade was calculated for each review and category.

The results, i.e. average grade for each evaluated review for each category was written back to the Hadoop database.

As a result of Hadoop aggregation of evaluated reviews for each hotel the average grade for every specified category on the level of single hotel is calculated.

Evaluation of reviews was performed periodically as was aggregation allowing a timely follow up of changes.

III. Development of Evaluation Method

Different evaluation systems can be implemented in the process of semantic analysis. The choice of the system will depend on the product of evaluation, its characteristics, on the end users of such cumulative information and on the level of information to be extracted.

In our first evaluation system we evaluated each term or phrase as negative and positive. Cumulative information about the number of positive and negative terms or phrases was generated on the level of the hotel and final evaluation of the hotel was generated in the following way:

- more than 70 percent of terms are positive – grade 3 (referring to very good)
- between 70 percent and 30 percent positive terms – grade 2 (referring to average)
- less than 30 percent positive terms – grade 1 (referring to bad)

The second evaluation system involved evaluation of terms and phrases with the help of grades from 1 to 5, where 1 referred to bad and 5 referred to excellent. Each term or phrase recognized in the review was evaluated according to the specification in the dictionary. Average grade for each single review was obtained. The total grade for each hotel was calculated as average grade of all reviews aggregated on hotel level.

As evaluation system with only one total average grade did not provide enough valuable information about hotels, its development was taken one step further by specification of evaluation categories.

Four different categories were introduced:

- location,
- service,
- atmosphere
- general (terms that did not fit in any of the specified categories).

Every term was categorized before ti was graded with marks in the range from 1 to 5. For every review all four categories were specified separately. The total grade for every hotel was determined as average grade of all reviews aggregated by category on the hotel level.

IV. Results

Reviews were collected using the sitemaps.org format from tripadvisor.com web site.

A total of 105 294 reviews were downloaded for a total of 3403 different hotels.

The developed dictionary comprised all together 1211 terms and phrases extracted from hotel reviews that expressed customer's specific opinion about hotels. In dictionary development, emphasis was on expressions because the same term within different expressions could have a positive and a negative meaning.

An overview of a number of terms and expressions by category, included in the dictionary, is given in Table I.

The use of dictionary resulted in evaluation of 40562 reviews in the tidiness category, of 50425 reviews in the

service category, of 31437 reviews in the atmosphere category and of 85883 reviews in the general category.

Results of review evaluations by category with percentages calculated with respect to total number of downloaded reviews are shown in Table II.

The reason for low percentages of evaluated reviews lies with the dictionary which is still under development. For a more complete and reliable review evaluation more efforts should be made to enhance development of dictionary.

Hotel evaluation is result of aggregation of review evaluations on the hotel level. Results of hotel evaluation by category with percentages calculated with respect to total number of hotels are shown in Table III.

TABLE I.. NUMBER OF TERMS AND PHRASES IN THE DICTIONARY BY EVALUTAION CATEGORY.

	Tidiness	*Service*	*Atmosphere*	*General*
Terms	7	13	7	25
Phrases	296	367	223	273

TABLE II. THE NUMBER OF REVIEWS EVALUATED BY CATEGORY, AND PERCENTAGES CALCULATED WITH RESPECT TO TOTAL NUMBER OF DOWNLOADED REVIEWS.

Number of downloaded reviews	*Tidiness*	*Service*	*Atmo-sphere*	*General*
105294	40562	50425	31437	85883
	38,52%	47,89%	29,86%	81,56%

TABLE III. RESULTS OF THE HOTEL EVALUATIONS BY CATEGORY WITH PERCENTAGES CALCULATED WITH RESPECT TO THE TOTAL NUMBER OF HOTELS

Number of evalauted hotels	*Tidiness*	*Service*	*Atmo-sphere*	*General*
3403	2729	2927	2469	3267
	80,19%	86,01%	72,55%	96,00%

TABLE IV. NUMBER OF TERMS AND PHRASES FOR EVERY CATEGORY EVALUATED WITH A CERTAIN GRADE AND IMPLEMENTED INTO THE DICTIONARY..

Category Grade	*Tidiness*	*Ser-vice*	*Atmo-sphere*	*Gener-al*	*Total by grade*
1	50	98	29	56	233
2	60	59	40	44	203
3	9	20	25	42	96
4	76	72	47	45	240
5	108	131	89	111	439
Total by category	303	380	230	298	

From the results in Table III, it is clear that the coverage of evaluations on the hotel level is much higher than on the review level. This is because single reviews which were evaluated according to one or two different categories, when aggregated on the hotel level cover in total more categories then single reviews.

Terms and phrases implemented in the dictionary were collected from a sample of downloaded reviews. In the process of dictionary creation, terms and phrases were assigned to categories and graded.

Table IV. shows the number of terms and phrases for every category evaluated with a certain grade, implemented into the dictionary and then searched for within the reviews structuring reviews and creating bases for reviews evaluation.

Statistics of grades appearing for the categories in the review evaluation are given in Figure 2. From the graph presentation, it is obvious that the terms and phrases most often found in the review evaluation are those that are graded within the category “general” with excellent mark, within the category “service” with a very good grade and within the categories “general” and “tidiness” with grades very good and excellent, respectively.

According to the statistics of terms and phrases belonging to the different categories and grades, shown in Table IV, it was expected for excellent grade to have the highest number of “hits” in the review evaluation, for most of the terms and phrases in the dictionary are graded as excellent. However, according to the statistics of terms and phrases implemented in the dictionary, shown in the Table IV, that for the “service” category graded with very good mark there would be much more hits for the number of implemented terms and phrases in the dictionary this category and grade is the highest, what does not appear in the graph showing grades appearing for categories in review evaluation (Figure 2). It appears that the terms implemented in the dictionary in that category-grade group and collected from the sample of reviews are not frequently used.

All this findings show us the importance of the dictionary development. The more time spent in the extraction of the terms and phrases from the existing reviews and categorization and grading of those terms, the better, more consistent and accurate the results of the review evaluation will be.

The aggregated evaluation results done on the hotel level show to the end users the current situation and position of the hotel on the market. However, of great importance is the analysis of change of customer's perception of the hotel with time. Since every review has a posting date specified, analysis of change of average grade for all the categories can be followed through time. In the figure 3. analysis for the hotel ‘Al Duca di Venezia’ is presented.

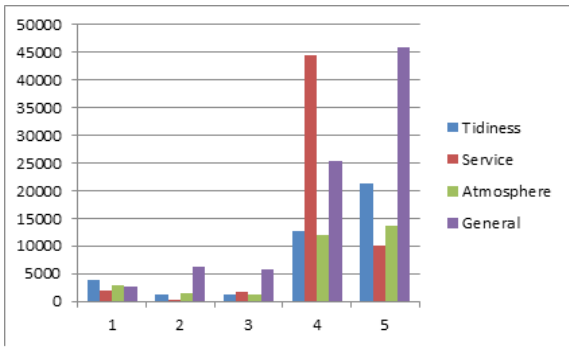

Figure 2. Grades acording to categories in review evaluation

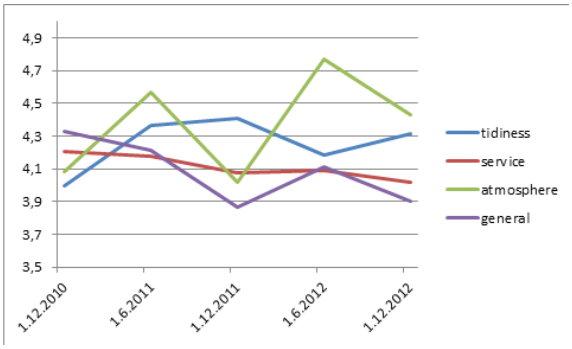

Figure 3. Average grade for all the categories followed through time for ' Al Duca di Venezia' hotel.

Since dictionary is not fully developed and only a limited number of reviews for evaluation was collected, results are not fully reliable, but they do clearly show the drop in evaluation of categories atmosphere and general at the end of 2011, while category service has slide but steady fall from the beginning of 2010 to the end of 2012.

V. CONCLUSION

The best way to decide about the most suitable product is to rely upon the opinions of others. With the growing availability and popularity of sources such as online review sites and personal blogs as well as with the ever advancing information technology which enable efficient processing of large scale unstructured data, new opportunities arise for finding and understanding different opinions and for facilitating the decision making process.

Systems for sentiment analysis of products and brands can be developed. Such systems can transform a vast amounts of unstructured data into an aggregated structured information. However, certain issues that come with processing large amounts of unstructured data in order to structure them have to be considered. One issue is data quality which deals not only with the recognition of the required terms and phases but also with the originality of the collected data. The later will depend on review authors but also on review multiplication which may generate misleading evaluations. Specificity of product also needs to be considered especially when developing a categorical evaluation system which needs to be adjusted to the product type. The terms and expressions contained in the dictionary should also be suited to the product type because of different terms and phrases, language forms and slang expressions which refer to different product types. In the future sentiment analysis system will bring new knowledge to the individuals as well as to businesses people. They are going to make managers more aware of the end users' perception of their products but also their competitors' products. In the environment of saturated market and recession aggregated information based on a vast amount of data will become invaluable for planning future business strategy planning and development.

[1] U. Gretzel, K. H. Yoo, M. Purify, Online Travel Review Study. Role & Impact Of Online Travel Reviews. Research Report. Laboratory For Intelligent Systems In Tourism. Texas A&M University, 2007.

[2] B. Pang, L. Lee, Opinion Mining and Sentiment Analysis, Foundations and Trends in Information Retrieval Vol. 2, Nos. 1–2, 2008

[3] S. LaValle, E. Lesser, R. Shockley, M. S. Hopkins, and N. Kruschwitz. "Big data, analytics and the path from insights to value." MIT sloan management review 52, no. 2, 2011, pp 21-32

[4] R. Feldman and J. Sanger. The text mining handbook: advanced approaches in analyzing unstructured data. Cambridge University Press, 2006.

[5] M. McGlohon, N. Glance, and Z. Reiter. "Star quality: Aggregating reviews to rank products and merchants." In International Conference on Weblogs and Social Media, 2010.

[6] L. Wei, Y. Hualiang, X. Jianguo, Automatically extracting user reviews from forum sites, Computers & Mathematics in Natural Computation and Knowledge Discovery, Volume 62, Issue 7, October 2011, pp 2779–2792

[7] http://nutch.apache.org/

[8] http://tika.apache.org

[9] http://knime.org/

[10] http://hadoop.apache.org/

[11] http://hbase.apache.org/