Opinion Mining Platform for Intelligence in Business

Cristian Bucur

Academy of Economic Studies, Piata Romana No.6, 1st district, Bucharest, Romania
Faculty of Economic Sciences, Petroleum-Gas University of Ploieşti, Bd. Bucureşti 39, 100680, Ploieşti, Romania
e-mail: cr.bucur@gmail.com

Abstract

The Web, the largest available database contains an enormous amount of relevant information. Difficulty accessing this information comes from many forms of presentation. Also the original purpose was to human users access information, so data cannot be automatically processed by machines. This paper represents a study on the various methods of analysis of information from web pages, and proposes a software system design of a web platform dedicated to extracting opinions of customers from online content, necessary as decision support in a company.

Keywords: opinion mining, web mining, business intelligence, web intelligence

JEL Classification: A12, M15, L21

Introduction

An important process in the discovery of knowledge in the content of web pages, is the detection and extraction of opinions or sentiment information on an online text. Determination of customer sentiment on a newly launched product, based on feedback from web pages is important for a company, for assessment of impact and decisions on the directions of development. Opinion mining is an area of research that deals with automatic methods for detection and extraction of opinions and emotions present in a text. Applications of opinion mining methods can result in the creation of effective referral systems, financial analysis, market research and product development.

Detection and retrieval of opinions from information present online is a hot topic in recent years. Consumers use the internet for information and submission of opinions on various issues. They form their opinions based on the opinions of others expressed online.

Background

According to internet live stats¹ number of internet users (as of July 1) is 2,925,249,355. The percent of users from the entire world population increased from less than 1% in 1995 to 40% today. It is estimated that this year the third billion users online will be reached (view figure 1 and table 1). As the web platforms of today encourage users to express their opinion online, all

¹ http://www.internetlivestats.com/internet-users/#trend (accessed 05 September 2014)
these users generate an immense volume of data in natural language intelligible by humans but not by machine capable of extracting knowledge from content.

Summarization of these opinions submitted online, allows those interested to react negative feelings on a particular theme and observing positive reactions. Automatic retrieval of opinions and feelings from multiple sources is becoming increasingly important in the context of the Web and the increasing popularity of online social networks. Manual processing of such information under the circumstances is virtually impossible.

Table 1. Evolution of internet users

<table>
<thead>
<tr>
<th>Year (July 1)</th>
<th>Internet Users</th>
<th>Users Growth</th>
<th>World Population</th>
<th>Population Growth</th>
<th>Penetration (% of Pop. with Internet)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2014*</td>
<td>2,925,249,355</td>
<td>7.9%</td>
<td>7,243,784,121</td>
<td>1.14%</td>
<td>40.4%</td>
</tr>
<tr>
<td>2013</td>
<td>2,712,239,573</td>
<td>8.0%</td>
<td>7,162,119,430</td>
<td>1.16%</td>
<td>37.9%</td>
</tr>
<tr>
<td>2012</td>
<td>2,511,615,523</td>
<td>10.5%</td>
<td>7,080,072,420</td>
<td>1.17%</td>
<td>35.3%</td>
</tr>
<tr>
<td>2011</td>
<td>2,272,463,038</td>
<td>11.7%</td>
<td>6,997,998,760</td>
<td>1.18%</td>
<td>32.5%</td>
</tr>
<tr>
<td>2010</td>
<td>2,034,259,368</td>
<td>16.1%</td>
<td>6,916,183,480</td>
<td>1.19%</td>
<td>29.4%</td>
</tr>
<tr>
<td>2009</td>
<td>1,752,333,178</td>
<td>12.2%</td>
<td>6,834,721,930</td>
<td>1.20%</td>
<td>25.6%</td>
</tr>
<tr>
<td>2008</td>
<td>1,562,067,594</td>
<td>13.8%</td>
<td>6,753,649,230</td>
<td>1.21%</td>
<td>23.1%</td>
</tr>
<tr>
<td>2007</td>
<td>1,373,040,542</td>
<td>18.6%</td>
<td>6,673,105,940</td>
<td>1.21%</td>
<td>20.6%</td>
</tr>
<tr>
<td>2006</td>
<td>1,157,500,065</td>
<td>12.4%</td>
<td>6,593,227,980</td>
<td>1.21%</td>
<td>17.6%</td>
</tr>
<tr>
<td>2005</td>
<td>1,029,717,906</td>
<td>13.1%</td>
<td>6,514,094,610</td>
<td>1.22%</td>
<td>15.8%</td>
</tr>
<tr>
<td>2004</td>
<td>910,060,180</td>
<td>16.9%</td>
<td>6,435,705,600</td>
<td>1.22%</td>
<td>14.1%</td>
</tr>
<tr>
<td>2003</td>
<td>778,555,680</td>
<td>17.5%</td>
<td>6,357,991,750</td>
<td>1.23%</td>
<td>12.2%</td>
</tr>
<tr>
<td>2002</td>
<td>662,663,600</td>
<td>32.4%</td>
<td>6,280,853,820</td>
<td>1.24%</td>
<td>10.6%</td>
</tr>
<tr>
<td>2001</td>
<td>506,609,240</td>
<td>21.1%</td>
<td>6,204,147,030</td>
<td>1.25%</td>
<td>8.1%</td>
</tr>
<tr>
<td>2000</td>
<td>413,425,190</td>
<td>47.2%</td>
<td>6,127,700,430</td>
<td>1.26%</td>
<td>6.7%</td>
</tr>
<tr>
<td>1999</td>
<td>280,866,670</td>
<td>49.4%</td>
<td>6,051,478,010</td>
<td>1.27%</td>
<td>4.6%</td>
</tr>
<tr>
<td>1998</td>
<td>188,023,930</td>
<td>55.7%</td>
<td>5,975,303,660</td>
<td>1.30%</td>
<td>3.1%</td>
</tr>
<tr>
<td>1997</td>
<td>120,758,310</td>
<td>56.0%</td>
<td>5,898,688,340</td>
<td>1.33%</td>
<td>2.0%</td>
</tr>
<tr>
<td>1996</td>
<td>77,433,860</td>
<td>72.7%</td>
<td>5,821,016,750</td>
<td>1.38%</td>
<td>1.3%</td>
</tr>
<tr>
<td>1995</td>
<td>44,838,900</td>
<td>76.2%</td>
<td>5,741,822,410</td>
<td>1.43%</td>
<td>0.8%</td>
</tr>
<tr>
<td>1994</td>
<td>25,454,590</td>
<td>79.7%</td>
<td>5,661,086,350</td>
<td>1.47%</td>
<td>0.4%</td>
</tr>
<tr>
<td>1993</td>
<td>14,161,570</td>
<td></td>
<td>5,578,865,110</td>
<td></td>
<td>0.3%</td>
</tr>
</tbody>
</table>

* estimate for July 1, 2014

Source: Internet Live Stats (elaboration of data by International Telecommunication Union (ITU) and United Nations Population Division)

The majority of online users are influenced by opinion expressed by others. Surveys on customers reveal that buying decision is influenced by online reviews more and more intensive.
A Comscore study from 2007\(^2\) revealed that 75\% percent of customers were influenced by online reviews and are willing to pay from 20\% to 99\% percent more on a product with 5 star review than on a 4 star rated one. A more recent study from Dimensional Research\(^3\) (2013) states that 90\% of respondents say they are influenced by online reviews.


Two-thirds of the respondents said they read online reviews. From the graph in fig. 3 we can see that a large part of reviews are accessed from social network platforms and online review sites. Majority of negative reviews were found on review sites, while social networks represent the leading source for positive reviews.

We can conclude that beside specialized review websites, the main source for online reviews is social networks. As we can see from below table (http://www.ebizmba.com/articles/social-networking-websites) monthly more than 75% percent of worldwide online users visit 15 most popular social networks (over 2.350 million unique visitors). So social networks represent a very important source for collecting content for opinion mining analysis.

<table>
<thead>
<tr>
<th>Social Networking Site</th>
<th>Estimated Unique Monthly Visitors (as of 1 September 2014)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Facebook</td>
<td>900,000,000</td>
</tr>
<tr>
<td>Twitter</td>
<td>310,000,000</td>
</tr>
<tr>
<td>LinkedIn</td>
<td>255,000,000</td>
</tr>
<tr>
<td>Pinterest</td>
<td>250,000,000</td>
</tr>
<tr>
<td>Google+</td>
<td>120,000,000</td>
</tr>
<tr>
<td>Tumblr</td>
<td>110,000,000</td>
</tr>
<tr>
<td>Instagram</td>
<td>100,000,000</td>
</tr>
<tr>
<td>VK</td>
<td>80,000,000</td>
</tr>
<tr>
<td>Flickr</td>
<td>65,000,000</td>
</tr>
<tr>
<td>Vine</td>
<td>42,000,000</td>
</tr>
<tr>
<td>Meetup</td>
<td>40,000,000</td>
</tr>
<tr>
<td>Tagged</td>
<td>38,000,000</td>
</tr>
<tr>
<td>Ask.fm</td>
<td>37,000,000</td>
</tr>
<tr>
<td>MeetMe</td>
<td>15,500,000</td>
</tr>
<tr>
<td>ClassMates</td>
<td>15,000,000</td>
</tr>
</tbody>
</table>


**Uses of Opinion Mining**

Opinion mining and sentiment analysis deal with polarity detection and recognizing of emotion in user generated text. For the purpose of extracting opinions in unstructured information in web pages the opinion mining field uses data mining algorithms and natural language processing techniques (NLP).

The process of opinion mining implies some problems specific to natural language processing like dealing with semantic and syntactical rules of languages, also common are negation handling and word sense disambiguation. But unlike the entire NLP domain, opinion mining deals only with limited portions of these specific issues due to the target of analysis, determining the entities or subjects the contextual opinions refers to and the polarity (positive or negative) of the sentiments regarding those entities.4

The opinion mining analysis use increased in last few years. Today, many companies use in their marketing studies opinion mining and sentiment analysis. Some of these uses are:

---

Opinion Mining Platform for Intelligence in Business

- Creating opinion aggregation and review websites;
- Creating recommendation systems;
- Studies about impact of new product development;
- Making studies about customers;
- Study perception of population regarding governmental policies;
- Improving of anti-spam systems;
- Market research about brand awareness.

Due to the massive quantity of user generated content on different online web platforms it is impossible for one to read all of them or make an opinion by summarizing parts of content. So to assist consumers, one trend is development of websites that automatically classifies and catalogue user reviews (opinion aggregation websites). An example of this type of web platform is recommendation sites, in which consumers can make an opinion about specific products they want to buy.

Companies are interested how their images is reflected on customers opinion, also, market perception on existing products or on release of new products is very important, so they invest in research for development of BI systems that use for their market analysis data from semi-structured or unstructured sources, besides corporate data (see picture below).

![Fig. 4. Use of unstructured data for BI analysis](image)

Source: Rafael Berlanga Llavori, María José Aramburu, Dolores Maria Llidó, Lisette García-Moya: Towards a Semantic Data Infrastructure for Social Business Intelligence. ADBIS (2) 2013: 319-327)

Opinion mining helps companies improve customer relationship management and make studies about attitude towards their brand. The knowledge extracted from web data helps them in effectuating predictions and assist managers in taking educated decisions.

Another important application of opinion mining is in assisting sociologists in studying population opinion. Making sociological studies implies using substantial funds for making polls. To be accurate these polls should be based on a large population. Opinion mining helps obtain results using substantial less effort and could target bigger masses of population.

Determining the opinion regarding adopted policies, or studying the impact of specific laws, or reform on population, could help government of a country to make better decisions or promote policies agreed by masses.

Also opinion mining, used in security systems, could help increase the accuracy of detecting spams in antispam systems.
Classification of Opinion Mining Process

By analyzing the methodology used to perform opinion mining we can classify opinion mining process in these categories:

- Keywords based;
- Lexical based;
- Use of statistical methods;
- Ontologies based methods.

The keywords based method is the simplest procedure and due to its simplicity is widely adopted. It is based on detecting specific words that have a strong sentiment polarity. In some studies regarding this method there were created some classes of opinionated words, or categories of affected words named Affective Lexicon. The main disadvantages of this procedure are the inability to treat correctly negations in text and meaning of a phrase rather than treating only composing words.

The lexical based method besides using a lexicon as in keyword based method, also assign to words a probability of indicating a positive or negative opinion. The assign of probability is made using a linguistic corpora. The results of this method are more accurate that of the above one, but there are also disadvantages, mainly because this method evaluate also on word level, and sentiment probability of a word is domain dependent.

The statistical method use machine learning algorithms as support vector machines (SVM) and Bayesian algorithms for classification and necessitate a training set. Usually the training set is composed by manually affect annotated text. The method has the advantage that also considers the words, or group of words apparition frequency, besides the opinion polarity of particular ones. The main disadvantage is, it necessitates a large training set for efficient classification.

Ontology based method rely on a knowledge base (ontology or semantic network) to make semantic associations, rather than base on simple words. This method can take into account expression that are related to concepts that have sentiment, even if is not explicit shown. The accuracy is dependent on the dimension and depth of knowledge base used, and is strongly domain dependent.

Proposed platform architecture

The proposed architecture is an online platform to extract and analyze the content of web pages on the Internet, for detecting opinions. Such a platform is useful for a company to monitor customer opinions and study market trends.

The objectives of the proposed system are:

- Visit and extract the content of web pages using a web spider;
- Parsing the contents extracted to obtain relevant information;

---

8 E. Cambria, T. Mazzocco, and A. Hussain, “Application of Multi-Dimensional Scaling and Artificial Neural Networks for Biologically Inspired Opinion Mining,” Biologically Inspired Cognitive Architectures, vol. 4, 2013, pp. 41–53
o Determining opinions from the extracted content;
o Determine the polarity of opinions expressed;
o Present the results of process graphical as modern business intelligence solutions.

![Platform architecture](source: made by the author)

Comments extracted from web pages and parsed using the extraction module, shall be subject to Opinion mining analysis, using assessment and classification module. The design of this module is shown in Figure 5. The proposed opinion mining process used in the platform is an unsupervised one, using a lexical resource.

The platform performs sentiment analysis at both the review and sentence level. Evaluation of comments and identification of opinions is performed using the SentiWordNet\(^{10}\) lexical dictionary as knowledge base.

SentiWordNet is a lexical resource used in opinion mining, which assigns to a set of synonyms, (called sysnet) from lexical base WordNet\(^{11}\) for English language, three numerical values: \(obj\) (objectivity of a word), \(pos\) (represents positivity) and \(neg\) (score negative meaning of a word). This lexical resource was achieved by automatic annotation of all sysnets in WordNet 3.0, consistent with notions of positivity, negativity and neutrality.

\(^{10}\) [http://sentiwordnet.isti.cnr.it/](http://sentiwordnet.isti.cnr.it/)
\(^{11}\) [http://wordnet.princeton.edu/](http://wordnet.princeton.edu/)
Classification of comments is done according to three classes (positive, negative and neutral) applying sentiment analysis at the document level. At the same time a classification and extraction of opinions at the level of sentences is performed. In this process are identified entities on which opinions are expressed in the comment. Also the process evaluates a set of domain-specific aspects initially pre-determined. Based on these initial aspects, the platform provides reports on the dominant sentiment of the users or customers, on a set of specifications, for a product or service in review.

To perform sentiment analysis process, the proposed system goes through several stages as described below:

- dividing comments into sentences;
- unique identification of each word in a comment;
- tagging parts of speech in the sentence, and store word order;
- polarity identification of words using a lexical dictionary;
- identifying words on which opinions are expressed;
- determine dominant sentiment of words regarding certain aspects;
- sentences classification;
- comments classification.

The platform uses SentiWordNet version 3.0 (SentiWordNet_3.0.0_20120206.txt). This lexical resource, originally in text version, was converted into a MySQL database table with the following fields: ID, term and score. For each word stored in the term field, it was calculated a value stored in the field score, according to the original values of positivity and negativity. The score is the highest of values, and is negative if the prevailing sense is negative or positive if positivity prevails. Depending on these values, the proposed classification module calculates...
positive, negative or neutral opinion of a comment, sentence or aspect, after some predetermined thresholds.

Conclusions

The need to obtain information in a short time and detailed is imperative for companies. Business intelligence systems that obtain information from corporate data warehouses are no longer sufficient, urging the need to integrate external data (e.g., Market information) in these systems. World Wide Web, the largest available database, contains an enormous amount of relevant information. Difficulty accessing this information comes from many forms of presentation and the fact that the primary goal was accessing information from human users.

The domain of opinion mining this article discuss is relatively new, and was chosen due to the growing interest in the development of Internet, web technologies and applications based on them. These applications based on extracting knowledge from web, have expanded in the last decade in many areas of economic environment and beyond. The proposed platform cover several research areas: Web technologies, algorithms for crawling, semantic web, business intelligence systems, opinion mining and sentiment analysis.

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