AN APPROACH FOR AUTOMATIC ANALYSIS OF ONLINE STORE PRODUCT AND SERVICES REVIEWS

Snezhana SULOVA

Abstract

One of the advantages of e-commerce systems is that they enable customers and merchants to become acquainted with product and services reviews. Currently in the most popular online stores there are hundreds and even thousands of reviews for certain goods, which contain valuable information about the quality of the offered assortment. This is the reason to look for ways for their computer processing. The article proposes an approach for automated analysis of customer reviews, based on natural language processing technology and application of methods of machine learning. A model for analysis and its implementation with the software product RapidMiner are proposed.

Introduction

In recent years Internet was established as one of the richest and most easily accessible sources of information. The global network has a large amount of documents, data, audio and video files, many recorded customer reviews. All these resources are carriers of knowledge about business and after appropriate computer processing they can contribute more detailed analyses and help to identify and explore new relationships.

In the sphere of e-commerce, core business activities are carried out through dynamic online systems. One of the main challenges for this type of business is making fast and accurate decisions in accordance with the changes in the market environment. The e-commerce systems generate detailed and varied reports which are based mostly on statistical processing of the data stored in the database. Lately for more detailed and in-depth analysis in this area there have been used intelligent business analysis based on both structured and unstructured data.

Practice has proven that nowadays new customers of online stores largely rely on the opinions posted from existing customers. Besides, manufacturers and service providers are also interested in analyzing customers' opinions to improve the quality.
and standards of products and services. All this requires the search for new and effective ways to transform unstructured data, such as customer opinions in detailed reports and analyses.

The purpose of this article is to propose an approach for automated analysis of online store product reviews, based on a study of existing technologies for natural language processing.

I. Theoretical foundations of computer technology for natural language processing

The concept of natural language processing (NLP) is a broad term that can be viewed as a synthesis between artificial intelligence and computational linguistics. It is “more than simple machine translation. It aims at the full understanding of the text, checking syntactic and semantic validity of linguistic input, using the real world knowledge to understand the participant’s goals and beliefs, and also speech acts, conversations and discourse structure.” (Kumar, 2011, p.4). Currently many researchers explore different aspects of intelligent text processing. In general, knowledge discovery in unstructured data in the literature is known as text mining (TM) (Fayyad, Piatetsky-Shapiro and Smith, 1996; Feldman, Sanger, 2007). This process is accomplished through the application of technology for data mining (DM) on unstructured text data. “Typical text mining tasks include text categorization, text clustering, concept/entity extraction, production of granular taxonomies, sentiment analysis, document summarization, and entity relation modeling” (Pena-Ayala, 2014, p.37).

The accumulation of more and more information on the Web becomes a prerequisite for extracting knowledge from Internet sources such as web pages. A new concept of extracting knowledge from web resources - web mining (WM) is born. Etzioni first used the term and defined it as “use of data mining techniques to automatically discover and extract information from World Wide Web documents and services (e.g., on-line travel agents, job listings, electronic malls, etc.)” (Etzioni, 1996, p. 65). Later the concept WM expanded and it now includes techniques for testing and analyzing data on the usability of the web resource. (Cooley, Mobasher, Srivastave, 1997; Markov, Larosed, 2007). Web mining is commonly divided into the following three sub-areas (Cooley, Mobasher, Srivastave, 1997):

- web content mining (WCM) – extracting useful knowledge from the contents of web documents;
- web structured mining (WSM) – extracting useful knowledge based on the structure of web sites;
• web usage mining (WUM) – extracting useful knowledge from data on the use of Internet resources.

Many researchers deal with WCM problems (Kosala, Blockeel, 2000; Navadiya, Patel, 2012; Markov, Larosed, 2007). The researches are differentiated depending on the specific research tasks and type of resources that are used. Automatic classification of web pages appears in scientific publications (Materna, 2008), grouping of documents (Markov, Larosed, 2007), detection of similarity between text documents (Huang, 2008; Lakshmi, 2013), extracting opinions from text and sentiment analysis (Liu 2012; Medhat, Hassan, Korashy, 2014; D’Avanzo, Pilato, 2015; Patel, Prabhu and Bhowmick, 2015).

In recent years, mainly thanks to the development of web applications and social networks, the Internet has accumulated a large amount of customer reviews, shared impressions, feelings, emotions. This is the reason for many researchers to focus their research on two interrelated areas such as: opinion mining (OM) and sentiment analysis (SA) (fig. 1).

![Fig. 1 Web Mining, Opinion mining and Sentiment analysis](image)

The term “opinion mining” is introduced by Dave, Lawrence and Pennock (2003). They define it as “a set of search results for a given item, generating a list of product attributes (quality features, etc.) and aggregating opinions about each of them (poor, mixed, good).” After that for opinion mining there are more comprehensive interpretations that involve different types of analysis and evaluation of text. Sadegh, Ibrahim and Othman say that these are techniques to detect and extract subjective
information in text documents. Opinion mining based on technology of natural language processing and focusing on determining the perceptions, views, ideas on a given subject, and automatic analysis aims to extract attributes, characteristics of the objects and on their basis to determine whether comments are positive, negative or neutral.

Sentiment analysis was mentioned for the first time in the works of Das, Chen and Tong (2001). They use the term in automatic analysis and evaluation of text. Subsequently plenty of studies discuss various aspects of the sentiment analysis and in many of them the term is used as a synonym for opinion mining, because emotions moods are extracted from the recorded comments (Pang, Lee, 2008; Sadegh, Ibrahim, Othman 2012; Ankitkumar, Badre, Kinikar, 2014; Rahmath, 2014).

To analyze sentiments in a text Liu (2012) makes a comprehensive study and divides methods depending on parts for analysis - a word, a phrase or the entire document. Document level sentiment analysis attempts to classify the general position in the document as positive, negative or neutral. It helps to establish the overall polarity of the text, but through it one cannot assess the different emotions about certain aspects of the document. Both supervised and unsupervised learning methods can be used for the document level classification (Varghese, Jayasree, 2013). In the sentence level sentiment analysis the polarity of each sentence is calculated using the same methods. Many authors also work on the problems of the word or sentence level sentiment analysis. In literature there are developments related to the grouping of words in dictionaries and lexicons, for automatic classification of words, extraction of keywords and identifying topics and others (Hu and Liu, 2004; Zhang, et. al. 2008; Heerschop, et. al. 2011; Taboada, et. al 2011).

One of the most comprehensive surveys for opinion mining is made by Ravi K. and Ravi V. (2015). They classify existing studies and methods used in them. According to them SA solve problems related to: the classification of text to its subjectivity; sentiment classification; review usefulness measurement; spam texts detection; dictionaries and lexicons creation; aspect extraction. In the presented systematization we cannot make a clear distinction between the first and second groups so we believe it is more appropriate SA tasks to be presented as shown in Fig. 2.
Fig. 2. Main tasks solved with sentiment analysis

The subject areas of application of opinion mining and sentiment analysis are defined well by Rahmat (2014). This type of analysis is mainly used:

- by customers of online stores when they make a decision to purchase a product or service;
- by manufacturers to improve the quality of goods;
- by marketers for research and analysis of moods and sentiments of customers;
- by analytical discussion groups, forums, blogs and social medias
- in detecting spam content on the web;
- in detecting view points on certain topics.

II. Approach to automated analysis of online store customer reviews

As mentioned already in the aim of our study, we will focus our research on analysis of customer reviews, recorded in online stores. The e-commerce systems are dynamic web based applications and they enable interactivity and communication with customers. Recorded customers’ opinions are one of the best measures of whether the e-shop service is good and what shoppers like or dislike, whether
products and services meet the descriptions and presentations, what else customers want to discover in the online store, and what the general clients’ assessments are. Opinion mining and sentiment analysis are important for traders, because they create prerequisites for individual marketing to each customer and to implement better service.

In this article for research and analysis of customers' reviews we suggest the use of the methods of classification, first to distinguish the views of various characteristics of the goods and then to evaluate the polarity of customer reviews about them. The model that we use for analysis is shown in fig. 3.

Fig. 3. Model for analysis of online store product and services reviews

Usually the process of opinion mining from text data is an unconventional task, because data is unstructured, it’s based on WCM and it is appropriate to analyze online customer reviews following these steps:
1. Collecting and recording product reviews.
2. Text preprocessing of product reviews.

The collection of customer reviews is a feature that is integrated into most e-commerce systems. In some systems the views we recorded directly in a web page in html format, but in others they are recorded in the database. However, regardless of the storage, reviews are unstructured data - text.

Text preprocessing is the process of preparing text for application data mining methods such as: classification, clustering and others. Text documents usually have a large amount of words that are not carriers of useful information and it is not appropriate so analyze all words. Recommended preprocessing steps for customer reviews are:

- tokenization – separating the full text review into a list of words;
- lemmatization – the process of bringing a word into its non-inflected dictionary form;
- stop word removal – these are auxiliary words that carry little information about the content of the text, such as: “for”, ”after“, ”so”, “later”, “back”, “against”;
- text capitalization - transformation of characters into lowercase.

The methods that can be used in text sentiment analysis are well summarized by Medhat et. al. (2014) (fig. 4).

Many of the cited surveys show that it is appropriate to use the methods of supervised machine learning for opinion mining and sentiment analysis for online store customer reviews, of which the most successfully applied methods for categorization of text are: Support Vector Machines (SVM) and Naive Bayes (NB) classifier (Singh, Husain, 2014; Verma, Kiranjyoti, 2015).

Many of the aforementioned studies show that for sentiment analysis for customer reviews for the products and services it is appropriate to use the methods of machine learning - supervised learning, of which the most successfully applied for categorization of text are linear classifier Support Vector Machines (SVM) classifier based on the algorithm Naive Bayes (NB) (Singh, Husain, 2014; Verma, Kiranjyoti, 2015).
Fig. 4. Sentiment analysis methods (Medhat, Hassan, Korashy, 2014)

SVM is an algorithm that builds a model based on a given set of training examples and assigns new examples into one category or another, making it a non-probabilistic binary linear classifier. In classification based on the NB algorithm, belonging to a class is calculated based on the formula for finding the conditional probability of assigning the object to a class. Both methods are suitable for creating and implementing a model for classifying customer reviews into two categories – positive and negative.

To test the automatic classification based on the SVM method we have used one of the most popular tools for data mining and predictive analysis – RapidMiner. It provides interactive graphical user interface and tools for links analysis, analysis of unstructured text and log files. The process of extracting knowledge is realized through the build operators. This software provides more than 400 operators for analyzing data, including classification, clustering, and association analysis.

The data used for testing the proposed approach are online reviews collected from the Amazon.com website for tablets of brand Apple. Training data is a very important element in the realization of machine learning, because validation of the model is based on it. The created Rapid Miner model shows the classification of customer views on polarity (Fig. 5).
Fig. 5. Classification customer reviews via SVM method

Based on the pattern, several posts are successfully tested and the result shows that from the analyzed 6 sample reviews only one was rated as negative (Fig. 6).

<table>
<thead>
<tr>
<th>Row No.</th>
<th>label</th>
<th>prediction(...)</th>
<th>confidence(neg)</th>
<th>confidence(pos)</th>
<th>metadata</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>reviews</td>
<td>pos</td>
<td>0.467</td>
<td>0.533</td>
<td>new1.txt</td>
</tr>
<tr>
<td>2</td>
<td>reviews</td>
<td>pos</td>
<td>0.473</td>
<td>0.527</td>
<td>new2.txt</td>
</tr>
<tr>
<td>3</td>
<td>reviews</td>
<td>pos</td>
<td>0.383</td>
<td>0.617</td>
<td>new3.txt</td>
</tr>
<tr>
<td>4</td>
<td>reviews</td>
<td>neg</td>
<td>0.532</td>
<td>0.468</td>
<td>new4.txt</td>
</tr>
<tr>
<td>5</td>
<td>reviews</td>
<td>pos</td>
<td>0.441</td>
<td>0.559</td>
<td>new5.txt</td>
</tr>
<tr>
<td>6</td>
<td>reviews</td>
<td>pos</td>
<td>0.295</td>
<td>0.705</td>
<td>new6.txt</td>
</tr>
</tbody>
</table>

Fig. 6. Result of an analysis of posts by RapidMiner
In practice, this is the last stage of the analysis of reviews. It shows the results and therefore the used software tools and the role of human interpretation of the results are important.

Similarly, the model that is based on the NB algorithm could be built. We made approbation and found that they obtain similar results.

**Conclusion**

The rapid development of social networking and sharing capabilities that provide many of the applications running on the Internet is a prerequisite for the generation of large collections of consumer reviews, impressions, shared feelings and emotions. Intelligent business analyses of these customer reviews is important to the business and therefore is subject to a research interest in the recent years. Since there is no specific algorithm available for viable search for knowledge in a text, based on the results of existing studies in this paper we propose an approach for analyzing the reviews of online stores’ customers, through which expressed opinions can be classified and conclusions about the quality of goods can be made.

The resulting new knowledge could help to improve the product range and customer satisfaction, and for e-commerce companies it is essential, because sales revenues largely depend on it. Furthermore, this kind of analysis can be used by managers to create successful business strategies based on the resulting in-depth and precise analyses and forecasts. Extracting new knowledge from Internet resources could be an important competitive advantage for companies involved in e-commerce, because in general it contributes to improving their business.

**References**


