ECAI 2015 - International Conference – 7th Edition
Electronics, Computers and Artificial Intelligence
25 June -27 June, 2015, Bucharest, ROMÂNIA

Sentiment analysis from product reviews using SentiWordNet as lexical resource

Alexandra Cernian, Valentin Sgarciu, Bogdan Martin
[*] University Politehnica of Bucharest, Faculty of Automatic Control and Computers, Bucharest, Romania
Alexandra.cernian@aii.pub.ro, vsgarciu@aii.pub.ro

***Abstract*— In the current social, technological and economic context, customers make their decisions based mostly on the opinion of other consumers. On the other side, companies need quick feedback from their customers in order to adapt to their needs in real time. The effective connection between these two aspects relies on opinion mining tools, which automatically process consumers' reviews and opinions about products or services. This paper presents a semantic approach for a sentiment analysis application, which is based on using the SentiWordNet lexical resource. The experimental validation proved a 61% average rate of success of the application.**

Keywords-sentiment analysis; opinion mining; SentiWordNet; lexical resource; product reviews

I. INTRODUCTION

Nowadays, customers make their decisions based mostly on the opinion of other consumers. A study published by KRC Research [1] revealed that 65% of electronic devices consumers are influenced by reviews when selecting brands and products. On average, they read 11 reviews before making a choice. Another interesting aspect revealed by the survey is that people's interest in other consumers' reviews is around 77%, whereas only 23% are influenced by an editorialist's opinion. According to surveys [1] and [2], the main aspects regarding customer reviews are the following: correct and reasonable (32%), well written (27%) and contain statistics, specifications and technical data (25%).

For a company, whether it offers products or services, it is very important to get feedback from clients or prospective clients. Feedback is a very powerful decision making instruments in this context, and it provides significant suggestions regarding products improvement or adapting marketing strategies.

The current trend is to create automatic tools to process consumers' reviews and opinions about products or services. That way, customer feedback quickly reaches other consumers or the company itself in a synthetic form, thus facilitating the decision making process.

This paper presents a sentiment analysis approach based on the SentiWordNet lexical resource. The experimental validation was conducted on a set of 300 product reviews from Amazon.

The rest of the paper is structured as follows: Section 2 presents some related work; Section 3 presents the application design and methodology; Section 4 discusses the results of the experimental validation and Section 5 draws some conclusions about the work presented in this paper.

II. BACKGROUND

An important aspect in opinion mining research has been sentiment analysis [7], which assumes developing automatic tools for extracting sentiments from text data sources. Sentiment analysis can be very helpful for organizations in making predictions about their products or services and establishing future strategies.

Nowadays, sentiment analysis is a powerful tool, providing valuable information in numerous fields of activity:

- marketing and advertising, for extracting the customers' opinion about products and services and effective use of advertising, in order to address the right target
- collaborative systems, in order to detect conflicts or aggressive behavior or to extract feedback
- social media, in order to extract opinions and predict evolutions
- recommender systems, which can provide customized recommendations based on the user's behavioral pattern and preferences
- management, in the decision-making process and setting strategic objectives for companies

Current approaches mainly focus on positive and negative sentiments, also analyzing the sentiment polarity [8]. An exhaustive presentation of the methods and approaches used in opinion mining and sentiment analysis is provided in [7].

One of the recently investigated methods is based on the use of lexical resources in order to provide a semantic analysis of text sources. SentiWordNet [5] is a lexical resource designed for processing and analyzing sentiment information for the English language [3]. Each term is assigned a score for each of the following characteristics: positivity, negativity and objectivity. Section 3 presents the design and architecture of the opinion mining system we propose for analyzing the opinions extracted from product reviews.

978-1-4673-6647-2/15/$31.00 ©2015 IEEE

III. APPLICATION DESIGN

The application is designed in order to process the opinions expressed by users in product reviews, in terms of positivity, negativity and objectivity, according to SentiWordNet classification [3]. The flow of the application consists of two main categories of steps, as shown in Figure 1:

1. Extracting and processing the text sources
2. Extract opinions and compute scores

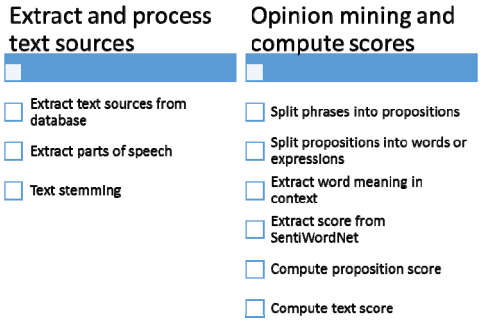

Figure 1. Overview of the application flow

Figure 2 depicts the main components of the application and their integration into its architectural model.

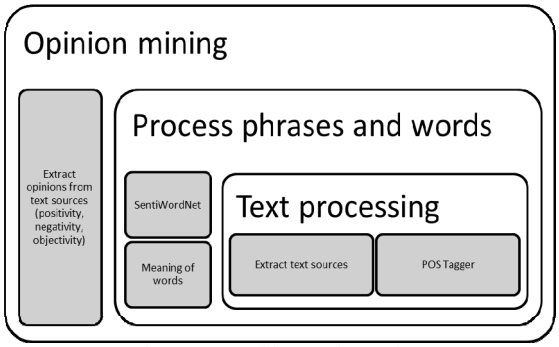

Figure 2. Architectural overview

A. *Parts of Speech Tagging*

A dedicated parts of speech tagging software [4] has been used in order to process products reviews, extracted from Amazon as text sources. This software application is able to read a text source written in English and associate parts of speech, such as noun, adjective, verb and so on, to each word or token. The result obtained at this step of the workflow is a vector containing the words from the product reviews and their corresponding parts of speech.

B. *SentiWordNet Lexicon*

The lexicon used in the hereby presented application is SentiWordNet[3]. SentiWordNet is a public lexical resource designed for the sole purpose of opinion mining and sentiment analysis. It assigns three types of scores to each word: positivity, negativity and objectivity.

It contains 207 000 words and expressions, which are represented in the following format [5]:

- #POS: the part of speech of the word of expression
- ID : unique ID for each part of speech
- PosScore and NegScore: the positivity score, namely the negativity score of the word. If the sum of those two scores is less than 1, the difference is considered the objectivity score of the word (ObjScore). Thus, the following formula determines the score of a specific word or expression from SentiWordNet [5]:

$$ObjScore = 1 - (PosScore + NegScore) \quad (1)$$

- SynsetTerms: The list of synonyms for each word The "#" character marks the most widely used term, in decreasing order
- Gloss: the definition of the word

In order to facilitate the analysis of text sources in our application, we have imported the SentiWordNet lexicon into a MySQL database. There are three main tables storing the information regarding the meaning of the words, their words and synset terms.

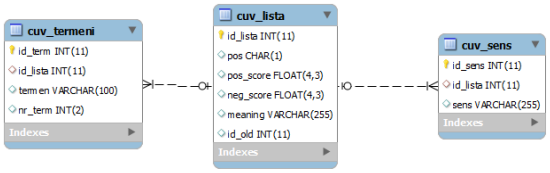

Figure 3. Importing SentiWordNet into database tables

C. *The corpus*

The corpus used for the experimental validation of this work consists of a set of 300 electronic devices product reviews from Amazon [6], dating from June 1995 until March 2013. Reviews on Amazon are rated with 0 to 5 stars.

The dataset is structured according to the following pattern:

- product/productId: the ID of the product, for example: amazon.com/dp/B00006HAXW
- product/title
- product/price:
- review/userId: the ID of the user, for example: A1RSDE90N6RSZF
- review/profileName: the name of the user
- review/helpfulness: users who have rated the review as useful
- review/score: the product rating
- review/time: the date and time of the review
- review/summary: the summary of the review
- review/text: the full text of the review

For the current work, we have used the text of the reviews in order to extract consumers' opinions and the review score, in order to validate the results generated by our application.

D. The methodology

Step 1: Extract product reviews from database.
Step 2: Determine parts of speech using Stanford POS Tagger [6]
Step 3: Split phrases into a vector of sentences.
Step 4: Split sentences into a vector of words.
Step 5: Extract meanings for each word, from SentiWordNet.
Step 6: Choose the correct meaning of each word in context. Thus, we perform an exhaustive search that involves the comparison of words that appear in the current sentence (context) and the different meanings of the words, as stated by the definitions. If the comparison returns a positive score, the ranking of that meaning of a word will be increased by 1.

In order to calculate the score of the comparison between the words that make up the meaning and the words that make up the context we use the following formula:

$$meaning\_score_1 = meaning\_score_1 + 1 \quad (2)$$

To calculate the result of the comparison between the words that make up the definition and the words that make up the context we use the following formula:

$$meaning\_score_2 = meaning\_score_2 + 1 \quad (3)$$

The final score of a meaning in the context of each phrase is calculated using the following formula:

$$score_{meaning} = \frac{meaning\_score_1 + meaning\_score_2}{2} \quad (4)$$

The meaning with the highest score will be associated to each word.

Step 7: Compute the sentence score. The score is calculated based on the score of the words forming the sentence and on a length index associated to each sentence. The shorter the sentence is, the easier it is to determine its meaning. Thus, for sentences with less than 10 words, we associate a length index of 1.4, for sentences having 10 to 15 words, we associate a length index of 1.2, and for sentences with less than 20 words we associate a length index of 1.1. Otherwise, the length index is 1.

The formula for calculating the score of a sentence is the ratio of the sum of the scores of the words in the sentence and the total number of words in the sentence.

$$sentence\_score = \frac{\sum_{i=0}^{n} item\_score(i)}{no\_words} * size\_index \quad (5)$$

Step 8: Compute the text score. To obtain the final result - the polarity of the text - the average score of the sentences is computed using the following formula:

$$text\_score = \frac{\sum_{i=0}^{n} sentence\_score(i)}{no\_of\_sentences} \quad (6)$$

Step 9: Evaluate the results. In order to evaluate the results obtained, we associate each score to a corresponding rating between 1 and 5, in order to match the Amazon ratings.

- Scores between [0.8 -1] determine a rating of 5 points
- Scores between [0.6 – 0.8] determine a rating of 4 points
- Scores between [0.4 – 0.8] determine a rating of 3 points
- Scores between [0.2 – 0.4] determine a rating of 2 points
- Scores between [0 – 0.2] determine a rating of 1 points

IV. EXPERIMENTAL EVALUATION

For the experimental evaluation, we have used a dataset of 300 product reviews from Amazon. The percentage expresses the accuracy of the algorithm in determining the consumers' opinion compared to the rating they have given to products. The score associated to each text source is expressed as a number between 1 and 5. We have analyzed the performance of the algorithm in terms of accuracy and generated a series of statistics in order to emphasize different correlations between specific conditions and the correctitude of the results. Therefore, we are interested in determining the influence of number of words, the number of sentences and the number of characters in the opinion mining process from text sources.

Table 1 illustrates the general statistics regarding the rate of success of the opinion mining method discussed in this paper.

Table 1. Accuracy – general statistics

Nr Crt	Accuracy	Number of reviews
1	75% – 100%	117
2	50% -75%	57
3	25%-50%	73
4	0%-25%	52

The lexicon based opinion mining methodology hereby discussed had an accuracy percentage of 75% to 100% for 117 reviews, as shown in Figure 4.

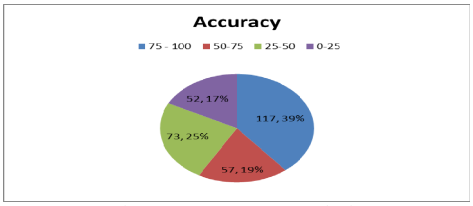

Figure 4. Accuracy analysis

Table 2 presents an analysis regarding the number of words per review, as identified in our dataset.

Table 2. Number of words/review

Number of words	Number of reviews
1-50	116
51-100	96
101-150	43
151-200	30
201-250	7
251-300	4
301-350	2
351-400	2
Grand Total	**300**

Table 3 presents the statistics regarding the correlation between the rate of success of the algorithm and the number of words of the reviews.

Table 3. Number of words/review

Number of reviews / Number of words	Accuracy 0-25	25-50	50-75	75-100	Grand Total
1-50	22	21	27	46	116
51-100	17	26	15	38	96
101-150	7	10	9	16	42
151-200	4	12		14	30
201-250		3	4		7
251-300		1		3	4
301-350	2				2
351-400			2		2
Grand Total	52	73	57	117	299

The accuracy is expressed as percentage in each column of table 3. As noticed, for each category of the classification based on the number of words per review, the accuracy of the algorithm was between 75% and 100% in a significant number of cases.

Figure 4 illustrates the distribution of accuracy according to the number of words classification. For short reviews with less than 50 words, 40% of the number of reviews produced ac accuracy of 75 to 100%, almost double compared to the second group of 50-75% accuracy.

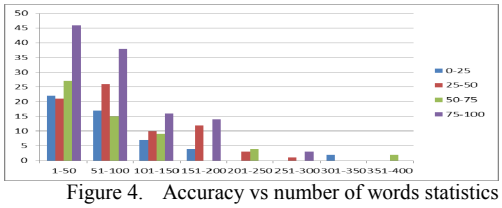

Figure 4. Accuracy vs number of words statistics

An interesting aspect here is the fact that as the number words per review increases, although the 75-100% accuracy remains the highest, the 25-50% accuracy group gets bigger.

If we look at the statistics based on the number of sentences in reviews, here is what we find (Table 4):

Table 4. Number of words/review

Nr. Crt	Accuracy	Number of sentences
1	75 - 100	6
2	50-75	6
3	25-50	5
4	0-25	5

For each interval of accuracy, the average number of sentences in reviews has been computed. The first conclusion was that if we take product reviews into consideration, the average number of sentences used by consumers is not very high, namely around the value of 5 or 6 in our dataset. The accuracy distribution did not vary significantly based on this criterion. Therefore, the semantic content is the most important aspect to influence the process of extracting consumers' opinions.

V. CONCLUSIONS

This paper presented a semantic approach for a sentiment analysis application, which is based on using the SentiWordNet lexical resource. The experimental validation proved a 61% average rate of success of the application, for a set of 300 product reviews from Amazon. For the validation methodology, we compared the results produced by the application with the star ratings from Amazon, in order to get an objective picture of the accuracy of our approach.

As future developments, we plan to optimize the algorithm so that it provides an extended semantic processing of text sources.

ACKNOWLEDGMENT

The work has been funded by the Sectoral Operational Programme Human Resources Development 2007-2013 of the Ministry of European Funds through the Financial Agreement POSDRU/159/1.5/S/132395.

REFERENCES

[1] Weber Shandwick, "Buy it, Try it, Rate it", KSC Research, 2012. http://www.webershandwick.com/uploads/news/files/ReviewsSurveyReportFINAL.pdfs

[2] Survey Zendesk, 2013 http://d16cvnquvjw7pr.cloudfront.net/resources/whitepapers/Zendesk_WP_Customer_Service_and_Business_Results.pdf

[3] S. Baccianella, Aa Esuli, and F Sebastian, "SENTIWORDNET 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining", http://nmis.isti.cnr.it/sebastiani/Publications/LREC10.pdf

[4] Kristina Toutanova, Dan Klein, Christopher Manning, and Yoram Singer, "Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network", In *Proceedings of HLT-NAACL 2003*, pp. 252-259

[5] SentiWordNet: http://sentiwordnet.isti.cnr.it/, last accessed 27.01.2015.

[6] Staford Network Analysis Project: http://snap.stanford.edu/index.html, last accesed 24.01.2015.

[7] Pang B, Lee L. (2008) "Opinion Mining and Sentiment Analysis", Foundations and Trends in Information Retrieval, Vol. 2, Nos. 1-2, pp. 1-135, 2008.

[8] Ohana, Bruno, "Opinion mining with the SentWordNet lexical resource", 2009. Dissertations.