Automatically conducted learning from textually expressed vacationers' opinions

Bartosz Jędrzejec[1,*], and *Krzysztof* Świder[1]

[1]Rzeszów University of Technology, Al. Powstańców Warszawy 12, 35-959-Rzeszów, Poland

Abstract. The automatically conducted consumers' opinions investigation is one of the most interesting potential applications of text analytics. In our study we perform a two steps procedure of learning from the textually expressed reviews concerning hotel services offered by a travel company. In the first stage we accomplish the necessary *Extract-Transform-Load* process utilizing one of the available web portals and required language resources. In the second stage each of the suitably pre-processed opinions is "linguistically evaluated", which results in a vector of numeric indicators characterizing its sentiment.

1 Introduction

1.1 Automated text analysis for consumer investigation

Automated text analysis (*text mining*) utilizes analytical methods to learn from collections of text data, like books, newspapers, emails or Web portals. Over the last two decades, we have observed an explosion of continuously growing consumer-generated content including discussions of products, services, hobbies, or brands. There are thousands of product websites which offer forums for consumer comment. Todays receivers of the information do not only consume the available content on the web, but in turn, actively annotate this content and generate new pieces of information. Within all this information lies knowledge about consumer opinions, decision-making, psychology, and culture that may be useful for consumer investigation [1, 2].

Although there are many ways to incorporate automated text analysis into consumer research, there is not much agreement on the standard set of methods, reporting procedures, steps of data inclusion, exclusion, sampling, and, where applicable, dictionary development and validation. Moreover, in order to provide the users, e.g. specialists in web marketing, with a truly operational tool, it seems to be a natural need to be able to develop text analysis applications not only for English but for any other language as well.

In our research, we developed the two steps procedure for learning from the textually expressed vacationers' opinions written in Polish. Essentially, we focused on *sentiment analysis*, which is the field of study that analyses human attitudes and emotions towards

[*] Corresponding author: bartoszj@prz.edu.pl

entities and their attributes expressed in a written text. In the following subsection we specify some significant features of sentiment analysis.

1.2 Sentiment analysis

Sentiment analysis is the computational study that analyses people's opinions, sentiments, emotions, and attitudes toward entities and their attributes expressed in a written text [3]. The entities can be products, services, organizations, individuals, events, issues, or topics. The problem is increasingly attractive in business and society offering numerous research challenges.

1.2.1 Opinions and sentiments

The main objective in sentiment analysis is typically to design effective algorithms and models to extract *opinions* from natural language text and to summarize them suitably. The term *opinion* can be considered as a broad concept that covers sentiment (evaluation, appraisal, or attitude) and associated information such as opinion target and the person who holds the opinion [3]. More formally, opinion O can be considered as the quintuple

$$O = (Entity, Aspect, Sentiment, Holder, Time).$$

The first element *Entity* denotes the opinion target, i.e. an object which the particular opinion refers to. Sometimes an opinion concerns a specific part or attribute of *Entity* called *Aspect*. Then follow: *Sentiment* which reflects an underlying positive or negative feeling implied by opinion, *Holder* – a person who formulated the opinion, and *Time* representing a moment when an opinion was issued. The simple example in Fig. 1 will perhaps be helpful in further clarification of crucial opinion components.

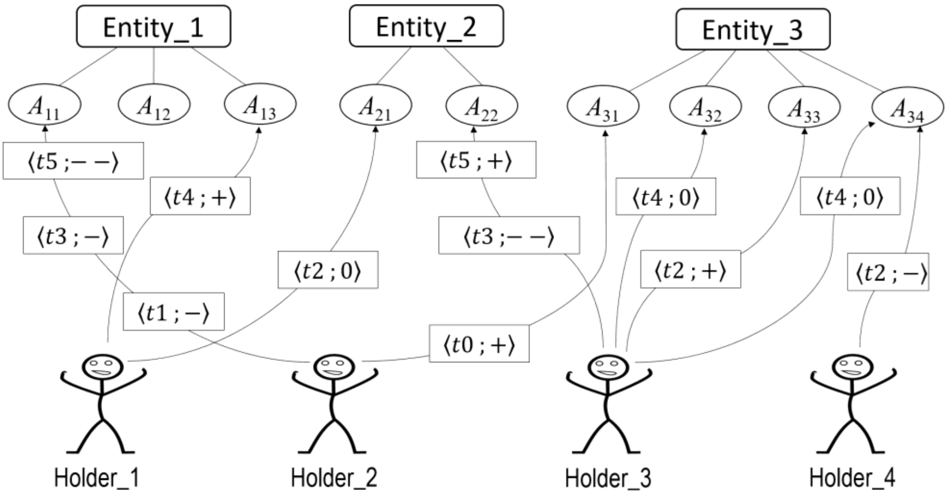

Fig. 1. An intuitive explanation of essential opinion components.

We assume that four persons depicted as *Holder_1, Holder_2*, *Holder_3*, and *Holder_4* are formulating a number of opinions referring to certain aspects of the three objects: *Entity_1*, *Entity_2*, and *Entity_3*. We further presume that the *Entity_1* is described by three aspects denoted by A_{11}, A_{12}, and A_{13}. Likewise the *Entity_2* has two aspects: A_{21} and A_{22}, etc. The fact that a *Holder_i* (i=1…4) has formulated an opinion about some aspect A_{jk} of an *Entity_j* (j=1…3) is represented by the line with an arrow connecting the opinion holder with the relevant aspect. Each opinion line is marked by one or more labels holding information about an opinion posting time and an assigned sentiment value. For example the label (t4 ; +) informs that at the moment t4 a positive opinion was formulated. Similarly (t3 ; – –) denotes a very negative opinion at the moment t3.

As reported in [3], sentiment analysis research has been mainly carried out at three levels of granularity: document level, sentence level, and aspect level. The task at the document level is to classify whether a whole opinion document expresses a positive or negative sentiment. The next level of granularity is to determine whether each *sentence* expresses a positive, negative, or neutral opinion. This level of analysis is closely related to *subjectivity classification*, which distinguishes sentences that express factual information (*objective sentences*) from sentences that express subjective views and opinions (*subjective sentences*). Neither document-level nor sentence-level analyses discover what people like and dislike exactly. In other words, they do not tell what each opinion is about, that is, the target of opinion. To obtain this level of fine-grained results, we need to go to the *aspect* level. Instead of looking at language units (documents, paragraphs, sentences, clauses, or phrases), aspect-level analysis directly looks at an opinion and its target.

1.2.2 Sentiment lexicon

The most important indicators of sentiments are *sentiment words*, also called *opinion words*. For example, *good*, *wonderful*, and *amazing* are positive sentiment words, and *bad*, *poor*, and *terrible* are negative sentiment words. Apart from individual words, there are also phrases and idioms, as for example, *cost an arm and a leg* in English. Sentiment words and phrases are instrumental to sentiment analysis. A list of such words and phrases is called a *sentiment lexicon* or *opinion lexicon*. Over the years, researchers have designed numerous algorithms to compile such lexicons.

In some cases a sentence containing sentiment words may not express any sentiment. This phenomenon happens in several types of sentences. Question (interrogative) sentences and conditional sentences are two main types, for example, "*Can you tell me which hotel in Crete is good*?" and "*If I can find a good hotel in Greece, I will book it*." Both these sentences contain the sentiment word *good*, but neither expresses a positive or negative opinion about any specific hotel.

1.3 Scope of the study

In order to test and introduce the analytical framework developed in this paper we used the real-world textual opinions posted by Polish speaking holidaymakers on a traveling company Web portal. In our work we concentrated on the *Extract-Transform-Load* process which is crucial for the whole analysis. The row data extracted from numerous websites was appropriately pre-processed and placed into regular structures suitable for analysis. Finally we performed tokenization chopping each review up into bigrams and applied a number of measures to get a numerical summarization of each opinion at the document level.

2 Data acquisition

In order to obtain data appropriate for the consecutive analysis we developed and applied a number of automatic procedures ready to accomplish the task.

2.1 Identifying data source

A number of popular social networking sites, such as Facebook and Instagram, as well as many portals, where users can present their opinions about products or services, are very good sources for developing an analytical database for sentiment analysis. In our study we exercised clients' reviews (opinions) presented on the websites of one of the Polish travel agencies referring to holidays spent in hotels offered by the agency. The general structure of the portal is illustrated in Fig. 2.

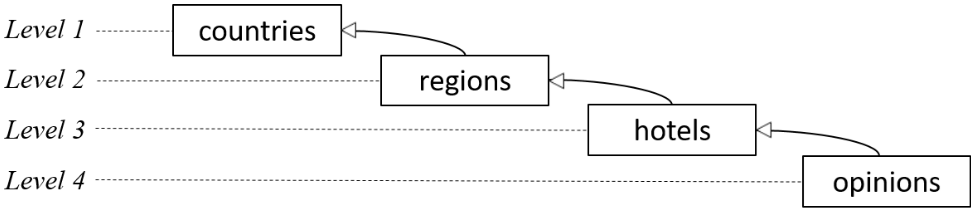

Fig. 2. The overall structure of the travel agency web portal.

Clients have an opportunity of entering the textual comments related to their stay in a hotel, as well as putting an overall numerical rating (1-6) determining the level of satisfaction - with the entire service as well as with selected aspects (room cleanliness, board, animations, etc.). Opinions are published on the website of a particular offer in a separate tab. The reviews are systematised in several levels including: destination (country), region, hotel, and particular opinion.

2.2 Extracting data

In order to create the reviews database i.e. the source repository for forthcoming analysis we used some web data extraction techniques also called *webscraping* (Fig. 3).

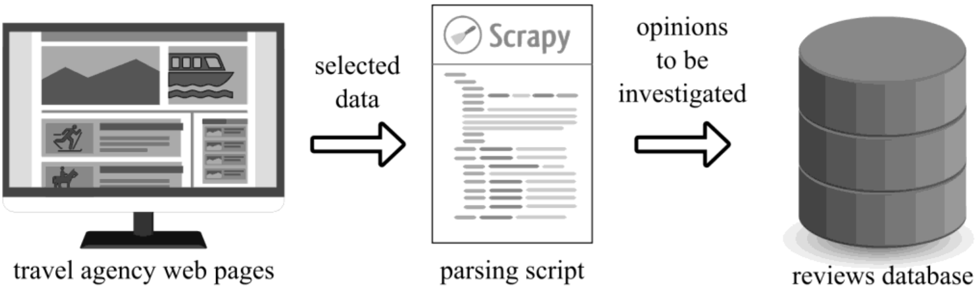

Fig. 3. The data extraction process.

The process of data extraction involves selecting elements related to desired data and parsing particular web pages. The application of this scheme requires a search through the website code consisting of HTML tags which mark relevant data. Very useful tools to accomplish the task are leading web browsers which provide appropriate functionalities and help to investigate even complex websites. After detecting elements responsible for displaying relevant data we need to connect to the websites of the travel agency in order to extract the desired values. For the purposes of this research, a parsing script has been developed to examine the applicable pages. It starts from the page with the list of destinations (countries to spent a holiday in), then deals with different regions in a country, and finally works with the list of hotels in a region. The script was limited to one destination (Greece) and resulted in 38,946 records containing information about hotels, regions, text opinions, and numerical ratings.

To implement the parsing script we used the Python programming language. It provides a number of programming libraries to support the website processing, from which, possibly the most popular one called *Scrapy*, was selected. The library is an open source product for creating website parsers, downloading data and storing it in a structured form. A properly prepared parser is able to process multiple sub-pages, at the same time significantly improving the preparation of the resulting database. The obtained data were saved in a CSV (*comma-separated values*) file, which is a popular format used both in analytical software and in processing with use of programming scripts.

3 Data transformation

Collecting data from external sources is usually only the first step in preparing them for analysis. Such data, especially when entered “manually" by users, is often not suitable for the construction of data mining models and needs to be modified accordingly to the requirements of the method chosen.

The most common operations on numerical data are: standardisation, discretization, identification of erroneous or empty values, etc. The pre-processing of textual data, such as opinions, requires preliminary operations different from those for numerical values. Typical text operations performed in this case are: deleting punctuation marks or case-sensitive unification. In addition, texts in languages other than English may require diacritical character conversion or the use of a basic form of a word. Our reviews database has been pre-processed in a few steps shown in Fig. 4.

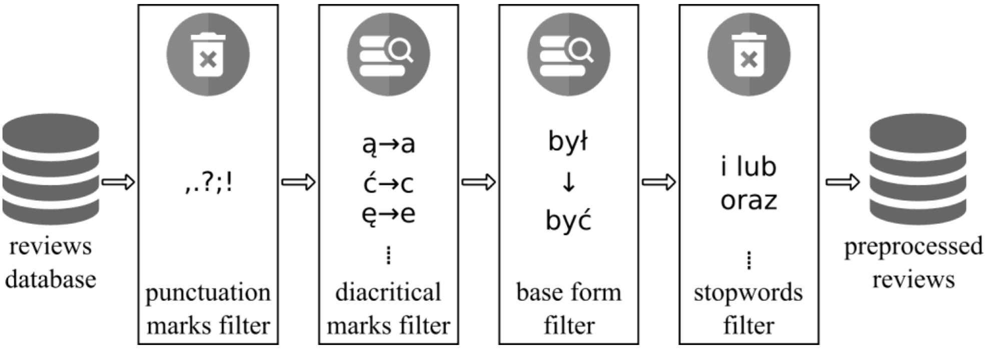

Fig. 4. The data pre-processing operations.

3.1 Removing punctuation and diacritical marks

At the first stage of opinion processing, all punctuation marks were removed from the text. This operation was performed in two sub-stages. First the punctuation marks were replaced by a space character, and then all the excess space marks were replaced by one. In this way, the problem is avoided in the case when the reviewer has forgotten to enter a space character after the punctuation mark, and deleting it would result in two words being merged into one.

The preliminary inspection of the source data was sufficient to discover that some text-based opinions were completed without Polish diacritical marks. This inconsistency needed to be unified so that all functions would work properly regardless of whether an author of any particular review used or not used the language specific characters. That is why we have decided to change the Polish signs to their equivalents without diacritical marks. To this end a necessary Python script was developed, which automatically modified the textually expressed opinions.

3.2 Transforming words into elementary form

Similarly to many other languages, together with the basic word form in Polish there is a number of derivative forms related to the flection. In the text analysis this could cause problems because as the same word can appear in multiple forms and at the same time should be identified as one unit. It is therefore important that the various forms that appear in the text be converted into their basic forms. For this purpose, a freely available dictionary of Polish words with their flection [4] was used. It is a collection created by hobbyists and used for spell-checkers and word games. The flection dictionary was used both to process textual opinions and to prepare a sentiment dictionary which included only basic forms with their emotional marks.

3.3 Eliminating stop-words

There exists a group of words which very often appear in sentences, but do not bring with them any content. This group includes, among others, conjunctions, most grammatical particles, numbers, etc. Such words are often omitted in the process of analysing texts although in the field of sentiment analysis, for example, some of them may have an impact on the emotional image of an utterance. Words that should not be removed in this case include "*not*", which reverses the meaning of the word as in *"not the best*", or "*very*", which enhances an emotional impression, such as *"very beautiful*".

4 Sentiment evaluation

The common and fundamental goal of sentiment analysis of free-text documents is to assign a predefined sentiment label as "positive" or "negative" to each document. Having collected and prepared textual data in section 3 we scheduled and programmed a number of operations aimed at sentiment evaluation and applied them for each of the pre-processed reviews. The result was the list of numerical values describing the review (see Fig. 5).

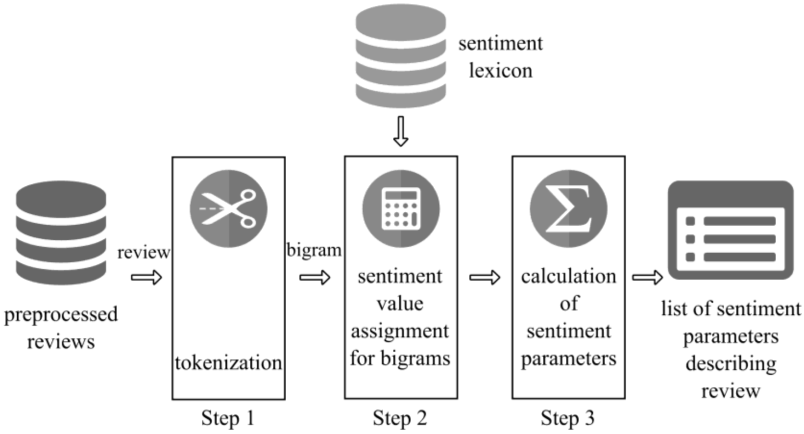
sentiment lexicon
preprocessed reviews
review
tokenization
bigram
sentiment value assignment for bigrams
calculation of sentiment parameters
list of sentiment parameters describing review
Step 1
Step 2
Step 3

Fig. 5. The main steps of sentiment evaluation.

One of the key components of the schema in Fig. 5 is the sentiment lexicon which consists of sentiment words. As indicated in section 1, sentiment words and phrases are helpful for sentiment analysis. In our project we used the plWordNet 3.0 database [5, 6] which is a semantical vocabulary reflecting the lexical system of the Polish language. For a number of words it provides information about their emotional divergence as: fully negative (-m), slightly negative (-s), neutral (0), slightly positive (+s), and fully positive (+m). The plWordNet 3.0 provided as XML document was used as an input data to form our sentiment lexicon. To accomplish this another Python script was developed and applied. Additionally, in order to include more words we involved another repository introduced in [7]. Merging this two resources resulted in a sentiment lexicon containing 27472 words in their basic form with sentiment marks assigned.

The sentiment parameters summarizing each review were calculated by applying a number of operations forming the three steps in Fig. 5. We shortly explain them in the following subsections.

4.1 Document tokenization

A *token* is an instance of a sequence of characters in some particular document that are grouped together as a useful semantic unit for processing [8]. Thus tokenization in Fig. 5 means chopping each review up into pieces called *tokens*, perhaps at the same time throwing away certain characters.

The major question of the tokenization phase is what are the correct tokens to be used for sentiment analysis. The important disadvantage of using separate words as tokens is that in such a case the negations such as “not bad” and “not good” will be not properly treated during sentiment evaluation. In order to avoid the handling of only single words in a review we decided to use *bigrams* [9-12] which are generally tokens consisting of two consecutive words of the document. We assume, that the application of bigrams will increase the overall accuracy of sentiment evaluation.

4.2 Assigning sentiment to bigrams

In the next step of sentiment evaluation we use the sentiment lexicon to check the sentiment mark of the second word of each bigram. The involvement of bigrams makes it possible to detect such combinations of words in a review where the first word effects the sentiment value of the second one. For example, if we consider “*beautiful*” as the second word in a bigram, the typical cases of such influence are: strengthening (“*very beautiful*”) or negation (“*not beautiful*”). If the second word of a bigram in not recorded in lexicon, the bigram is ignored.

4.3 Computing sentiment indicators at the document level

As mentioned in section 1, the document-level sentiment analysis aims to classify an opinion document (e.g. a product review) as expressing a positive or a negative attitude (sentiment). Following the approach we consider each document as a whole and do not study entities or aspects inside the document or determine sentiments expressed about them. Instead, in order to get more information about sentiment of each individual review a number of numerical values were calculated for this review through the third step in Fig. 5. The calculations were performed for each document (opinion) and resulted in 16 numerical indicators referring to sentiment. The examples of the calculated values are:

- number of words with positive, negative, and neutral sentiment,

- mean value of sentiment separately for the positive and negative words,
- (number of positive words) / (total number of words in review) ratio, etc.

Document sentiment classification can be considered as a traditional text classification problem with sentiment orientations or polarities as the classes. It is the relative simple sentiment analysis task and any supervised learning algorithms can be applied directly to solve the problem [3]. In such a case our numerical indicators for each individual review appear to be helpful in preparation the necessary training data.

5 Summary

We investigated the real-life textual opinions formulated by Polish speaking vacationers and published on a traveling company Web portal. The special emphasis was given to the *Extract-Transform-Load* process which is traditionally recognized as crucial for data analysis. The row data extracted from numerous websites was properly pre-processed and recorded as regular forms suitable for exploration. Finally, we applied a number of numerical measures to get a summarization of each opinion at the document level.

In our view, sentiment analysis of free-text documents is a common task but most likely it has to be solved separately for any particular language. This concerns first of all such language-specific issues as document pre-processing and sentiment lexicon generation. At the same time we believe that the parsing and pre-processing schema presented and discussed in this study could be applicable and helpful for a number of particular solutions.

References

1. A. Humphreys, R. Jen-Hui Wang, J Consum Res, **44**(6), 1274–1306 (2018)
2. S. Padmaja, S. S. Fatima, Int J Ad Hoc Sensor Ubiq Co, **4**(1), 21–33 (2013)
3. B. Liu, *Sentiment analysis. Mining opinions, Sentiments, and Emotions* (Cambridge University Press, 2015)
4. Zespół SJP, Lista słów z odmianami (*A repository of Polish vocabulary resources available on* https://sjp.pl/slownik/odmiany/) (2018)
5. M. Maziarz, M. Piasecki, E. Rudnicka, S. Szpakowicz, P. Kedzia, *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, 2259-2268 (2016)
6. M. Zasko-Zielinska, M. Piasecki, S. Szpakowicz, *Proceedings of the International Conference Recent Advances in Natural Language Processing*, 721-730 (2015)
7. A. Wawer, D. Rogozinska, *2012 IEEE 12th International Conference on Data Mining Workshops*, 724–730 (2012)
8. C. D. Manning, P. Raghavan, H. Schütze, *An Introduction to Information Retrieval* (Cambridge University Press, 2009)
9. S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, P. Woodland, *HTK Book* (2005)
10. P. Lamere, P. Kwok, E. Gouvea, B. Raj, R. Singh, W. Walker, P. Wolf, *The CMU Sphinx-4 speech recognition system* (2004)
11. E. Whittaker, P. Woodland, Comput Speech Lang, **17**(1), 87–104 (2003)
12. T. Hirsimäki, J. Pylkkönen, M. Kurimo, IEEE T Audio Speech, **17**(4), 724–732 (2009)