Sentiment Analysis of Hotel Reviews in Greek: A Comparison of Unigram Features

George Markopoulos, George Mikros, Anastasia Iliadi, and Michalis Liontos

Abstract Web 2.0 has become a very useful information resource nowadays, as people are strongly inclined to express online their opinion in social media, blogs and review sites. Sentiment analysis aims at classifying documents as positive or negative according to their overall expressed sentiment. In this paper, we create a sentiment classifier applying Support Vector Machines on hotel reviews written in Modern Greek. Using a unigram language model, we compare two different methodologies and the emerging results look very promising.

Keywords Sentiment analysis • Text mining • Information retrieval • Machine learning • Natural language processing

1 Introduction

Whenever people need to make a decision, they usually ask for other people's opinion. Since their decision involves spending time or/and money, what other people think receives great significance.

The appearance of World Wide Web initially and its development later into Web 2.0 changed the existing situation up to this time. While some years ago there were few resources from where one had the option to ask for an opinion (e.g. family, friends, etc.), nowadays through the web a huge amount of data is accessible to everyone. The proliferation of Web 2.0 led to an excess increase of user-generated content as users are now provided with the potentiality to express online their opinion for different events, persons, products or services in blogs, forums, social media and review sites.

G. Markopoulos (✉) • A. Iliadi • M. Liontos
Department of Linguistics, School of Philosophy, University of Athens, 157 84 Zografou, Greece
e-mail: gmarkop@phil.uoa.gr; anasmusicpiano@gmail.com; mixalisl7@hotmail.com

G. Mikros
Department of Italian Language and Literature, University of Athens, School of Philosophy, 157 84 Zografou, Greece
e-mail: gmikros@ill.uoa.gr

V. Katsoni (ed.), *Cultural Tourism in a Digital Era*, Springer Proceedings in Business and Economics, DOI 10.1007/978-3-319-15859-4_31

A significant amount of research has been carried out in recent years into online reviews because they contain rich opinion information. Especially in the case of hotel reviews, exploiting such information can be proved very useful for both customers and service providers in different ways. On the one hand, viewing other people's travel experiences in a given destination or comments about a certain hotel can crucially influence a potential customer in his/her travel planning and booking. Kasper and Vela (2011) highlight the importance of hotel reviews as information sources especially for hotel booking. They mention that "such user reviews are relevant since they are more actual and detailed than reviews found in traditional printed hotel guides etc., they are not biased by marketing considerations as e.g. the hotels' home pages or catalog descriptions and reflects actual experiences of guests" (p. 45). Hotel managers, on the other hand, can readily gather feedback from their costumers concerning what they liked or disliked in their hotels in order to improve the quality of the services provided.

However, the process and manipulation of such information can indeed be a laborious and time-consuming task for humans because of its vastness. Thus, the need for automated opinion detection and extraction systems has led to the emerging field of sentiment analysis. Sentiment analysis provides techniques for the computational study of sentiments and opinions expressed in text utilizing various natural language processing and text analysis tools. Within sentiment analysis, two are among others the basic subtasks: (a) determining the polarity of a given text, i.e. whether it expresses a positive or a negative opinion on a certain topic and (b) identifying whether a given text (usually a sentence) is subjective or objective, i.e. whether it contains or not an expression of opinion.

Classifying a document according to the overall sentiment of its content is perhaps the most widely studied problem in the academic community nowadays. The greater part of the research in sentiment analysis has been focused on online texts written in English and especially on movie and product reviews and, thus, the literature on other languages and domains is rather limited. A typical example of a very challenging domain which has gained little research attention is the hotel domain.

Motivated by this observation and given that tourism is a very popular industry in Greece, we decided to examine hotel reviews, and by doing so we developed a prototype for predicting sentiment polarity in hotel reviews written in Modern Greek. Using unigram language modeling, we trained a machine learning algorithm following two different methodologies: (a) the frequency of individual words using the TF-IDF weighting scheme and (b) the occurrence of selected polarity words. The results of our study show that we can classify reviews written in Greek based on their sentiment polarity in a considerably efficient manner.

2 Literary Review

In recent years, there have been a large number of studies on sentiment-based classification. The approaches adopted by researchers could be grouped in two main categories: machine learning and semantic orientation approaches. The machine learning approach is a supervised task as it involves the training of a classifier using a collection of representative data. On the other hand, the semantic orientation approach involves the determination of the document's overall sentiment from the semantic orientation of words it contains without prior training and, thus, it is an unsupervised method.

Chaovalit and Zhou (2005) compare the two aforementioned methods using reviews from the movie domain. The results show that the unsupervised semantic orientation approach achieves low accuracy, but is much more efficient when used in real-time applications. In contrast, the supervised machine learning approach provides more accurate classification results but has the drawback that the training of the classifier tends to be very time-consuming. On account of this, researchers many times apply unsupervised techniques in order to label a corpus which is later used for supervised learning.

Most of the early work within sentiment classification used words as the processing unit. Hatzivassiloglou and McKeown (1997) propose a method that automatically determines the semantic orientation of adjectives. They utilize the use of conjunctions between adjectives in order to extract information indirectly from the corpus. More specific, when two adjectives are linked by conjunctions such as 'and', they are probably of the same orientation (e.g. *fair and legitimate*, * *fair and brutal*), while, when they are linked by 'but' or other similar conjunctions, they have different orientation (e.g. *fair but brutal*). Using these constraints, combined with supplementary morphological rules, they achieve 82 % accuracy in predicting whether two conjoined adjectives are of the same or different orientation.

Closely related to the previous work is the method presented by Turney (2002). In this study, given that adjectives and adverbs are considered good indicators of opinions, two-word phrase patterns containing these categories were extracted with the second word providing the context. The semantic orientation of each extracted phrase is then estimated with the *pointwise mutual information* (PMI) measure. Using the NEAR operator of the AltaVista search engine, which constrains the search to documents that contain the words within ten words of one another, he examined whether a phrase has the tendency to co-occur in the context of the word 'excellent' or the word 'poor'. A phrase would have positive or negative semantic orientation if it was strongly associated either with 'excellent' or with 'poor' respectively. Finally, after calculating the average semantic orientation of all extracted phrases in a given review, the review is classified accordingly as recommended or not recommended.

The earliest work in automatic sentiment classification problem using supervised learning at document level has been carried out by Pang et al. (2002). They compare

the performance of three machine learning algorithms (Naive Bayes (NB), Maximum Entropy (MaxEnt) and Support Vector Machines (SVMs)) on a movie review corpus using different features such as unigrams, bigrams, part-of-speech information, position information, etc. The main findings of their study are that: (a) SVMs give better results than other classifiers (82.9 %); (b) unigram presence information is more effective in comparison to unigram frequency and (c) the accuracy in sentiment classification drops when bigrams are used.

A related study is presented by Boiy et al. (2007). They first give an overview of the various techniques that can be used to detect the sentiment of a text and later they compare the performance of SVMs, MaxEnt and NB on Pang and Lee's (2004) movie review corpus with the selected features being unigrams, unigrams along with subjectivity analysis, bigrams and adjectives. The frequencies of the features are used in the feature vector for SVMs and NB, while feature presence is used for MaxEnt. Their results show that there is little difference in accuracy of the three compared algorithms.

Several types of features or feature selection schemes have been also used in opinion mining research studies. In one of them, Mullen and Collier (2004) use SVMs to bring together several favorability measures for adjectives and phrases, the unigram model of Pang et al. (2002), lemmatized versions of the unigram models and, where available, knowledge of the topic of the text. Their hybrid SVMs reach an accuracy of 84.6 % on movie reviews data.

Ng et al. (2006) examine the role of four types of simple linguistic knowledge sources in the automatic polarity classification of movie reviews using a SVMs classifier. Their results show that bigrams and trigrams selected according to the weighted log-likelihood ratio as well as the manually tagged term polarity information are very useful features for the task.

Kennedy and Inkpen (2006) present a combined method for determining the sentiment of movie reviews. First, they use two different methods separately: a term-counting approach (66.5 % accuracy) and a machine learning approach using SVMs with unigrams as features (84.9 % accuracy). Then, by combining the two methods together, they achieve better results (85.4 % accuracy).

Finally, Rushdi et al. (2011) apply SVMs on three datasets with different sizes and domains; namely, they use the movie review corpus of Pang and Lee (2004), the multi-domain corpus of Taboada and Grieve (2004) and a digital camera review corpus (SINAI) created by them. They use three different weighting schemes (i.e. word frequency in document and in the entire corpus (TF-IDF), Term Occurrences (TO) and Binary Term Occurrences (BTO)) and three different n-gram techniques (i.e. unigrams, bigrams and trigrams) in order to examine how these features affect the sentiment classification task. Their results show that TO is the worst weighting method while TF-IDF and BTO give similar results. As far as n-gram techniques are concerned, trigrams are superior for the first two corpora while bigrams perform better in the SINAI corpus.

3 Methodology

The aim of our study is to build a prototype for the classification of hotel reviews based on the sentiment expressed in them. We preferred to apply a machine learning approach which has been shown that is more accurate than semantic orientation approaches (Chaovalit and Zhou 2005; Boiy et al. 2007; Kennedy and Inkpen 2006). Therefore, we started by collecting our data set, which was then used for the training of the SVMs classifier. Selecting unigrams (single words) as features, we followed two different methods. In the first method, the classification algorithm computes the frequency of individual words by applying the TF-IDF weighting scheme (TF-IDF bag-of-words model), while in the second method the algorithm counts the occurrence of selected individual words which express positive or negative sentiment.

3.1 Data Set

The corpus of hotel reviews was collected from the Greek version of *Tripadvisor* which is one of the world's largest travel sites (www.tripadvisor.com.gr). Our data set consists of 1,800 reviews (900 positive and 900 negative). Reviews translated in Greek were not taken into consideration as they contained grammatical and syntactic errors. Extremely short (i.e. less than 30 words) or very lengthy (i.e. more than 250 words) reviews were also excluded from the corpus. In order to ensure a proper training set, the data were manually checked by processing of the selected reviews, namely the correction of spelling and punctuation errors. The labeling of the reviews as either positive or negative derived from the combination of the reviewers' ratings and our personal intuition. Finally, in order to have a balanced typology of hotel reviews, we tried to include an equal distribution among different travel destinations as well as accommodations in Greece i.e. hotels, villas or apartments located close to mountain, sea or city centers.

3.2 Classification Algorithm and Features Selection

The majority of machine learning approaches treat sentiment classification problems by building SVM classifiers, which have been proved to produce better results than other machine learning techniques (Vapnik 1998; Pang et al. 2002; Pang and Lee 2004; O'Keefe and Koprinska 2009). Joachims (1998) mentions the significance of SVMs in text categorization tasks; he claims that "SVMs are robust and, with their ability to generalize well in high dimensional feature spaces, eliminate the need for feature selection" (p. 142).

In our research, data training is performed by a binary SVMs classification algorithm which labels sentiment polarity (positive or negative) on texts represented as feature vectors using feature selection on unigrams (Pang et al. 2002).

3.3 Experimental Setup

In order to run our experiments we made use of the RapidMiner software version 5 (www.rapidminer.com) with its text mining extension which provides different tools that are necessary for statistical text analysis. RapidMiner is an open source analytics platform which exploits statistics, machine learning, and natural language processing techniques to automate sentiment analysis on large collections of texts.

In both classification methods, we applied the SVM operator which is provided by RapidMiner and we also implemented the bag-of-words language model in which a text is represented as the bag of the words it contains, where each individual word (unigram) is considered as one feature. Furthermore, due to the fact that the raw data we have collected were not directly readable by the algorithm, which requires numerical feature vectors, some pre-processing of the data was needed.

In order to address this issue as well as to avoid unnecessarily large feature vectors, each text was automatically tokenized and filtered in relation to the length of its tokens. Tokens that consisted of less than four or more than 25 characters were removed. In addition, we used a list of Greek stop words in order to remove semantically empty tokens such as articles, pronouns, and prepositions. At the end of this procedure, our data set can be represented by a matrix with one row per document and one column per token that occurs in the corpus.

In respect to term weighting, RapidMiner uses four weighting methods for unigrams: Term Frequency (TF), TF-IDF, TO and BTO. We have decided to process our documents with the second and third weighting scheme.

3.4 The TF-IDF Bag-of-Words Approach

In the bag-of-words model that we have adopted, each document is represented as an unordered collection of features. In order to generate the feature vector, we used the TF-IDF weighting scheme. TF-IDF is the most common term weighting method in the field of Information Retrieval and previous research has demonstrated that it can significantly increase the classification accuracy of sentiment analysis systems (Paltoglou and Thelwall 2010).

The TF-IDF weighting scheme estimates the informativeness of a given term in a given document by combining two scores: its TF weight and its IDF weight. TF gives measure of the importance of the term within the particular document and is calculated by dividing the number of occurrences of a given term into the number of

total words in that document. IDF estimates the rarity of a term in the whole document collection and it is calculated by dividing the total number of documents by the number of the documents, in which a given term is occurred. The key idea behind IDF is that words that appear infrequently in a collection of documents tend to be more informative than the words that appear frequently across many documents. Each term in a document receives, hence, a specific weight by multiplying these two scores.

TF-IDF weight is higher when a term occurs either many times within a given document or a few times in a large number of documents. Conversely, a lower weight in TF-IDF is reached when a term occurs few times in a given document or many times in many documents. If a term appears in almost all the documents of the collection, then its IDF is close to one.

In our research, we have selected the first 1,000 features with the higher TF-IDF weight in the corpus regardless of their positive or negative label. Based on these features the algorithm is trained to predict the polarity of new unclassified documents as either positive or negative.

3.5 The Term Occurrence (TO) Approach

The term occurrence (TO) approach is the simplest approach that has been used in determining the sentiment of a document. In this approach each document is classified as either positive or negative according to the number of polarity terms that it contains. More specifically, if a document contains more positive than negative words, it is assumed to have positive semantic orientation. Alternatively, when in a document there are more negative than positive words, it is considered to express negative sentiment. Finally, if the number of polarity terms is equal, the document is considered to be neutral.

In order to apply this approach in our study for hotel reviews, we first had to manually build a sentiment lexicon with Greek words with positive or negative meaning. Subsequently, since Greek is an inflected language, we had to count for all the inflected types of each word; we utilized Wordforms Applet 0.2 (http://users.otenet.gr/~nikkas/grammar/wordforms.html), which is an open-source tool that inflects Greek words in a semi-automatic manner. Our resulting lexicon includes verbs, nouns, adjectives, adverbs, comparatives, superlatives, and participles and comprises a total of 27.388 types of positive words and 41.410 types of negative words.

In the next step, the two lists with the polarity words were imported in the algorithm and we selected the TO weighting method, which gives us the exact number of occurrences of a given polarity term in a document. Finally, the algorithm counts the total occurrences of positive and negative terms in a given document and classifies it into the respective category.

4 Results

In order to validate our data set, we applied tenfold cross validation, i.e. our data were randomly separated into ten equal size folds with each of them containing 180 hotel reviews. Ninefold function as training data and the remaining functions as the validation data for testing the algorithm. The cross validation process is then repeated ten times resulting in the evaluation of the whole corpus.

The cross-validated performance of the classifier was evaluated using the measures of accuracy, recall and precision. Accuracy indicates how well our classifier can predict the category that a review belongs to. It is calculated by the ratio of the number of correctly classified positive and negative hotel reviews to the total number of hotel reviews being used. Recall is estimated as the ratio of the number of hotel reviews correctly classified as positive to the total number of hotel reviews that belong to that category. Finally, precision is defined as the ratio of the number of positive hotel reviews that are classified correctly to the total number of the reviews that are predicted to be positive.

Tables 1 and 2 present the results of the TF-IDF bag-of-words and the TO approach respectively according to the average accuracy over a tenfold cross-validation. Table 1 shows that the TF-IDF bag-of-words method achieves a remarkably satisfactory accuracy (95.78 %) as the algorithm classified correctly 1,724 out of 1,800 hotel reviews. The recall and precision rates are quite high too. More specifically, the recall rate is 93.78 % (844 out of 900 positive hotel reviews) and the precision rate is 97.69 % (844 out of 864 predicted positive hotel reviews). Table 2 shows the results of the TO method. In this case, the accuracy is 71.76 %, namely 1,222 correctly classified reviews. The value of recall is 100 % as all positive reviews were correctly classified by the algorithm (899 out of 899 positive reviews) while precision is 65.14 % (899 correctly classified as positive out of 1,380 predicted to be positive hotel reviews).

Table 1 Results of the TF-IDF bag-of-words approach

	Predicted negative	Predicted positive
Negative cases	880	20
Positive cases	56	844
Accuracy	95.78 %	
Recall	93.78 %	
Precision	97.69 %	

Table 2 Results of the TO approach

	Predicted negative	Predicted positive
Negative cases	323	481
Positive cases	0	899
Accuracy	71.76 %	
Recall	100 %	
Precision	65.14 %	

5 Discussion

By examining the results in Tables 1 and 2, a comparison of the two classification methods reveals that the TF-IDF bag-of-words language model obviously performs much better than the TO approach. Even though the second method gives to some degree good results, they are not as satisfactory as those of the first one. From the results of the second table, we can observe that there is a great deviation between recall and the other two measures. Although this method obtains 100 % recall, the precision rate remains relatively low. More specifically, while all positive reviews, except one, are correctly classified as such, the greatest part of the negative reviews was incorrectly predicted to be positive. It is notable, however, that whenever the output of the classifier is negative, the prediction is always correct.

This significant difference between recall and precision, which results in a decrease of the overall accuracy of the TO approach, occurs possibly due to two reasons. Firstly, as the results show, the selection of unigrams as features affects mainly the performance of the second method. By not taking into account the context of the selected polarity words, the algorithm faces problems in the classification task. For instance, the shift of the semantic orientation of a clause, which may be caused by the occurrence of negatives and intensifiers such as *not* and *but*, is not identifiable by the algorithm. Secondly, a substantial part of the reviews remain unclassified as a result of the occurrence of equal number of positive and negative sentiment words. The fact that 97 reviews (96 negative and 1 positive) fail to be classified in the respective category definitely affects the results of the TO method.

To sum up, as far as the classification of hotel reviews is concerned, it becomes clear that the TF-IDF bag-of-words method is more robust than the TO method.

6 Limitations

In a classification task the performance of the machine learning algorithm depends on the features that have been selected. In the present study, the bag-of-words representation of the documents entails that we did not take into account any word order dependency. In particular, we did not use any computational grammar and thus the effect of contextual valence shifters like negatives, intensifiers and diminishers is not examined.

Another reason which may constrain the results of our study is the sparsity of the feature vector due to the great number of features that are irrelevant and could be considered as noise to the classification task. More precisely, in order to reduce noise, we could first determine whether the sentences of the reviews express an opinion or some factual information and then examine only the opinionated sentences. Furthermore, we did not distinguish between on-topic and off-topic passages in our data. In many hotel reviews the authors usually provide information

such as descriptions of either their trip or the travel destination which are redundant and irrelevant.

The application of our automatic sentiment classifier in other domains should be done carefully as our results are domain-specific.

Finally, in relation to the second method, one additional limitation arises from the manual generation of the polarity lexicon, as it is possible that we may have omitted some sentiment words. Therefore, it is useful to extent the lexicon by adding more sentiment words including certain domain-specific entries.

7 Conclusion

Within the field of sentiment analysis little research has been done in the hotel domain. In this study we tried to develop an automatic sentiment classifier for hotel reviews written in Greek. Two different classification methods were compared, namely the TF-IDF bag-of-words model and the TO approach. Experimental results have shown the effectiveness of the first method which can be compared with state-of-the-art existing approaches. The resulting polarity classifier could be easily deployed in many domains and produce good results without using sophisticated, hand-picked sentiment wordlists.

The developed prototype could be exploited in many different and practical ways. Firstly, by integrating the algorithm into a recommender system, we could facilitate the classification of the hotels as either recommended or not recommended. Hotels that receive a lot of positive reviews will be recommended as opposed to hotels that receive many negative reviews. The exploitation of such kind of information could benefit both individuals and hoteliers; individuals for collecting more focused information for their travel plans, and hoteliers for gathering important feedback so that they can improve the quality of their services. Secondly, an expansion in the use of our classifier in reviews from product or service domains may facilitate further the understanding of how each product or service is perceived by customers.

References

Boiy, E., Hens, P., Deschacht, K., & Moens, M. F. (2007). Automatic sentiment analysis of on-line text. In L. Chan & B. Martens (Eds.), *Openness in digital publishing: Awareness, discovery and access* (pp. 349–360). Vienna: IRIS-ISIS.

Chaovalit, P., & Zhou, L. (2005). *Movie review mining: A comparison between supervised and unsupervised classification approaches*. In Proceedings of the 38th Hawaii international conference on system sciences, 112.3. doi:10.1109/HICSS.2005.445

Hatzivassiloglou, V., & McKeown, K. (1997). *Predicting the semantic orientation of adjectives*. In Proceedings of the 35th annual meeting of the association for computational linguistics (pp. 174–181). doi: 10.3115/976909.979640

Joachims, T. (1998). Text categorization with support vector machines: learning with many relevant features. In C. Nédellec & C. Rouveirol (Eds.), *Machine learning: ECML-98: 10th European conference on machine learning* (pp. 137–142). Berlin: Springer.

Kasper, W., & Vela, M. (2011). Sentiment analysis for hotel reviews. In K. Jassem, P. Fuglewicz, M. Piasecki, & A. Przepiorkowski (Eds.), *Proceedings of the computational linguistics-applications conference* (pp 45–52). Polskie Towarzystwo Informatyczne.

Kennedy, A., & Inkpen, D. (2006). Sentiment classification of movie reviews using contextual valence shifters. *Computational Intelligence, 22*(2), 110–125.

Mullen, T., & Collier, N. (2004). *Sentiment analysis using support vector machines with diverse information sources*. In Proceedings of the 9th conference on empirical methods in natural language processing (pp. 412–418). Retrieved from http://acl.ldc.upenn.edu/acl2004/emnlp/

Ng, V., Dasgupta, S., & Arifin, S. M. N. (2006). *Examining the role of linguistic knowledge sources in the automatic identification and classification of reviews*. In Proceedings of COLING/ACL 2006 main conference poster sessions (pp. 611–618). Association for Computational Linguistics.

O'Keefe, T., & Koprinska, I. (2009). *Feature selection and weighting methods in sentiment analysis*. In Proceedings of the 14th Australasian document computing symposium. Retrieved from http://es.csiro.au/adcs2009/proceedings/

Paltoglou, G., & Thelwall, M (2010). *A study of informational retrieval weighting schemes for sentiment analysis*. In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (pp. 1386–1395). Association for Computational Linguistics.

Pang, B., & Lee, L. (2004). *A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts*. In Proceedings of the 42nd association of computational linguistics (pp. 271–278). doi: 10.3115/1218955.1218990

Pang, B., Lee, L., & Vaithyanathan, S. (2002). *Thumbs up? Sentiment classification using machine learning techniques*. In Proceedings of the conference on empirical methods in natural language processing (pp. 79–86). doi: 10.3115/1118693.1118704

Rushdi, S. M., Martín-Valdivia, M. T., Montejo-Ráez, A., & Ureña-López, L. A. (2011). Experiments with SVM to classify opinions in different domains. *Expert Systems with Applications, 38*(12), 14799–14804.

Taboada, M., & Grieve, J. (2004). *Analyzing appraisal automatically*. In Proceedings of the association for the advancement of artificial intelligence spring symposium on exploring attitude and affect in text: Theories and applications (pp. 158–161). Retrieved from http://www.sfu.ca/~mtaboada/research/pubs.html

Turney, P. D. (2002). *Thumbs up or thumps down? Semantic orientation applied to unsupervised classification of reviews*. In Proceedings of the 40th annual meeting of the association for computational linguistics (pp. 417–424). doi: 10.3115/1073083.1073153.

Vapnik, V. (1998). *Statistical learning theory*. New York, NY: Wiley.