Design of a Sentiment Lexicon for the Greek Food and Beverage Sector

Anastasios Liapakis, Theodore Tsiligiridis, and Constantine Yialouris

Abstract Sentiment Analysis is a computational method aiming to extract opinions/evaluations of individuals for an entity such as a product, a service etc. In social media networks and other online sources, as for example the food websites, sentiment analysis is able to identify all possible terms, such as simple words, combinations of words, or phrases (pre-processing stage) that can be used to express the feelings of a user for a specific entity. Then, by considering the characteristics of these media, such as time sensitivity, text size limitation and unstructured expressions, converts them giving a positive or negative significance. For the analysis, a set of linguistic, statistical and machine learning techniques are usually considered to structure the information contained in text sources. The main purpose of this paper is twofold. First is to provide a literature review of the sentiment analysis techniques and second is to design a sentiment lexicon as a preliminary step to further analyze the customers' reviews of some leading companies in the Greek Food and Beverage industry as they uploaded in the most common Opinion Social Networks in Greece (fb). Existing research has focused mainly on the recognition on English characters, while to our knowledge, limited research papers have been published so far concerning the Greek language, concentrating mainly on the banking and financial sector, neglecting contributions on food industry. Note that since significant portion of online text-based Greek communications ignore the rules of spelling and grammar the study takes into account this trend and improves the calculation of a sentiment score accordingly. As appears, the findings are expected to contribute in the design issues of a sentiment lexicon particularly devoted to the Greek food and beverage industry and to be used for further analysis.

Keywords Sentiment analysis · Modern Greek · Food and beverage sector

A. Liapakis (✉) · T. Tsiligiridis · C. Yialouris
Informatics Laboratory, Department of Agricultural Economics and Rural Development, School of Applied Economics and Social Sciences, Agricultural University of Athens, Athens, Greece
e-mail: liapakisanastasios@aua.gr; tsili@aua.gr; yialouris@aua.gr

E. Krassadaki et al. (eds.), *Operational Research in Agriculture and Tourism*, Cooperative Management, https://doi.org/10.1007/978-3-030-38766-2_3

1 Introduction

Social Media networks are usually changed consumers' behavior. Nowadays, more and more companies use social media marketing in order to attract more customers. This modifies consumers' attitudes and companies cannot detect these modifications due to the big volume and the diversity of the produced information. According to surveys (comScore/the Kelsey group, 2007; Horrigan, 2008), 81% of the Internet users in USA have done online research on a product at least once, the vast majority of questioned are willing to pay from 20 to 99% more for a five-star-rated item or a service than a low-star-rated item and interestingly, Rainie and Hiltin (2004), report that "Individuals who have rated something online are also more skeptical of the information that is available on the Web". Other surveys have shown that the majority of Internet users, usually do researches on opinion networks for products that they are willing to buy and claim that reviews influence their purchase decision. Obviously, the variety of social media networks creates multimedia data as text, audio, images and videos files which are difficult to be edited or sorted by the companies or other users.

Facebook is the most popular social media platform all over the world and in Greece too. Companies are using it for reading what other users say about their products or services, responding to users' messages and distributing content about their product or services. Table 1, shows according the Eurostat, the most common social media (Hu & Liu, 2012) tools that are used from the Greek companies and individuals too.

In the case of Greek Food and Beverage (F&B) industry, the use of social media networks is very high to multinational and large companies. This is due to the fact that this sector, is especially prone to problems in sustainability given its high impact and dependence on natural, human and physical resources (Liapakis, Costopoulou, Tsiligiridis, & Sideridis, 2017). The challenges that face are numerous, including environmental sustainability (usage of natural resources, animal welfare etc.), social sustainability (labor and work conditions, food safety etc.) and economic sustainability (energy usage, waste management etc.) (Genier, Stamp, & Pfitzer, 2009). Every day, various campaigns regarding the previous aspects are shown in social media networks by the companies in order to eliminate the attacks concerning the good reputation of their brand names. In this framework the quality of services, marketing and maximization of sales will be enforced by considering the textual content that is generated by Internet users. Sentiment analysis (subjective analysis, opinion mining, and appraisal extraction with some connections to affective

Table 1 Categorization of the most common social media tools in Greece

Category	Platforms
Blogging	Blogger
Micro-blogging	Twitter
Opinion and reviews	Trip advisor, efood
Media sharing	YouTube, Instagram
Social networking	Facebook, LinkedIn

computing are also alternative names) can approach methodologically the problem by identifying relevant information from the huge communication of stakeholders over the Internet. The data created has resulted in the development of web opinion mining, as a concept in web intelligence focusing on extracting, analyzing and combining web data about user thoughts. The sentiment analysis is significant since the users' opinions provide information of how people feel about a topic of interest and understand how this was acknowledged by the market.

Sentiment analysis can be seen as a classification process divided into three main levels; document-level, sentence-level, and aspect-level. Document-level aims to classify a document by expressing a positive or negative opinion (sentiment). It assumes that the whole document is a basic information unit and it refers about one topic only. Sentence-level classifies opinion expressed in each sentence. Only subjective sentences need to be considered and, in such cases, sentence-level will determine whether the sentence expresses positive or negative opinions. Since sentences are short documents there is no substantial difference between document and sentence level classifications. However, to provide detail opinions on all aspects of the entity which is needed in many applications, we have to classify the sentiment with respect to the specific aspects of entities, namely to move to the aspect level sentiment analysis. This level of analysis requires in the first place to identify the entities and their aspects. The sentiment holders may provide various opinions for different aspects of the same entity like, for example: "the food quality of this restaurant is very good, but the service time is long".

Sentiment analysis is, in fact, a sentiment classification problem and therefore it is necessary to extract and select text features. According to Aggarwal and Zhai (2012), some of these features are the terms presence and frequency which are individual words or word n-grams and their frequency counts. It either gives the words binary weighting (zero if the word appears, or one if otherwise) or uses term frequency weights to indicate the relative importance of features (Yelena & Padmini, 2011); the Parts of Speech (POS), tags the words of a sentence as verb, adverb, adjective and noun based on some POS tagging rules. The following example could help us to understand the method precisely. Assuming the sentence: "The place is clean and the food is very good." Then in POS method, the review should be written as: The/***DT***, place/***NN***, is/***VB***, clean/***JJ***, and/***CC***, the/***DT***, food/***NN***, is/***VB***, very/***RB***, good/***JJ*** where, **DT**: Determiner, **NN**: Noun, **VB**: Verb, **JJ**: Adjective, **CC**: Conjunction, **VB**: Verb in present tense, and **RB**: Adverb.

Finally, negations, which are negative words that may change the opinion orientation, like 'not good' an alternative to 'bad'. Feature selection approaches can be lexicon-based that usually start with a small set of seed words, and extend this set through synonym detection, online resources and or by using a large corpus of documents from a single domain in order to obtain a larger lexicon or statistical-based automatic methods that are used more frequently. The feature selection techniques treat the documents either a group of words (Bag of Words (BoWs) or as a string which retains the sequence of words in the document. BoW is used more often because of its simplicity for the classification process. The most common feature selection step is the removal of stop-words and stemming (returning the word

to its stem or root i.e. 'flies' to 'fly'). Another popular term-weighting numerical statistic used in this context is the Term Frequency–Inverse Document Frequency (TF-IDF), which reflects the importance of a word in a document of a collection or corpus. The value of the TF-IDF statistic increases proportionally to the number of times a word appears in the document and is balanced by the number of documents in the corpus that contain the word. This helps to adjust for the fact that some words appear more frequently in general. Frequently used statistical methods in feature selection are the Point-wise Mutual Information (PMI) (Cover & Thomas, 1991), the Chi-square $\chi 2$ (Fan & Chang, 2012) and the Latin Semantic Indexing (LSI) (Deerwester, Dumais, Landauer, Furnas, & Harshman, 1990). Other methods are the Gini index (Yelena & Padmini, 2011), the Hidden Markov Model (HMM) and the Latent Dirichlet Allocation that simultaneously model topics and syntactic structures in a collection of documents (Duric & Song, 2012; Griffiths, Mark, Blei, & Tenenbaum, 2005) as well as an approach to detect irony in customer reviews (Reyes & Rosso, 2012). Some discussion on the methods above can be seen in Medhat, Hassan, and Korashy (2008) and Feldman (2013).

Sentiment classification analysis relies on three types of techniques, i.e., machine learning-based and lexicon-based techniques as shows in Fig. 1.

Machine learning based techniques apply some well-known machine learning algorithms and use linguistic features (Singh, Singh, & Singh, 2016). To describe the corresponding text classification problem, we assume that each record of the training set of records is labeled to a class. Then, for a given instance of an unknown class, a classification model is used to predict a class label for it. Obviously, the model is related to the features in the underlying record to one of the class labels. A label can be assigned to an instance either directly or indirectly, by selecting it from a set of labels in accordance to some predefined probability. The text classification machine

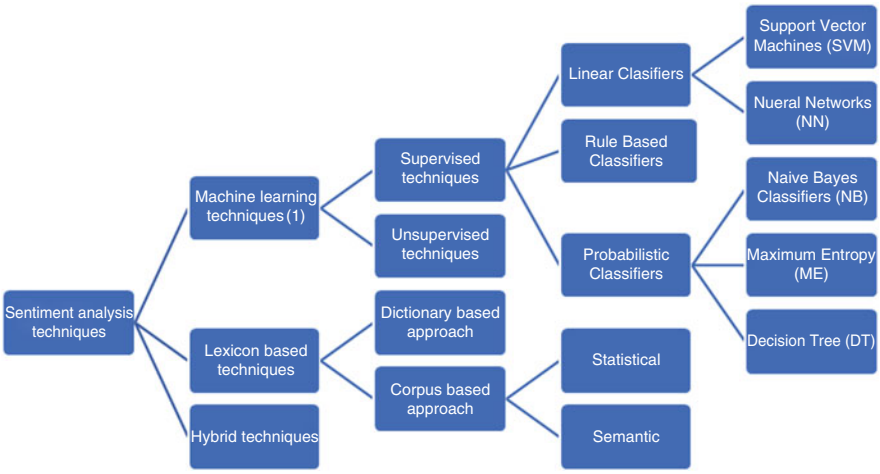

Fig. 1 Types of sentiment analysis classification techniques

learning methods, can be distinguished by the approach used; namely supervised, semi-supervised and unsupervised.

The supervised methods make use of a large number of labeled training documents and include four categories of classifiers; the decision tree, the linear, the rule-based and the probabilistic. Insight on these categories subdivides them further; however, the most interesting cases include the linear classifiers which are split into Support Vector Machines (SVM) and Neural Networks (NN) as well as the probabilistic classifiers into Naive Bayes (NB), Bayesian Network (BN), Decision Tree (DT), and Maximum Entropy (ME).

As for the unsupervised text classification methods, which are used when it is difficult to find the labeled training documents and consequently it is difficult to classify documents into categories, alternative methods are used (Ko & Seo, 2000; Read & Carroll, 2009; Turney, 2002; Xianghua, Guo, Yanyan, & Zhiqiang, 2013). Details for the semi-supervised classification can be seen in He and Zhou (2011).

The lexicon-based approach (2) relies on sentiment or opinion lexicon, a collection of known and precompiled sentiment terms. It is divided into manual, dictionary-based and corpus-based approaches. The first one is quite consuming and it is used only in particular cases. The dictionary-based approach depends on finding opinion seed words and then searches the dictionary of their synonyms and antonyms. The corpus-based approach begins with a seed list of opinion words and then finds other opinion words in a large corpus to help in finding opinion words with context-specific orientations. This could be done by using statistical or semantic methods.

The remaining two use statistical or semantic methods to find sentiment polarity. Corpus-based techniques are based on decision trees such as *k*-Nearest Neighbors (*k*-NN), Conditional Random Field (CRF), Hidden Markov Model (HMM), Single Dimensional Classification (SDC), Sequential Minimal Optimization (SMO), and related to methodologies of sentiment classification. The hybrid approach combines both approaches and is very common with sentiment lexicons playing a key role in the majority of methods. Other approaches to sentiment analysis may also be found in the literature. Some of the most characteristic studies include (Gräbner, Zanker, Fliedl, & Fuchs, 2012) in which sentiment analysis does what every user is required to do after writing a product review: to quantify the opinion represented by the text with percentages or stars; Yu, Zhou, Zhang, and Cao (2017) in which it aims to automatically classify the text of written reviews from customers into positive or negative opinions; and Vinodhini and Chandrasekaran (2012) in which the type of natural language processing for tracking the mood of the public about a particular product or topic.

To conclude, sentiment analysis is contextual mining of text which identifies and extracts information from online reviews, websites, and social media networks. It is applied in many fields such as consumer information, marketing, books and websites (Griffiths et al. 2005). It helps businesses to compute the social sentiment of their brand, products or services while was examining online conversations. To facilitate processing and storage in databases, an opinion/review of a user about an entity can be analyzed as a five-dimensional vector of the form:

Opinion (*Entity Name, Entity Part, Polarity, User, Date*).

In the case of the Food and Beverage industry and for the purposes of this study we suggest the following form:

Opinion (*Name*, *Function*, *Polarity*, *User*, *Date*)

where, *Name* is the restaurant's or cafeteria's name; *Function* could be the image of the store (decoration etc.), the service facilities (location, working hours etc.), the staff (availability, information offered etc.) and the products (quality, variety etc.) taking values *positive* or *negative*; *Polarity* could be *positive* or *negative*; *User* is the user's name; and *Date* is the date that the review was written. For the better understanding, a review for a known Athens restaurant (named: *X*) on Tripadvisor could be used: "The place is clean and the food is very good. We like the atmosphere and the way its look. Very good coffee and breakfast"—Reviewed July 18, 2018 by *Name A*. According to the suggested form and for further analysis, the review could be shown as follows:

Opinion1 = {*X*, image of the store: *positive*, *A*, *July 18, 2018*}
Opinion 1.1 = {*X*, products: *positive*, *A*, *July 18, 2018*}

The structure of the article is as follows: The second section presents a literature review of the most important supervised techniques that are used in sentiment analysis as well as an overview of the related to our study articles. Section 3 presents the methodology used for identifying the most common Greek adjectives in social media networks (written in modern Greek and Greeklish too). The final section ends with a discussion on the findings of this study.

2 Literature Review

2.1 Sentiment Analysis Techniques

As it has been pointed out in sentiment analysis, lexicon-based and machine learning are two techniques that are used in common; in lexicon-based schemes, a dictionary with selected sentiment words is proposed, whereas, in machine learning techniques various methods are used for sentiment classification (supervised and unsupervised algorithms) (Haseena & Tanvir, 2014).

BoW and TF-IDF are two of the most popular feature selection schemes used so far and will be described in the sequel. In addition, for the machine learning approach the interesting is focused on the super-vised classification techniques, consisting of three consecutive phases, namely, the determination of the data set, the selection of the appropriate classifier and finally the feature selection. The selection classifiers to be considered in this section are the *Supporting Vector Machines* (SVM) (Joachims, 2001), the *Naïve Bayes* (NB) (Lewis, 1998), the *Maximum Entropy* (ME) (Pang, Lee, & Vaithyanathan, 2002) and the *K-nearest neighbors' algorithm (*Adeniyi, Wei, & Yongquan, 2016*)*.

The *Bag of Words* (BoW) method (Gabryel, Damaševičius, & Przybyszewski, 2018) is the most known algorithm in SVM. It aims to deal document with linear algebraic operator by transforming a text into sparse numeric vectors. Terms are stored in dictionaries and can be represented by simple words, (1-gram) or composed words (2, 3, . . ., n-gram) that occur in various documents. In n-grams dictionaries the priority of the words composing the term (phrase) is very important as it changes the semantic, e.g. for the 4-gram term "The food was beautiful" or "Was the food beautiful?" Each term is used as an attribute of the data set represented in the attribute-value form. Thus, in BoW model, a term is represented as a separate variable having numeric weight of varying importance. In particular, after preprocessing the set of documents d_i; $i = 1, 2, \ldots, n$, can be represented in a matrix of the form $[u_{ij}]$, with u_{ij}; $i = 1, 2, \ldots, n$; $j = 1, 2, \ldots, m$ be a count of the number of occurrences of term t_j in document d_i, where n documents are represented and each document consists of m terms. Each document d_i can be seen as a histogram $d_i = [u_{i1}, u_{i2}, \ldots, u_{im}]$, with u_{ij} referring to the value of the j-th term of the i-th document. This results in a numerical vector which can be utilized as inputs in the various machine learning algorithms in order to classify documents into topics, to be used for many other machine learning applications where text is the initial input. This simple model is proven very flexible and effective in various scenarios.

The *Term Frequency—Inverse Document Frequency* (TFIDF or *tf-idf*) (Gabryel et al., 2018) is another most popular term-weighting scheme used today in information retrieval, text mining, and user modelling. This weight is a statistic showing how important a term is to a document in a collection or corpus of examined online reviews. It is the product of two factors: the first computes the normalized *Term Frequency* (TF), e.g., the number of times a term appears in a document, divided by the total number of terms in that document; the second term is the *Inverse Document Frequency* (IDF), computed as the logarithm of the number of the documents in the collection or corpus divided by the number of documents where the specific term appears. The simplest choice is of course to use the raw count of a term in a document, i.e., the number of times that term t occurs in document d. Thus,

$$tf - idf = tf \times idf = f^1_{t,d} \times f^2_{t,D} = f^1_{t,d} \times \log\left(\frac{N}{n_t}\right).$$

where: $f^1_{t,d}$, and $f^2_{t,d}$ are the *tf* and *idf* respectively, with $t = 1, 2, \ldots, T$; $d = 1, 2, \ldots, N$; $N = |D|$; and $n_t = |\{d \in D : t \in d\}|$. In the above T is the total number of terms appeared in the corpus N is the total number of documents the corpus contains and n_t is the number of documents containing the term t. Note that in the usual case, when the term is a simple word (1-gram) then the above formula becomes:

$$tf - idf = tf \times idf = f^1_{w,d} \times f^2_{w,D} = f^1_{w,d} \times \log\left(\frac{N}{n_w}\right)$$

Now *tf-idf* is the relative weight of the feature in the vector, *tf* presents the number of words occurrences in total of reviews, n_w is the number of documents in the

corpus containing the word, N is the number of reviews in the corpus. The *tf–idf* value increases proportionally to the number of times a word appears in the document but is offset by the number of documents in the corpus that contain the word. This helps to adjust for the fact that some words appear more frequently in general.

As a working example consider a document containing 300 words wherein the words 'lovely' and 'badly' appear 15 and 45 times, respectively. The term (word) frequencies, i.e., $tf = f^1_{w,d}$ for $w =$ 'lovely' and $w =$ 'badly' is then $15/300 = 0.05$ and $45/300 = 0.15$, respectively. Now, assume we have two million documents and the word 'lovely' and 'badly' appear in 50 and 450 thousands of these, respectively. Then, the inverse document frequency, i.e., $idf = f^2_{w,D}$ is calculated as $\log(2{,}000{,}000/50{,}000) = 1.6$ and $\log(2{,}000{,}000/400{,}000) = 0.699$ and the *tf-idf* weight is the product of these quantities: $0.05 \times 1.6 = 0.08$ and $0.15 \times 0.699 = 0.105$ respectively.

Variations of this scheme are often used by search engines as a central tool in scoring and ranking a document's relevance given a user query. A simplest ranking function is computed by summing the *tf-idf* for each query term, however, many more sophisticated rank functions are variants of this simple model. The advantages are that the *tf-idf* metric, as well as, the similarity between two documents are computed easily, and in addition, it extracts in a flexible way the most descriptive terms in a document. In contrast, *tf-idf* is based on BoW model, and therefore it is only useful as a lexical level feature and cannot capture position in text, semantics, co-occurrences in different documents, etc.

The *Support Vector Machine* (SVM) classifier is a supervised machine learning algorithm or kernel methods that can be used for binary classification or regression. It constructs an optimal hyperplane as a decision surface such that the margin of separation between the two classes in the data is maximized. To introduce SVM (Joachims, 2001) we consider a training dataset $S = \{(\boldsymbol{x}, \boldsymbol{y})\} = \{(x_i, y_i)\}_{i=1}^m$. where $x_i \in R^n$ is the input vector, and $y_i \in \{+1, -1\}$ is the target value. SVMs map these input vectors into a high dimensional reproducing kernel space, where a linear machine is constructed by minimizing a regularized functional. The linear machine is of the form $f(\boldsymbol{x}) = <\boldsymbol{w} \cdot \varphi(\boldsymbol{x})> + \mathrm{b}$, where $\varphi(\cdot)$ the mapping function, b is the bias, and the dot product $<\varphi(\boldsymbol{x}) \cdot \varphi(\boldsymbol{x}^T)>$ is also reproducing kernel $\mathrm{K}(\boldsymbol{x}, \boldsymbol{x}^T)$. Then, the standard SVM can be stated as the following objective function to be minimized:

$$\min_{w \in R^n} \frac{\lambda}{2} \|\boldsymbol{w}\|^2 + \frac{1}{m} \sum_{(x,y) \in S} \max \{0.1 - \boldsymbol{y} < \boldsymbol{w}, \boldsymbol{x} >\}$$

According to Shalev-Shwartz, Singer, Srebro, and Cotter (2011), Pegasos algorithm can be used to solve the SVM since it was proved to be faster, having higher computational efficiency than the standard SVM. Then the modified objective SVM function above becomes:

$$\min_{w \in R^n} \frac{\lambda}{2} \|\boldsymbol{w}\|^2 + \frac{1}{m} \sum_{i=1}^{m} \max \{0.1 - y_i \boldsymbol{w}^T x_i\}$$

In the above, λ is the regularization term, $\boldsymbol{w}$ is the estimated vector, indicating a score for each word, m is the number of reviews, y_i is 1 if the review is positive or -1 if the review is negative, and x_i is the feature vector of each review. Pegasos is a stochastic sub-gradient descent method with varied step size with the following pseudocode:

Input $\lambda > 0$
Choose $w_1 = 0, t = 0$
While epoch $<$ max_epochs
  For $j = 1, \ldots, m$
    t = t + 1
    $n_t = 1/(t * \lambda)$
    If $y_j * w_t^T < 1$ *then*
      $w_{t+1} = (1 - \eta_t * \lambda) * w_t + \eta_t * y_j * x_j$
    Else
      $w_{t+1} = (1 - \eta_t * \lambda) * w_t$
    End if
  Next j

In the proposed model, each word in a review treats as an individual feature and a sparse feature matrix with very high dimensions would be generated. To avoid numerous zeroes in the list of words that is generated, a dictionary would be set up for each review, in order to provide information only for the words appear in the review. The sentiment score of each word in a sample of reviews, is multiplied by its average frequency among all reviews, namely:

$$\textit{Sentiment score of a word } (t) = \textit{score}(t) \times \frac{\textit{total frequency } (t)}{\textit{total number of reviews}}$$

In the above, *score*(*t*) is the word score calculated from the model, and *total frequency*(*t*) is the total frequency of word t in all the reviews.

The Naïve Bayes (NB) classifier is based on the so-called Bayesian theorem and is particularly suited when the dimensionality of the inputs is high. Despite its simplicity, Naive Bayes can often outperform more sophisticated classification methods. It computes the posterior probability of a class, based on the probability of a class, say *Pclass*, and the conditional probability of a class, say *P*(*w*/*class*) in a review, is computed as *Pclass* = *Nclass*/*N* where, *Nclass* is the probability of a class (positive or negative) *N* is the total count of class in the training set (Sharma & Dey, 2012):

$$P(w|class) = \frac{count(w, class) + 1}{count(class) + |V|}$$

In the above w is the word attribute, c is the class, $count(w, class)$ is the total count of word attribute occurs in class, + 1 is the Laplace smoothing, $count(class)$ is the total count of word in a particular class occurs in the data set, and $|V|$ is the vocabulary T.

The Maximum Entropy (ME) is a probabilistic classifier which belongs to the class of exponential models that is used to classify text documents. Unlike the NB classifier the ME does not assume that the features are conditionally independent of each other. From all the models that fit the training data, the ME selects the one which has the largest entropy. It can be used to solve a large variety of text classification problems such as language detection, topic classification, sentiment analysis and more. According to Pang et al. (2002) it outperforms NB in standard text classification tasks. It estimates the conditional probability of a class by the following form:

$$P(class|w) = \frac{1}{Z(w)} \exp\left(\sum_i \lambda_{i,c} F_{i,c}(w, c)\right)$$

In the above $Z(w)$ is a normalization function, $\lambda_{i,\ c}{}'s$ are the feature $-$ weight parameters, and $F_{i,\ c}$ is defined as follows:

$$F_{i,c}(w, c) = \begin{cases} 1, & n_i(d) > 0 \\ 0, & otherwise \end{cases}$$

The parameter values (Reyes & Rosso, 2012) are set in a way that the maximum entropy classifiers maximize the entropy of the induced distribution while maintaining the constraints enforced by the training data.

The K-nearest neighbors' algorithm (KNN) (Shalev-Shwartz et al., 2011), is a similarity-based learning algorithm that has been shown to be very effective for a variety of problem domains including text categorization. It compares the unknown data with the training data comparing the distance between them. The distance could be computed by the following form:

$$d(x_i - x_t) = \sqrt{(x_{i1} - x_{t1})^2 + (x_{i2} - x_{t2})^2 + \cdots + (x_{ip} - x_{tp})^2}$$

In the above, x_i is an input of reviews with p features $x_i = (x_{i1},\ \ x_{i2}, \cdots, x_{ip})$, x_t is the other input of reviews with p features $x_t = (x_{t1}, x_{t2}, \cdots, x_{tp})$, p the total number of inputs $j = 1, 2, 3, \ldots, p$ and n is the total number of inputs $i = 1, 2, 3, \ldots, n$. Given a test document, the KNN algorithm finds the K-nearest neighbors among the training documents, and uses the categories of the K neighbors to weight the category candidates. The similarity score of each neighbor document to the test document is used as the weight of the categories of the neighbor document.

According to Feldman (2013), the most important resource for the sentiment analysis techniques is the design and development of a sentiment lexicon. There are three approaches that are used for the acquisition of the sentiment lexicon. The first is

the manual approach in which the coding of the proposed lexicon is done by hand. The second is the dictionary-based approach in which a set of words is used by utilizing external resources as for example the WordNet dictionary, and the third is the corpus-based approach in which a large set of documents is examined in order to propose a dictionary.

In the following section, we propose a lexicon in Greek and Greeklish (Greek words written with English letters) language after the analysis of reviews in the data set. This lexicon should be used in the lexicon-based techniques and especially, in the sentence-level sentiment analysis techniques.

2.2 Related Articles

For the purpose of this work, we examined a number of related articles presented in Table 2. Since our interest is for the food and beverage of the Greek sector, we restrict the review in English and Greek language. The first column indicates the methodology adopted by each article and it is divided into two categories (Liu, 2010), namely, the unsupervised learning methods (methods that perform classifications based on some fixed syntactic phrases which are likely to be used to express opinions), and the supervised learning methods (methods that should be readily applied to sentiment classification), such as Naive Bayesian (NB) and Support Vector Machines (SVM). The second column specifies the language that analyzed in each article. The third column determines the data source of each case. The sources could be online reviews, forecast etc. The fourth column specifies the part of speech detected, namely, nouns, pronouns, adjectives, verbs, participles and adverbs. The fifth column shows the industry the corresponding article is referred and the final column specifies the examined study. We should note that all the following articles use the binary polarity (positive or negative).

The main result that comes out from the above summary is that existing research has focused mainly on the recognition of English characters and that very limited research work has been published on the Greek language and Greeklish idioms so far. Most of the work focuses on the technology and movie sector, while there was a very limited work on food industry.

3 Design of Greek Sentiment Lexicon

3.1 Data Set

We created a corpus of Greek and Greeklish (Latin letters) reviews by retrieving individuals reviews from Facebook and Twitter social media platforms. Specifically, we examine 275 customer reviews of 10 Greek leading companies in the Greek food industry. For a proper training set, the reviews were manually checked. The small

Table 2 Related articles

Methods used	Language	Sources	Part of speech	Industry	References
Unsupervised	English	Reviews	Adj and Adv	N/A	Liu (2010)
Super/ unsupervised	English	Forecasts	All	N/A	Kang, Yoo, and Han (2012)
Supervised	English	Restaurant reviews	All	Fast food	Li and Wu (2010)
Supervised	English	Product reputation	All	Mobile phones	Morinaga, Yamanishi, Tateishi, and Fukushima (2002)
Unsupervised	English	Reviews	Adj and Adv	Movies	Turney (2002)
Supervised	Greek	Reviews	All	–	Kalamatianos, Mallis, Symeonidis, and Arampatzis (2015)
Super/ unsupervised	Greek	Reviews	All	–	Adam (2018)
Supervised	Greek	Reviews	All	Movies and Tech.	Giatsoglou et al. (2017)
Supervised	Cantonese	Restaurant reviews	All	F&B	Zhang, Ye, Zhang, and Li (2011)
Supervised	English	Hotel reviews	All	Hotels	Duan, Cao, Yu, and Levy (2013)
Supervised	English	Hotel reviews	All	Hotels	Gräbner et al. (2012)
Supervised	English	Restaurant reviews	All	F&B	Vinodhini and Chandrasekaran (2012)
Supervised	English	Reviews	All	F&B	Yu et al. (2017)
Lexicon based	Greek	Reviews	All	Technology	Agathangelou, Katakis, Kokkoras, and Ntonas (2014)
–	All	Reviews	All	All	Kokkoras, Ntonas, and Bassiliades (2013)

volume of the examined reviews comes as a result of the General Data Protection Regulation (GDPR). This regulation may cause problems in the mining of published online reviews from companies' social media accounts.

3.2 Results

Greek is a typical high inflection language due to its complex grammatical and syntax rules. It is a particularly challenging language for Neuro-linguistic programming and specifically for sentiment analysis. For instance, according to Giatsoglou

et al. (2017), in English, there are only 4 forms of the regular verb ask (ask, asks, asked and asking), while there are 93 different forms corresponding to the regular Greek verb "ρωτώ". According to the data set, we found the most used positive adjectives and we present them in Tables 3 and 4.

The most used word is the adjective Exypiretiko which means helpful. However, we should note that this adjective has also negative meaning if for example precedes the word not (δεν) as presented in the two following examined reviews:

The restaurant's staff was very helpful (To proswpiko tou estiatoriou itan poly exypiretiko)
The restaurant's staff was not so helpful (To proswpiko tou estiatorioy den itan poly exypiretiko)

In the case of Greek language, there are many adjectives whose semantic orientations depend on contexts in which they appear (fifth column). This problem can be resolved with the negation rules (e.g.: Not helpful/Oxi exypiretiko). The negation words or phrases usually reverses the opinion expresses in a sentence such as no, not and never. The basic rule that is applied for negation is: Negation Positive transforms to negative.

We should also note that each of the adjective that is shown in Table 3, may describe different *Function* (Facilities, Staff, Products and Image of the Store) of a restaurant (Table 5).

In order to solve this restriction, the proposed lexicon should include word sequences. The basic idea of this method is to compute sentiment scores in

Table 3 Proposed Greek adjectives

Adjective (Greek)	Pronunciation (in English)	Adjective (Greeklish)	Possible similar writing
Καλός/η/ο	kalos	Kalo (Good)	–
Ωραίος/α/ο	oreos	Oraios (Nice)	Oreo
Νόστιμος/η/ο	nostimos	Nostimo (Tasty)	Nosthmo, Nostymo, Nosteimo,Nostoimo
Τέλειος/α/ο	telios	Teleio (Perfect)	Telio
Φιλικός/η/ο	filikos	Filiko (Friendly)	–
Αγαπημένος/η/ο	agapimenos	Agapimeno (Favourite)	–
Εξαιρετικός/η/ο	ekseretikos	Exairetiko (Brilliant)	Exeretiko
Υπέροχος/η/ο	iperohos	Iperoho (Fabulous)	Iperoxo, Yperoxo, Yperoho
Εξυπηρετικός/η/ο	ekseepiretikos	Exypiretiko (Helpful)	Exipiretiko
Οικονομικός/η/ο	ikonomikos	Oikonomiko (Cheap)	Ikonomiko
Όμορφος/η/ο	Omorfos	Omorfo (Nice)	–
Θαυμάσιος/α/ο	θaumasios	Thaumasio (Terrific)	Thaymasio
Καταπληκτικός/η/ο	katapliktikos	Kataplitiko (Exceptional)	–
Γρήγορος/η/ο	γriγoros	Grigoro (Fast)	Grhgoro
Ευγενικός/η/ο	evjenikos	Evgeniko (Polite)	Ebgeniko

Table 4 Statistics of proposed Greek adjectives

Adjective	Frequency positive meaning	Relative frequency positive meaning	Frequency negative meaning	Relative frequency negative meaning	tf-idf positive meaning	tf-idf negative meaning
Καλός/η/ο	37.0	0.13	3.0	0.14	0.117	0.021
Ωραίος/α/ο	40.0	0.15	–	–	0.122	–
Νόστιμος/η/ο	18.0	0.07	–	–	0.078	–
Τέλειος/α/ο	4.0	0.01	–	–	0.027	–
Φιλικός/η/ο	5.0	0.02	4.0	0.18	0.032	0.027
Αγαπημένος/η/ο	6.0	0.02	–	–	0.036	–
Εξαιρετικός/η/ο	26.0	0.09	–	–	0.097	–
Υπέροχος/η/ο	6.0	0.02	–	–	0.036	–
Εξυπηρετικός/η/ο	50.0	0.18	6.0	0.27	0.135	0.036
Οικονομικός/η/ο	10.0	0.04	3.0	0.14	0.052	0.021
Όμορφος/η/ο	40.0	0.15	2.0	0.09	0.122	0.016
Θαυμάσιος/α/ο	5.0	0.02	–	–	0.032	–
Καταπληκτικός/η/ο	5.0	0.02	–	–	0.032	–
Γρήγορος/η/ο	8.0	0.03	2.0	0.09	0.045	0.016
Ευγενικός/η/ο	15.0	0.05	2.0	0.09	0.069	0.016

Table 5 Alternative functions that each adjective describes

Adjective	Function that referred most	Alternative function 1	Alternative function 2
Kalo (Good)	Products	Image of the store	Staff
Oraio (Nice)	Products	Image of the store	–
Nostimo (Tasty)	Products	–	–
Teleio (Perfect)	Products	Image of the store	Staff
Filiko (Friendly)	Staff	–	–
Agapimeno (Favourite)	Products	Staff	Image of the store
Exairetiko (Brilliant)	Products	Staff	Image of the store
Iperoho (Fabulous)	Products	Staff	Image of the store
Exypiretiko (Helpful)	Staff	–	–
Oikonomiko (Cheap)	Products	–	–
Omorfo (Nice)	Products	Image of the store	–
Thaumasio (Terrific)	Products	Image of the store	Staff
Kataplitiko (Exceptional)	Products	Image of the store	Staff
Grigoro (Fast)	Staff	–	–
Evgeniko (Polite)	Staff	–	–

sequences of words depending the function that describes. Another obstacle with the Greek language is that the vast majority of online text-based communications ignore the rules of spelling and grammar. After the analysis of the examined reviews, are identified alternative ways to write the same adjective (column 4, Table 3). For

Table 6 Transcription of Greeklish into Latin letters

Greek letters	Latin letters	Greeklish
Αι, αι	Ee	Ai/ai, Ee
Αυ/αυ	Av/av, Au/au	Av, au, Ay, Af
Ββ	Vv	Bb, Vv
Γγ	Gg, Yy	Gg, Yy
Εε	Ee	Ee
Ει, ει	Ii	Ii, Ei/ei
Ευ, ευ	Ev/ev, Eu/eu	Ev, Eu, Ey, Ef
Ηη	Ii	Ii, Hh
Ιι	Ii	Ii
Κκ	Kk, Cc	Kk, Cc
Ξξ	Xx	Xx, Jj, Ks/ks
Οο	Oo	Oo
Οι, οι	Oi, oi	Oi, oi
Ου, ου	Ou, ou	Ou, Oy
Υυ	Ii	IiYi
Υι, υι	Ii/ii	Ii, Ui, Yi
Φφ	Ff	Ff, Ph/ph
Χχ	Xx	Xx, Ch/ch
Ωω	Oo	Oo, Ww

instance, the adjective tasty was presented with the following five different ways: Nostimo, Nosthmo, Nostymo, Nosteimo, Nostoimo. This is an outcome of the complexity of the Greek alphabet. Table 6, cites an equivalent standard for transcription of Greek letters into Latin letters (Pedersen, 2009) and into Greeklish from the examined reviews.

Adapting the abbreviation Greek/Greeklish there are two ways to write the letter e (ε/e and αι/ai), four ways to write the letter i (ι/i, η/h, υ/y or υ/u and ει/ei), two ways to write the letter o (ο/o, ω/o or ω/w), and two ways to write the letter v (β/b and β/v), since in each one of the above cases the corresponding Greek letters are pronounced identically. All the above could help us to suggest a sentiment dictionary with Greek sentiment adjectives which may be used in lexicon-based techniques.

According to the proposed analysis, the transcription of Greek letters into Greeklish enables us to use the Greek dictionary in any recommended program that has already used in sentiment analysis in English language. The following procedure should be ideal. Firstly, the proposed lexicon should be saved in a .csv file (greeklexicon.csv). The alternative ways to write the same adjective must be registered as different (column 1 and 2, Table 3). Each row of the first column has to contain an adjective. Secondly, the data set of the examined reviews should be saved also as a .csv file (examinedreviews.csv). Each row has to contain a review that has to be written with Latin letters (Table 4). In order to avoid the double binary semantic orientation, all neutral and negative sentiments should be removed. Fourthly, from

each row in the file examined reviews.csv the frequencies of each adjective have to be computed. Finally, the reviews which contain one of the adjectives in the greeklexicon.csv should be saved in a new file as follows: (Company's Name, Adjective, Positive, User, and Date of review).

4 Conclusions

In Greece, more and more people use the social media networking to make a decision about purchasing a product or a service. On the other hand, the vast majority of companies use the social media marketing to improve their brands. Due to the large and continuous volume of data that produced users face difficulties to identify all the necessary information. The problem can be approached with the method of sentiment analysis. Sentiment analysis has a wide variety of techniques (supervised, unsupervised and lexicon based) to detect users' positive or negative sentiments. While a lot have been written and researched about sentiment analysis in various languages, Greek has not drawn researcher's attention. After the analysis of 275 customer reviews of 10 leader companies in Greece, a mini vocabulary of 15 positive adjectives and a transcription of Greek into Latin letters (and Greeklish idioms) was suggested. The proposed vocabulary should be used in any recommended technique to detect comments that are written in Greek language (or Greeklish) in social media networks in case of food and beverage sector. Future work will focus, primarily, on the extension of this study regarding all the parts of speech of Greek language (verbs, nouns, adverbs) and secondly, on the restrictions as presented in Sect. 3 (negation rules and spelling mistakes). Especially, a bigger sample of online reviews will help us to predict all the alternative ways to write Greek words as presented in Greeklish idioms in the case of social media networks.

Acknowledgments This research is co-financed by Greece and the European Union (European Social Fund- ESF) through the Operational Programme "Human Resources Development, Education and Lifelong Learning 2014–2020" in the context of the project "Strengthening Human Resources Research Potential via Doctorate Research—2nd Cycle" (MIS 5000432).

References

Adam, T. *Greek sentiment lexicon*. Accessed May 23, 2018., from http://socialsensor.eu/results/datasets/147-greek-sentiment-lexicon

Adeniyi, D., Wei, Z., & Yongquan, Y. (2016). Automated web usage data mining and recommendation system using K-nearest neighbor (KNN) classification method. *Applied Computing and Informatics, 12*(1), 90–108.

Agathangelou, P., Katakis, I., Kokkoras, F., & Ntonas, K. (2014). Mining domain-specific dictionaries of opinion words. In *International conference on web information systems engineering* (pp. 47–62). Cham: Springer.

Aggarwal, C. C., & Zhai, C. X. (2012). *Mining text data*. New York: Springer Science.
comScore/the Kelsey group. (2007). *Online consumer-generated reviews have significant impact on offline purchase behavior*. Press Release, Accessed May 23, 2018, from http://www.comscore.com/press/release.asp?press=1928
Cover, T. M., & Thomas, J. A. (1991). *Elements of information theory*. New York: Wiley.
Deerwester, S., Dumais, S., Landauer, T., Furnas, G., & Harshman, R. (1990). Indexing by latent semantic analysis. *JASIS, 41*, 391–407.
Duan, W., Cao, Q., Yu, Y., & Levy, S. (2013). Mining online user-generated content: Using sentiment analysis technique to study hotel service quality. In *System Sciences (HICSS), 46th Hawaii International Conference* (pp. 3119–3128).
Duric, A., & Song, F. (2012). Feature selection for sentiment analysis based on content and syntax models. *Decision Support Systems, 53*, 704–711.
Fan, T., & Chang, C. (2012). Blogger-centric contextual advertising. *Expert Systems, 38*, 1777–1788.
Feldman, R. (2013). Techniques and applications for sentiment analysis. *Communications of the ACM, 56*, 82–89.
Gabryel, M., Damaševičius, R., & Przybyszewski, K. (2018, October). Application of the bag-of-words algorithm in classification the quality of sales leads application. In *ICAISC 2018, LNAI 10841* (pp. 615–622). New York: Springer International.
Genier, C., Stamp, M., & Pfitzer, M. (2009). Corporate social responsibility for agro-industries development. In C. Da Silva, D. Baker, A. Shepard, C. Jenane, & S. Miranda-da-Cruz (Eds.), *Agro-industries for development* (pp. 223–251). Oxfordshire, UK: CABI.
Giatsoglou, M., Vozalis, M. G., Diamantaras, K., Vakali, A., Sarigiannidis, G., & Chatzisavvas, K. C. (2017). Sentiment analysis leveraging emotions and word embeddings. *Expert Systems with Applications, 69*, 214–224.
Gräbner, D., Zanker, M., Fliedl, G., & Fuchs, M. (2012). Classification of customer reviews based on sentiment analysis. In M. Fuchs, F. Ricci, & L. Cantoni (Eds.), *Information and communication technologies in tourism* (pp. 460–470). New York: Springer.
Griffiths, T. L., Mark, S., Blei, D. M., & Tenenbaum, J. B. (2005). Integrating topics and syntax. *Advances in Neural Information Processing Systems, 17*, 537–544.
Haseena, R., & Tanvir, A. (2014). Opinion mining and sentiment analysis - challenges and applications. *International Journal of Application or Innovation in Engineering & Management (IJAIEM), 17*, 25–29.
He, Y., & Zhou, D. (2011). Self-training from labeled features for sentiment analysis. *Information Processing and Management, 47*, 606–616.
Horrigan, J. A. (2008). *Online shopping pew internet & American life project report*. https://www.pewinternet.org/2008/02/13/online-shopping/
Hu, X., & Liu, H. (2012). Text analytics in social media. In *Mining text data* (pp. 385–414). Boston: Springer.
Joachims, T. (2001). A statistical learning model of text classification for support vector machines. In *Proceedings of SIGIR-01, 24th ACM International Conference on Research and Development in Information Retrieval* (pp. 128–136). New York: ACM.
Kalamatianos, G., Mallis, D., Symeonidis, S., & Arampatzis, A. (2015) Sentiment analysis of Greek tweets and hashtags using a sentiment lexicon. In *Proceedings of the 19th Panhellenic Conference on Informatics, Athens* (pp. 63–68). New York: ACM.
Kang, H., Yoo, S. J., & Han, D. (2012). Senti-lexicon and improved Naïve Bayes algorithms for sentiment analysis of restaurant reviews. *Expert Systems with Applications, 39*, 6000–6010.
Ko, Y., & Seo, J. (2000). Automatic text categorization by unsupervised learning. In *Proceedings of the 18th Conference on Computational Linguistics* (pp. 453–459). Stroudsburg, PA: Association for Computational Linguistics.
Kokkoras, F., Ntonas, K., & Bassiliades, N. (2013). DEiXTo: A web data extraction suite. In *Proceedings of the 6th Balkan Conference in Informatics, Thessaloniki* (pp. 9–12). New York: ACM.

Lewis, D. D. (1998). Naïve (Bayes) at forty: The independent assumption in information retrieval. In *Proceedings of ECML-98, 10th European Conference on Machine Learning* (pp. 4–15). Berlin: Springer.

Li, N., & Wu, D. D. (2010). Using text mining and sentiment analysis for online forums hotspot detection and forecast. *Decision Support Systems, 48*, 354–368.

Liapakis, A., Costopoulou, C., Tsiligiridis, T., & Sideridis, A. (2017). Studying corporate social responsibility activities in the agri-food sector: The Greek case. *International Journal of Agricultural and Environmental Information Systems (IJAEIS), 8*, 1–13.

Liu, B. (2010). Sentiment analysis and subjectivity. In N. Indurkhya & F. J. Damerau (Eds.), *Handbook of natural language processing* (Vol. 2, pp. 627–666). Boca Raton: Chapman and Hall/CRC.

Medhat, W., Hassan, A., & Korashy, H. (2008). Combined algorithm for data mining using association rules. *Ain Shams Journal of Electrical Engineering, 1*(1), 1–12.

Morinaga, S., Yamanishi, K., Tateishi, K., & Fukushima, T. (2002). Mining product reputations on the web. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Edmonton* (pp. 341–349). New York: ACM.

Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up? Sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing* (Vol. 10, pp. 79–86). Stroudsburg, PA: Association for Computational Linguistics.

Pedersen, T. T. (2009). *Transliteration of non-roman scripts*. Accessed May 23, 2018.

Rainie, L., & Hitlin, P. (2004). *The use of online reputation and rating systems*. Washington, DC: Pew Internet and American Life Project.

Read, J., & Carroll, J. (2009). Weakly supervised techniques for domain in dependent sentiment classification. In *Proceeding of the 1st International CIKM Workshop on Topic-Sentiment Analysis for Mass Opinion* (pp. 45–52). New York: ACM.

Reyes, A., & Rosso, P. (2012). Making objective decisions from subjective data: Detecting irony in customer reviews. *Decision Support Systems, 537*, 754–760.

Shalev-Shwartz, S., Singer, Y., Srebro, N., & Cotter, A. (2011). Pegasos: Primal estimated sub-gradient solver for SVM. *Mathematical Programming, 127*, 3–30.

Sharma, A., & Dey, S. (2012). A comparative study of feature selection and machine learning techniques for sentiment analysis. In *Proceedings of the 2012 ACM Research in Applied Computation Symposium* (pp. 1–7). Raleigh, NC.

Singh, J., Singh, G., & Singh, R. (2016). A review of sentiment analysis techniques for opinionated web text. *CSI Transaction on ICT, 4*, 241–247.

Turney, P. D. (2002). Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics* (pp. 417–424). Philadelphia, PA.

Vinodhini, G., & Chandrasekaran, R. M. (2012). Sentiment analysis and opinion mining: A survey. *International Journal of Advanced Research in Computer Science and Software Engineering, 2*, 282–292.

Xianghua, F., Guo, L., Yanyan, G., & Zhiqiang, W. (2013). Multi-aspect sentiment analysis for Chinese online social reviews based on topic modeling and how net lexicon. *Knowledge-Based Systems, 37*, 186–195.

Yelena, M., & Padmini, S. (2011). Exploring feature definition and selection for sentiment classifiers. In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*. Barcelona, Catalonia.

Yu, B., Zhou, J., Zhang, Y., & Cao, Y. (2017). Identifying restaurant features via sentiment analysis on yelp reviews. *arXiv preprint arXiv:1709.08698*.

Zhang, Z., Ye, Q., Zhang, Z., & Li, Y. (2011). Sentiment classification of internet restaurant reviews written in Cantonese. *Expert Systems with Applications, 38*, 7674–7768.