A Sentiment-Based Hotel Review Summarization Using Machine Learning Techniques

Agorakis Bompotas, Aristidis Ilias, Andreas Kanavos(✉), Christos Makris,
Gerasimos Rompolas, and Alkiviadis Savvopoulos

Computer Engineering and Informatics Department,
University of Patras, Patras, Greece
{mpompotas,aristeid,kanavos,makri,robolas,asavv}@ceid.upatras.gr

Abstract. With the advent of social media, there is a data abundance so that analytics can be reliably designed for ultimately providing valuable information towards a given product or service. In this paper, we examine the problem of classifying hotel critiques using views expressed in users' reviews. There is a massive development of opinions and reviews on the web, which invariably include assessments of products and services, and beliefs about events and persons. In this study, we aim to face the problem of the forever increasing amount of opinionated data that is published in a variety of data sources. The intuition is the extraction of meaningful services despite the lack of sufficient existing architectures. Another important aspect that needs to be taken into consideration when dealing with brand monitoring, relates to the rapid heterogeneous data processing, which is vital to be implemented in real-time in order for the business to react in a more immediate way.

Keywords: Classification · Machine learning · Neural networks · Opinion mining · Sentiment analysis

1 Introduction

In the emergent field of Web 2.0, users' reviews, comments and reports have constituted a crucial area of interest for tourism businesses. As an escalating number of consumers are inclined to share as well as exchange their personal experiences in social networks, forums and websites, the tourism industry has radically altered the way of increasing and influencing customers' engagement with tourism brands [10]. In this new era of e-tourism, businesses, in order to keep up with the upsurging competition, have to develop innovative marketing strategies and techniques focused on customers' needs and satisfaction. Subsequently, in the last years, there has been a wide interest in extracting actionable insights on customers' behaviour and sentiment by leveraging user-generated content, that will enable businesses to identify and in following predict the usefulness of online reviews [4].

I. Maglogiannis et al. (Eds.): AIAI 2020 Workshops, IFIP AICT 585, pp. 155–164, 2020.
https://doi.org/10.1007/978-3-030-49190-1_14

The widespread use of social media platforms has significantly contributed to the growth of the electronic word-of-mouth (eWOM) communication, which has notable impact on the tourism industry. This textual kind of communication is encapsulated in online reviews and has attracted researchers' interest in various domains. In particular, text and opinion mining systems have been proposed in the literature in order to analyse and classify customers' reviews, providing thus businesses with the capability of monitoring their online brand reputation [17,20]. Moreover, due to the large volume of user-generated data, text summarization techniques have also been proposed in order to effectively and efficiently identify the top-k most informative sentences of hotel reviews [12].

In general, despite the textual information, reviews consist of a score rating mechanism, which can reflect the overall customer satisfaction in a very explicit way. Although customers' ratings have been found to be highly correlated with the sentiment polarity of the corresponding textual content of the reviews [5], there is still a strong interest in further examining and evaluating the textual content under specific technical attributes, which can influence customer ratings [22]. In any case, it is obvious that customers' reviews are a vital source of information for the tourism industry, as they enable businesses to have a clear view of the most important aspects deriving from them and thus they can better prioritize and optimize their marketing strategies.

The vast amount of user-generated content has led to the need of NoSQL databases in order to manipulate them in a scalable and productive way [21]. Therefore, in this paper, a NoSQL system with an automated tool that generates an intelligent mechanism for analysing data and exporting useful knowledge and insights to tourism traders, is proposed. As a result, businesses will be able to adjust their marketing strategies and adapt to the customer needs in time, and simultaneously reducing their human resource requirements.

The contribution of this work lies in the design of a new approach for analyzing hotel reviews using Latent Dirichlet Allocation (LDA) for aspect mining and Neural Networks (NN) for sentiment analysis. A dynamic architecture, which receives the data stream, on-line or off-line in order not to overload the systems of the participating hotels or their service providers, is proposed. It extracts the aspects along with the sentiment of the hotel reviewers by applying LDA and NN modules accordingly, then stores the data and finally, attempts to correlate the data with the reviewers. The process is not obvious, given the anonymity of the reviewers, but the attempt to correlate them can be implemented with extensive training of the NN. Our architecture proposes a novel platform utilizing the benefits of both algorithms, so that it can be used in an effective way in data forecasting.

The paper is organized as follows. Section 2 describes related work. The key design ideas and concepts of the architecture of the proposed Sentiment Analysis system are presented in Sect. 3. In Sect. 4 we present implementation details regarding the system infrastructure, while in Sect. 5 we provide our conclusions and thoughts relating to future work.

2 Related Work

Due to the growing available data that are generated on a daily basis from hotels worldwide, a turn of attention has been observed in academic literature in adopting new ways of managing the insightful hotel data, and extrapolate important and valuable information, which can later be used for sustainable economic growth. The nature of most of these data are mainly in the form of text, accompanied by a certain numerical grading. These two characteristics constitute a modern review regarding the user's accommodation. Such reviews contain in their main body the reviewer's opinion of the hotel as well as a grade that indicates the polarity or the sentiment towards the accommodation, and wholly characterizes the experience itself.

In [17], an overview of a review management tool is shown where a variety of hotel comments were collected, in order to hark the visitors' points and views of the hotel quality. At the same point, the work presented in [11] introduces a more general and non context-specific approach for opinion mining, based on customer reviews. More specifically, authors performed a summarizing of the numerous comments and reviews regarding particular products, and in following extracted a comprehensive polarity percentage that represents the sentiment of the buyers as a whole. Altogether, the sentiment of a review as well as the general sentiment that hotel reviews accumulate, can produce a meaningful abstraction and pinpoint either problems that the management can solve or aid potential customers in choosing their next hotel [18].

Social media compose platforms that welcome a vast amount of product and service reviews, which have an unbeatable advantage over classic comments under the designated product; more to the point, the graph representation and the links between reviews can provide deeper latent connections between sentiment and review. One such corresponding work has been implemented in [14], where authors demonstrated numerous academic researches that concentrate on consequential sentiment analysis through the scope of social media. Coauthoring an extension of their previous work, authors in [16] also showed the scalability of their methodologies, where massive collections of review data were processed with the aid of distributed computing frameworks, while maintaining robustness in terms of velocity of processing and accuracy of sentiment prediction.

However, the process of analyzing the sentiment in general and in retrospect, the sentiment polarity of reviews is not straightly performed into the raw collected data. A number of pre-processing layers must be beforehand executed so in the work presented in [9], this importance is greatly highlighted. Before advancing to the classification and performance evaluation steps, two important layers take place; data transformation and filtering. Data were cleaned and stripped of useless tags, and in following, stemming and lemmatization procedures occurred. During the filtering step, a statistical analysis to measure the dependency between word and category that the word is included, was performed with the aid of the Chi-square test. As shown in the performance evaluation step, all measurements were improved when considering the pre-processing procedure in comparison to completely avoiding that step, in terms of the three basic evaluation metrics, namely $F1$-measure, accuracy and recall.

As a result, the review management is immediately dependent on the aforementioned nature of the reviews, which is none other than a text collection. Thus, the branch of text mining as a tool to aid this process is deemed essential. There has been much academic literature throughout the latest decades regarding text mining and opinion mining techniques, either of probabilistic nature or not [1,3]. As a previous work on opinion clustering emerging in reviews, one can consider the setup presented in [6]. Other existing works that deal with customers' buying habits is presented in [13,15].

The machine learning algorithms have the advantage of dealing with high dimensional and nonlinear relationships, which is especially suitable for establishing train dynamic model and train speed prediction on account of the dynamic and nonlinear nature [19]. One of the most classic text mining techniques that composed the foundation for modern opinion mining is the Latent Dirichlet Allocation (LDA) [2,7,8]. LDA is a probabilistic algorithm that can discover the latent topics that may exist within the reviews of the collection. More specifically, LDA extracts the top N topics that are most common in a review, based on the representations of the most frequent words with the input being a term document matrix, whereas two distributions are considered as output; one for document-topic relations and the other for topic-word ones.

3 System Architecture

In this section, we will elaborate on the key design ideas and concepts regarding the architecture of the proposed Sentiment Analysis platform. As depicted in Fig. 1, our system consists of an Application Programming Interface (API) that serves as the gateway to an online hotel booking platform (or a channel manager), a NoSQL database and the Sentiment Analysis Infrastructure which in turn is divided into five different modules.

The flow of data within the system is relatively simple. Initially, hotel reviews are inserted in the database through the corresponding API, where they become available to the Sentiment Analysis Infrastructure. The Natural Language Processing module initially parses the stored reviews, transforms them into the appropriate form and eventually passes them to the Aspect Mining and Sentiment Analysis modules. Subsequently, the intermediate results are given as input to the Results Combiner module, which produces the final outputs and stores them back to the database. Finally, both the initial reviews and the results of the analysis are easily accessible through the API.

3.1 Hotel Reviews Sentiment Analysis API

For exposing the Sentiment Analysis Infrastructure to the systems that will access it, a RESTful web service was implemented. Concretely, it supports both GET and POST requests and the data exchanged over the API are expected to be in the form of JavaScript Object Notation (JSON). The API allows the user

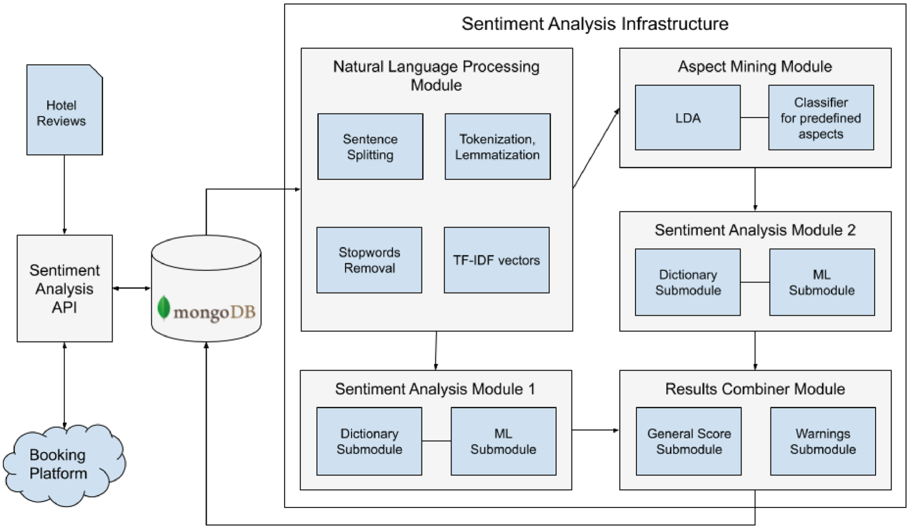

Fig. 1. Hotel reviews sentiment analysis' platform architecture

to insert new hotel reviews to the system and update or simply retrieve information about the old ones. The aforementioned functionality can be extended as well to the authors of the reviews. The access to the API is restricted and an authentication method is utilized for avoiding illwanted visiting. Regarding the API implementation, Python programming language in combination with the popular Flask microframework were selected, because of their lightweight and powerful properties.

3.2 Storage

For storing the hotel reviews along with the intermediate and final results produced by the Sentiment Analysis Infrastructure, a fast but flexible database system was required. Furthermore, the risk that the number of reviews can grow exponentially created the additional need for scale-out, which dictated us to swift away from the traditional RDBMs and to seek a NoSQL alternative. As no join operations were compulsory for our analysis, a document-based database was deemed as the best suited solution. More specifically, we opted to use the MongoDB, because of its maturity level along with the robustness it provides.

3.3 Sentiment Analysis Infrastructure

3.3.1 Natural Language Processing Module

The first module of the Sentiment Analyzer is responsible for processing the raw text of the hotel reviews with the aim of producing the vectors that will be used as input for the next modules. This process, which is also known as feature extraction, can be further analyzed into four separate stages: Sentence Splitting, Tokenization, Stopwords Removal and Tf/Idf Calculation.

In most cases, the importance of a review content is condensed within the first few sentences of a text. Therefore, tokenization of sentences is deemed as crucial, as well as extracting the sentences where the necessary information resides. Often, a simple split of the text string on every dot will suffice. However, there are many cases in the English language, where the period punctuation is used for reasons other than ending a sentence such as timestamp depiction (a.m., p.m.). That is why the sentence splitting module must be context-aware.

Regarding the tokenization of the text body, the first step is splitting the sentence into separate words by using the space string to perform the splitting. These words are generally referred to as tokens, which is no other than a sequence of characters grouped together to form a semantic unit. However, not all terms or units are useful for the final analysis, since there are words that are deemed trivial or others that contribute to the general context far more heavily.

A useful technique to apply to the extracted tokens is the stemming and lemmatization technique. Many words contained in the same document correlate, since they belong to similar derivational families and have similar meanings. Thus, it is deemed useful to extract a general lemma that applies to all words with similar prefixes, and remove the corresponding suffix. This procedure allows any Natural Processing Module to work faster, since it handles exceptionally less words in a smaller vocabulary, without sacrificing any percentage of context or meaningful information.

After the application of the above mentioned procedures, *Tf/Idf* weighting can be performed. There the term frequency definition is combined with the inverse document frequency, in order to produce a specific weight for every word that exists inside each document. This assigned weight takes higher values when the terms occur many times inside a few documents, and lower when the term occurs less times in a document or generally occurs in various documents. When the word exists in every document, it is deemed borderline meaningless.

Each stage is assigned to a sub-module designed and developed specifically for the corresponding task. In addition, these sub-modules are connected in order to form a pipeline that receives as input the collection of the reviews and produces the final vectors.

3.3.2 Aspect Mining Module

The Aspect Mining Module aims to detect the aspect where each sentence of the review refers to. In order to achieve this goal, both a supervised and an unsupervised learning approach are employed. Initially, each sentence is labeled by a Multiclass Classifier. The aspects are simply considered as the predefined labels, which are commonly found in hotel reviews (e.g., cleanliness, facilities, etc.) with the addition of an "undefined" class. However, some extra analysis is required in order to discover potential aspects that were omitted.

This analysis is performed by the second submodule of the Aspect Mining Module, which employs Latent Dirichlet Allocation (LDA). LDA is considered a generative probabilistic model of a collection of composites, made up of parts. It is designed based on the idea that each document in a collection can be described by a distribution of topics and simultaneously, each topic can be described by a distribution of words. In our scenario, where hotel reviews are in place of documents and the aim is to discover the aspects that characterize them, it is evident why LDA constitutes a perfect fit to our problem.

3.3.3 Sentiment Analysis Module

In the proposed system, there are two Sentiment Analysis Modules; the first one characterizes the whole review based on the polarity of the sentiment that expresses, while the second one attempts to do the same but for every aspect mentioned. The sentiments are both detected with the use of predefined rule-based sentiment annotators as well as machine learning models. The output of these modules constitutes a vector of sentiment scores that are then passed to the Results Combiner Module.

Regarding the training of the aforementioned models, a number of classifiers was employed and the Long Short Term Memory Neural Networks (LSTMs) emerged as the most efficient solution. Compared to the standard Neural Networks, which only allow information flowing forward from the input nodes to the output nodes, LSTMs are equipped with feedback connections. More to the point, based on this feature, LSTMs are able to effectively handle entire sequences of data and are best suited in fields such as speech recognition and text processing. Furthermore, LSTMs use "exploit" for regulating the flow of data within their cells in order to deal with the exploding and vanishing gradient problems that are the most common shortcomings of the traditional Recurrent Neural Networks (RNN).

3.3.4 Results Combiner Module

As derived from its name, the Results Combiner Module gathers the information about the extracted sentiment of reviews by the previous modules and attempts to produce an insight useful to the end user. It consists of two different submodules; the first one calculates a score of the review in order to quantify the overall customer satisfaction, whereas the second one issues warnings that might help the hoteliers to understand their shortcomings.

4 Implementation

The functionality of each aforementioned module is widely enhanced through the advancements of machine learning algorithms, whereas their efficiency, robustness and velocity of training have been proved throughout the scientific literature of the last decade. In this experimental work, the detection of the sentiment in each hotel review, as well as their polarity related to aspects that pertain them, is made feasible with the use of the following algorithms.

4.1 Long Short Term Memory (LSTM) Neural Networks

Neural Networks' capability has made possible the successful supervised training of classifiers on collections of data big enough; this training can ensure a wide captivation of patterns in the corresponding dataset. Given well structured data, the Long Short Term Memory Neural Networks (LSTM) can provide consequential classifications, especially when sequences must be managed. The case at hand can be directly associated with sequences since hotel review collections consist of text data, whilst their labels are the polarized sentiments of the document that need to be detected.

LSTMs' architecture is based on "*cell state*" and "*gates*" through which the input information is propagated. More accurately, there are three gates and two states in LSTMs: the *forget gate (f_t)* whose responsibility is to remove unnecessary information from the *cell state* taking as input the hidden state of the previous cell h_{t-1} and data record x_t. Next, the *input gate (i_t)* adds new information on the cell state by creating a vector of all possible values and multiplying them with the *tanh* function. In following, the *output gate (o_t)* transfers a percentage of information to the hidden state, so that the LSTM can maintain its robustness by preserving the long-term dependencies. Finally, regarding the two states, the *cell state (c_t)* represents the internal memory of the cell and the *hidden state (h_t)* decides the time frame of the recalled dependency. These characteristics are presented in the following Eq. 1.

$$
\begin{aligned}
f_t &= \sigma(W_f(x_t + h_{t-1}) + b_f) \\
i_t &= \sigma(W_i(x_t + h_{t-1}) + b_i) \\
o_t &= \sigma(W_o(x_t + h_{t-1}) + b_o) \\
c_t &= f_t c_{t-1} + i_t \sigma(W_c(x_t + h_{t-1}) + b_c) \\
h_t &= o_t tanh(c_t)
\end{aligned}
\tag{1}
$$

4.2 Latent Dirichlet Allocation (LDA)

The fundamental functionality of Latent Dirichlet Allocation (LDA) resides in the idea that each document has a number of words (terms as a subset of the total word collection), which are partially involved in the said document as well as a number of topics, which are elaborated in this document. The detection of topics that the document consists of as well as the correlation of its words with those topics, can be considered as the utmost goal of LDA.

In this work, the hotel reviews can be seen as documents, whereas the aspects as topics, which the review is meant to analyze. LDA will parse each document and allocate each review word to a specific topic through a parameterized Dirichlet allocation, allowing thus a degree of initial randomness. Afterward, two calculations are performed: initially the number of words in the review that were assigned a specific topic divided by the number of words in any topic, as well as the percentage of allocations given a topic that were derived through a given word.

As a result, the probability of a word belonging to a topic is updated by the multiplication of the two aforementioned values. The re-assignment of documents to words and topics takes place until convergence and a stable mixture of topics are produced.

5 Conclusions and Future Work

The aim of the current study is the design of a new schema based on NoSQL databases for the manipulation of hotel reviewers' comments along with the appropriate modules based on LDA in terms of aspect mining as well as Neural Networks for hotel reviewers' sentiment. In this schema, the data stream is initially received, then the aspects along with the sentiment of the hotel reviewers are extracted and finally, the data with the reviewers are correlated. We proposed a novel architecture utilizing the benefits of Neural Networks along with LDA algorithm so that it can be used in an effective way in data forecasting.

For further work, we would like to compare the effectiveness of our architecture in larger sample. In addition, new classifiers can be considered in order to be compared with the current ones, such as Random Forest, Support Vector Machines, etc.

Acknowledgement. Agorakis Bompotas, Aristidis Ilias, Andreas Kanavos, Christos Makris, Gerasimos Rompolas and Alkiviadis Savvopoulos have been co-financed by the European Union and Greek national funds through the Regional Operational Program "Western Greece 2014-2020", under the Call "Regional Research and Innovation Strategies for Smart Specialisation - RIS3 in Information and Communication Technologies" (project: 5038701 entitled "Reviews Manager: Hotel Reviews Intelligent Impact Assessment Platform").

References

1. Blei, D.M.: Probabilistic topic models. Commun. ACM **55**(4), 77–84 (2012)
2. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet allocation. J. Mach. Learn. Res. **3**, 993–1022 (2003)
3. García, S., Luengo, J., Herrera, F.: Data Preprocessing in Data Mining. Intelligent Systems Reference Library, vol. 72. Springer, Heidelberg (2015). https://doi.org/10.1007/978-3-319-10247-4
4. Gavilan, D., Avello, M., Martinez-Navarro, G.: The influence of online ratings and reviews on hotel booking consideration. Tour. Manag. **66**, 53–61 (2018)
5. Geetha, M., Singha, P., Sinha, S.: Relationship between customer sentiment and online customer ratings for hotels - an empirical analysis. Tour. Manag. **61**, 43–54 (2017)
6. Gourgaris, P., Kanavos, A., Makris, C., Perrakis, G.: Review-based entity-ranking refinement. In: 11th International Conference on Web Information Systems and Technologies (WEBIST), pp. 402–410 (2015)
7. Griffiths, T.L.: Gibbs sampling in the generative model of latent Dirichlet allocation (2002)

8. Griffiths, T.L., Steyvers, M.: Finding scientific topics. Proc. Nat. Acad. Sci. **101**(suppl 1), 5228–5235 (2004)
9. Haddi, E., Liu, X., Shi, Y.: The role of text pre-processing in sentiment analysis. In: 1st International Conference on Information Technology and Quantitative Management (ITQM), pp. 26–32 (2013)
10. Harrigan, P., Evers, U., Miles, M., Daly, T.: Customer engagement with tourism social media brands. Tour. Manag. **59**, 597–609 (2017)
11. Hu, M., Liu, B.: Mining opinion features in customer reviews. In: Proceedings of the Nineteenth National Conference on Artificial Intelligence (AAAI), pp. 755–760 (2004)
12. Hu, Y., Chen, Y., Chou, H.: Opinion mining from online hotel reviews - a text summarization approach. Inf. Process. Manag. **53**(2), 436–449 (2017)
13. Iakovou, S.A., Kanavos, A., Tsakalidis, A.: Customer behaviour analysis for recommendation of supermarket ware. In: Iliadis, L., Maglogiannis, I. (eds.) AIAI 2016. IAICT, vol. 475, pp. 471–480. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-44944-9_41
14. Kanavos, A., Perikos, I., Hatzilygeroudis, I., Tsakalidis, A.: Emotional community detection in social networks. Comput. Electr. Eng. **65**, 449–460 (2018)
15. Kanavos, A., Iakovou, S.A., Sioutas, S., Tampakas, V.: Large scale product recommendation of supermarket ware based on customer behaviour analysis. Big Data Cognit. Comput. **2**(2), 11 (2018)
16. Kanavos, A., Nodarakis, N., Sioutas, S., Tsakalidis, A., Tsolis, D., Tzimas, G.: Large scale implementations for twitter sentiment classification. Algorithms **10**(1), 33 (2017)
17. Kasper, W., Vela, M.: Sentiment analysis for hotel reviews. In: Computational Linguistics-Applications Conference, pp. 45–52 (2011)
18. Liu, B., Zhang, L.: A survey of opinion mining and sentiment analysis. In: Aggarwal, C., Zhai, C. (eds.) Mining Text Data, pp. 415–463. Springer, Heidelberg (2012). https://doi.org/10.1007/978-1-4614-3223-4_13
19. Savvopoulos, A., Kanavos, A., Mylonas, P., Sioutas, S.: LSTM accelerator for convolutional object identification. Algorithms **11**(10), 157 (2018)
20. Sun, Q., Niu, J., Yao, Z., Yan, H.: Exploring ewom in online customer reviews: sentiment analysis at a fine-grained level. Eng. Appl. Artif. Intell. **81**, 68–78 (2019)
21. Vonitsanos, G., Kanavos, A., Mylonas, P., Sioutas, S.: A NoSQL database approach for modeling heterogeneous and semi-structured information. In: 9th International Conference on Information, Intelligence, Systems and Applications (IISA), pp. 1–8 (2018)
22. Zhao, Y., Xu, X., Wang, M.: Predicting overall customer satisfaction: big data evidence from hotel online textual reviews. Int. J. Hosp. Manag. **76**, 111–121 (2019)