Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 196 (2022) 199–206

Procedia Computer Science

www.elsevier.com/locate/procedia

CENTERIS - International Conference on ENTERprise Information Systems / ProjMAN - International Conference on Project MANagement / HCist - International Conference on Health and Social Care Information Systems and Technologies 2021

Customer reviews sentiment-based analysis and clustering for market-oriented tourism services and products development or positioning

Sandra Jardim*, Carlos Mora

Smart Cities Research Center, Polytechnic Institute of Tomar, Quinta do Contador, 2300-313 Tomar, Portugal

Abstract

This paper proposes a method that allows the clustering and identification of similarities between users of a digital tourism platform, through the extraction of the sentiments expressed by them in the reviews or comments registered and the subsequent automatic clustering of the users, according to the polarity of sentiments subjectively expressed in their posts. This research fills a gap in the text mining literature for the development, improvement and/or reorientation of services and products in the field of tourism, providing a method to explore the needs and desires of the client based on their digital footprint drawn from posts and reviews about the service or product in question. The sentiment analysis is detailed, comprehending language detection and some specific language syntax treatment, with a subsequent explanation of the clustering algorithm used. The developed algorithm was tested in the user's segmentation and sentiment analysis of their publications on a digital tourism platform. The results obtained demonstrate the efficiency of the solution, which presents a high accuracy in the classification of publications in four different languages and in the user's segmentation process.

Keywords: sentiment analysis; automatic sentiment classification; lexicon-based approach; customer reviews; customer segmentation.

* Corresponding author. Tel.: +351 249 328 100; fax: +351 249 328 186.
E-mail address: sandra.jardim@ipt.pt

10.1016/j.procs.2021.12.006

1. Introduction

Tourism, which has positive impacts and advantages on countries' economy and infrastructure, bringing growth and development, presents itself as an industry strongly connected to a range of global industries and sectors ranging from commerce to cultural preservation and ecological conservation. Tourism has a high economic value, boosting a country's GDP through different ways, also promoting social gains, such as the preservation of local culture and the strengthening and empowerment of communities, and environmental gains, such as the preservation of nature, as well as political gains, through the creation of opportunities for international collaboration, partnerships, or agreements between countries. In addition to the tourism's economic value, it has high importance for those who practice it, the tourists. Engaging in tourism, between one of the more than 150 different types of tourism available, tends to improve people's quality of life, physical and emotional well-being, and increase educational value.

Choosing a touristic destination involves making a decision that depends on several factors, from personal taste, economic capabilities, time availability, professional and family responsibilities, among others. Usually, decision making is defined as a conscious process of making choices among one or more alternatives with the intention of moving towards some desired goal [1]. However, and taking in account that human beings tend to give great importance to what others think and feel, both in the personal and professional fields [2, 3], not only pure rational facts interfere in the decision-making process. Other people's opinions tend to carry important weight when making an individual's decision. This is true in many contexts, including tourism. In the process that leads us to choose a touristic destination, we tend to analyze the comments or reviews of those who have already been to the destination we are considering, whose opinions tend to influence our decision.

In a global economy that is characterized by globalization and a fast technological evolution, the commitment of organizations on the search for greater competitiveness translated into gains in productivity and improvement in quality, is increasing. The success of organizations depends not only on factors such as the quality of the services provided or the products sold, but also on their ability to meet the needs of customers. In this sense, it is extremely important to know the customers and their opinions regarding the services obtained, to act as to prioritize the processes that create organizational value. This is true in the most varied contexts, going from the area of services, such as healthcare, to the area of product commercialization, being tourism one of the sectors where the analysis of customer's assessments and reviews leads to the acquisition of complementary knowledge that can be used by organizations to refine their marketing strategies and improve their services, leading to the improvement of their performance, strengthening their credibility and competitiveness.

The article is organized as it follows. Section 2 reviews the theoretical basis of the proposed method. Section 3, which explains the proposed method, presents three subsections, in which the method used to capture the sentiments expressed by users regarding a given item (service or product), the clustering of users based on sentiment analysis and its subdivision based on certain characteristics are explained. Section 4 presents the results of the proposed method when applied to a regional tourism platform and section 5 presents conclusions and future work.

2. Theoretical background

Knowing customer's opinions, attitudes and emotions regarding products and services, helps organizations to understand their customers' satisfaction degree, which is very important in the decision-making process [4]. With this knowledge, organizations can anticipate or change their commercial strategies and adapt themselves to the evolution of the market and their needs. On the other hand, the increasing use of digital technologies in the relationship between customers and organizations, provides a large amount of data, in the form of reviews, opinion texts and complaints, from which is possible to extract the knowledge needed to improve the organizational decision-making process or the services and products market positioning.

2.1. Sentiment Analysis

Sentiment analysis [5], also known as opinion mining, has attracted great interest from the scientific community, enhancing the investigation of automatic classification methods of sentiments subjectively expressed in customers' reviews, opinion texts and assessments [6, 7]. For organizations, its relevance applies in different areas, such as:

analysis of consumer's buying patterns [8, 9]; collecting customer's feedback on social media, websites or online forms [10]; obtaining knowledge about the stimuli that create the greatest impact on people [11]; understanding the factors that motivate people to like a product or service [12]; conducting research market [13]; categorizing customer's service requests; predicting consumer's behavior, among others [14].

Sentiment analysis is a textual and visual automated process that classifies information by detecting, extracting, and classifying opinions, according to the data polarity (positive, negative, and neutral) [15, 16] but also on sentiments and emotions (angry, happy, sad, etc.), urgency (urgent, not urgent) and even intentions (interested vs. not interested). It can be performed at different levels [17], such as: document level [18], capturing the overall sentiments expressed in the text, sentence level [19], classifying the polarity of each sentence in the text, characteristic level [20], analyzing the polarity of opinions on characteristics/attributes of the object, and aspect level [21], finding and aggregating sentiment on entities mentioned within documents or aspects of them.

A sentiment analysis process can be structured in five main procedures: i) data extraction; ii) pre-processing; iii) sentiment detection; iv) sentiment classification; and v) polarity report, that displays the results of a sentiment analysis in several possible ways.

In a sentiment analysis process, one of the main applied techniques are lexicon-based approaches, where the classification task is made using semantic-based unsupervised methods. Lexicon-based semantic orientation approach, also known as dictionary or knowledge-based, calculates the sentimental orientation of a document from the semantic orientation of the words or phrases that compose it [22]. As it is an unsupervised approach, it does not require an initial training of data, using instead a pre-defined list of words, to which a specific sentiment is associated. The overall sentiment of the text is calculated based on the count of positive and negative words present in it [23]. The analysis of sentiments using lexicon-based approaches has the advantages of simplicity of understanding, implementation, and efficiency, both in the use of computational resources and in the ability to predict. However, this approach has a limitation in terms of generalization since each solution is created according to the context in which it will be applied. This approach does not require labeled data, but implies the construction of a lexical dictionary, which constitutes one of its main challenges, since the need to consider application contexts makes it difficult to use a single and board lexicon. Added to this difficulty is the large volume of data that normally has to be analyzed, as well as the variability of the language used. In lexical approaches is essential a pre-processing step on the textual content, to extract relevant information about the text for later textual and sentiment analysis. This is particularly important given that the change in the grammatical characteristic of a word can change the meaning and intensity of the sentiment involved in it. Another important pre-processing technique is text segmentation (tokenization), to transform the text into a set of terms extracted from the texts (tokens), both at word and phrase level. Other pre-processing techniques are also extremely important in lexical approaches such as stemming, stop words removal, and uppercase and lowercase letters transformation.

2.2. Clustering for Customer Segmentation

Customer segmentation is a technique used to separate the customer base into groups where customers belonging to a certain group have similar characteristics, while customers from different groups have as different characteristics as possible. For organizations it is an important process, as it allows them to divide their customers into groups according to their preferences, characteristics, and behavior towards the organization [24], allowing them to study groups of customers with similar needs, preferences, habits, e.g., and thus adapt their strategy to improve its relationship with customers and hence their business performance [25].

Clustering, also known as cluster analysis, has great importance in data mining [26] and can be considered one of the most important unsupervised learning problems. Its objective is the organization of unlabeled data into similarity groups called clusters. In addition to detecting similarities in the data, it is the objective of unsupervised clustering methods to determine the appropriate number of clusters, a repeatable and iterative task, where large amounts of raw data are analyzed in search of similarities and patterns [27]. To achieve good results, the choice of the clustering algorithm to be used must consider the application domain [28]. Among the various existing hierarchical and non-hierarchical clustering algorithms, the most used in customer segmentation are K-Means [29] and Agglomerative Hierarchical Clustering [30].

3. Methodology

In this paper, is described a method for users automatic clustering based on the sentiment polarities of their reviews on a digital tourism platform. Theses reviews are extracted from a relational database with metadata detailing the point of interest addressed by the reviews and their author.

For the user clustering process, we propose a method that uses the K-Means algorithm, using the Elbow method to obtain the number of user clusters, allowing the creation of different groups of users, according to some characteristics, detailed further in this paper.

For the classification of sentiment polarity of posts, we propose an unsupervised automatic method of multilingual lexicon-based sentiment analysis algorithm, where a dictionary-based approach is used. This approach consists of using a predefined lexicon that contains positive and negative words. The frequencies of the words extracted from each of the texts to classify are calculated, as to be possible scoring the global sentiment of each text in the data set.

The developed algorithm classifies the sentiments or opinions of the digital platform users, by analyzing their publications, comments, or reviews and referencing these with the user clusters previously calculated, allowing for a more rich and contextualized analysis.

The performance of the developed algorithm was tested using data from the cultural and social information system named Cuscarias, a digital platform based on the concepts of co-experience and co-creation, which is accessible to users through a mobile device [31]. This information system allows the worldwide creation of as many Cuscarias as desired, where each one identifies a touristic point of interest, such as a place, a monument, an event, among others.

The core architecture of the proposed approach is illustrated in Fig. 1.

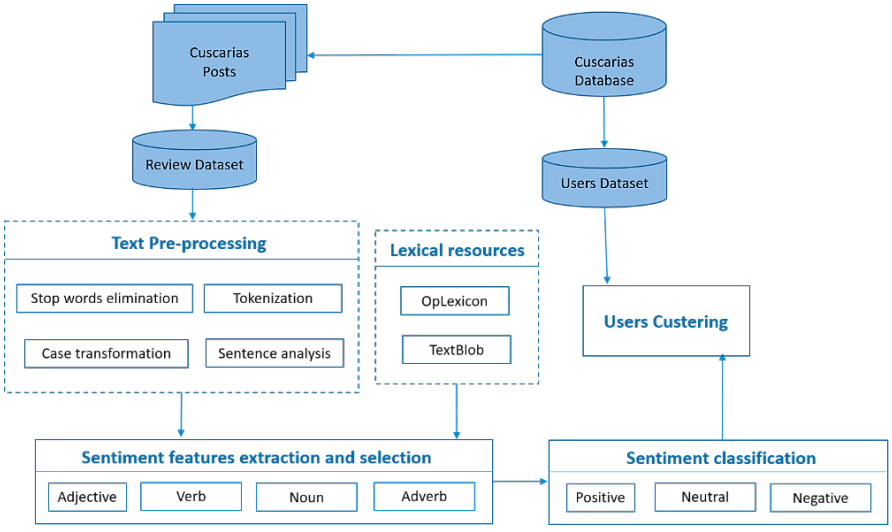

Figure 1. Architecture of the proposed solution.

4. Proposed approach

To implement the proposed algorithm, it was used the Python programming language, where the possibility of importing modules facilitates and simplifies the evocation of methods to be used. The solution architecture follows a file organization made up of 5 Python files – API_sentiment, Cron_sentiment, EN_sentiment, PT_sentiment, and Translation – and a text file, corresponding to the dataset used in the sentiment analysis of Portuguese publications (or comments).

As previously mentioned, in this work we are focusing on the data stored in the Cuscarias cultural and social information system, so the data set is built from the comments recorded for each of the existing Cuscarias

installations. Since the proposed approach considers posts in different languages, before applying the respective sentiment classification method, it is carried out a detection of the language of the text to be processed. Language detection is performed using the detect method, defined in the language-detection Python library LangDetect.

For the sentiment classification of a Portuguese text, we used the dataset OpLexicon, a data structure with the polarities of adjectives and verbs used in the Portuguese language, with more than 32000 entries. It was created a dictionary with the keywords and respective polarities, considering the value -1 for negative polarity, 0 for neutral polarity and 1 for positive polarity.

After extracting users' posts, pre-processing techniques are applied to prepare the texts for the sentiment features extraction and selection, such as the elimination of non-relevant and non-text information and case transformation. For sentiment identification and extraction, nouns, adjectives, verbs, and adverbs are identified, for later classification. The classification of the text's sentiment, corresponding to a post, is performed by adding the polarity values of each word in the text. For this purpose, it is checked whether each word in the text has an assigned polarity, according to the dictionary used. The polarity of the text is determined by adding the polarities of the words that belong to the dictionary. The analysis carried out considers the text's syntax, where words that deny statements are considered. In the Portuguese language, the presence of a comma changes the meaning of the text, which must also be considered. The developed algorithm considers the negation "não" ("no") in particular locations, performing a verification of its existence after the sum of the text's polarity has been performed.

The algorithm also considers the position of the sentiment feature, in relation to the one that affects it, which may not be adjacent. The presence of a comma can influence the final polarity of an expression. See the example "Não é bom" ("Not good"), whose polarity is negative, where the introduction of a comma after the negation "Não, é bom" (No, it is good), changes the polarity to positive.

For English texts, was used the TextBlob text processing library, through which it is possible to determine the polarity and subjectivity of a text. The sentence construction in the English language is not as complex nor enjoys as much variability as in the Portuguese language, which means that the algorithm has a lower number of exceptions.

To enable the sentiment classification in publications, written in other languages than Portuguese or English, it was decided to translate them into English, using the Python library GoogleTrans.

To perform the sentiment analysis of user's posts in an automated way, the solution implemented consists of using a cron, defining the action to be performed, and the day and time when it should be performed. This approach eliminates possible security flaws in the access to the server, which would arise from the need to assign special permissions to the service that runs PHP (the language used in the construction of the webservice for the Cuscarias information system [31]).

Two user grouping processes were implemented. The first aims to group users who have visited a Cuscaria (when the user accesses the content of a Cuscaria) or who have just visualized it (when the user sees the Cuscaria, not accessing its content). Making an analogy with an online sales platform, visiting a Cuscaria may correspond to purchasing a product, while viewing a Cuscaria may correspond to viewing a product. The second clustering process is performed only on users who have visited a Cuscaria, and the grouping is carried out based on the classification of the polarity of the sentiments of their comments. In both users' clustering processes are considered the attributes gender, nationality, and age.

For users' clustering we used the K-Means algorithm, which is an unsupervised machine learning algorithm used in unlabeled data, that is, data without predefined categories or groups. The objective is to find groups of data (clusters) in which the numbers of groups are represented by the variable K, which refers to the number of centroids in the dataset The algorithm works iteratively to assign each data to one of the K groups based on the features provided, based on their similarity. Since the K-Means clustering is "isotropic" in all directions of space, it tends to produce round, rather than elongated, clusters. Thus, leaving variances unequal is equivalent to giving more weight to variables with less variance. To avoid this situation, it is recommended to standardize the data, otherwise the range of values in each resource will act as a weight when determining how to group the data, which is usually undesirable. When standardizing, we try to give all variables an equal weight, to achieve objectivity. In the developed algorithm, we used the StandardScaler utility class from the preprocessing package of scikit-learn library, to standardize features by removing the mean and scaling to unit variance.

A fundamental step for a clustering unsupervised algorithm is to determine the optimal number of clusters into which the data may be clustered. In the proposed approach we used the Elbow method, that consists of plotting the

explained variation as a function of the number of clusters and picking the elbow of the curve as the optimal number of clusters to use.

5. Results and discussion

Table 1 shows four comments in Portuguese and English, and the respective texts after the pre-processing process, which present only the information relevant to the sentiment analysis. Table 2 presents the results of the proposed sentiment analysis algorithm, in which the polarities of words, or groups of words, in the text are observed, as well as the text global polarity, which corresponds to the sentiment classification of the publication.

Note that, in the case of the last text, the sentiment of the publication is classified as neutral, which does not correspond to its correct classification. This is due to the incapacity of the algorithm to classify the term "don't agree" as positive, since it is unable to classify the term "disagree" as positive. The disagreement with something is not easy to classify, as it depends on the polarity of the sentiment on which the assessment is made.

Table 1. Pre-processing of Portuguese text.

Original text	Pre-processed text
Este hotel tem quartos muito bonitos e um pequeno-almoço maravilhoso mas é muito caro	Este hotel tem quartos muito bonitos e um pequeno-almoço maravilhoso mas é muito caro
Não, não penso que seja de voltar a este restaurante	Não, não penso que seja de voltar a este restaurante
Not a very nice hotel. The bedroom was very tiny but the breakfast was very good	Not a very nice hotel. The bedroom was very tiny but the breakfast was very good
No, I don't agree. Lisbon is a beautiful city	No, I don't agree. Lisbon is a beautiful city

Table 2. Sentiment classification for texts in Table 1.

Pre-processed text	Polarity Value
hotel quartos muito bonitos (1) pequeno-almoço maravilhoso (1) muito caro (-1) não penso (-1) voltar restaurante	1
Not (-1) very nice (1) hotel bedroom very tiny (-1) breakfast very good (1)	0
don't agree (-1) Lisbon beautiful (1) city	0

Fig. 2a) illustrates users' clustering results obtained for Cuscaria A which, at the experiments time, had a total of 298 users with posts, 33% male and 77% female, aged between 18 and 80 years, and from 4 different European regions. Considering Cuscaria B, Fig. 2b) presents users' segmentation according to views and visits.

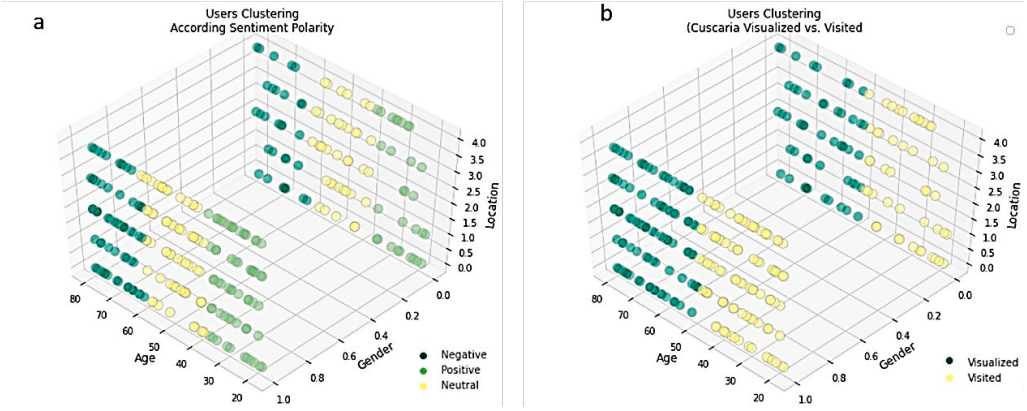

Fig. 2. (a) Users' clustering given the sentiment polarity of their posts; (b) Users' clustering according to the view or visit of a Cuscaria.

In Fig 2a) we can observe that users are segmented in three clusters where, users in the age group [18, 34] expressed a positive opinion, while users aged between 60 and 80 years old and 35 and 79, expressed a negative and neutral opinion, respectively, regardless the age and the region where they come from. With this information the entity responsible for Cuscaria A has the possibility to act as to improve its services and/or products, meeting the needs of costumer groups that demonstrate dissatisfaction.

From the analysis of Fig. 2b), it is possible to observe that, for the most part, users over 50 years of age only see Cuscaria B. With this information, Cuscaria B can change its strategy to capture the interest of tourists in these age groups, taking them to visit it.

The performance of the sentiment analysis algorithm was evaluated on a dataset of 3,460 posts in Portuguese, 1,940 in English, and 1,820 in different languages (Spanish and French), with an accuracy in the sentiment classification of 91% in the case of Portuguese posts, 89% for English posts, and 75% and 71% for posts in Spanish and French, respectively.

6. Conclusions and future work

In this paper, we described an unsupervised automatic user clustering method based on the sentiment polarities of their reviews. The architecture of the proposed solution consists of four main levels, being the first for the pre-processing of the texts to be analyzed and classified, the second for sentiment identification and extraction, using the dictionaries corresponding to the language in which the texts are written, the third for the classification of the identified sentiments and the calculation of the global polarity of the expressed sentiment, and the fourth for the users clustering, based on the sentiment polarity of their posts. The implemented solution was tested in the analysis of user posts, written in four different languages, showing a high accuracy for sentiment classification and user segmentation. It should be stressed that all analyzed data was extract from a single digital platform, Cuscarias, although we did not identify any platform specific aspect that could bias the analysis or the conclusions.

In future developments we would like to process native languages other than Portuguese and English, for we consider that the translation of the languages not directly processed may incur in loss of literal and subjectivity content, thus biasing the classification and subsequent clustering.

Another interesting development would be to perform macro clustering, that is to apply the same clustering technics used for posts and comments of a specific service or product, in aggregated information from several similar services or products, and to perform orthogonal clustering, that is to cluster post clustering information from specific services or products and identify if any correlation exist between the sentiment based identified clusters and the type of service or product under analysis.

Acknowledgements

This work has been funded by (Portuguese) Foundation for Science and Technology (FCT), under the Project UIDB/05567/2020.

References

[1] McShane, S. L., and M. A. von Glinow. (2000) Organizational Behavior. McGraw-Hill, Boston.

[2] E. Bericat, (2015). “The sociology of emotions: four decades of progress”, Current Sociology, vol. 64, pp. 491–513.

[3] T. L. Saaty, and L. G. Vargas (2012). “Models, Methods, Concepts & Applications of the Analytic Hierarchy Process”, International Series in Operations Research & Management Science. Boston: Springer US, vol. 175.

[4] J. I. Peláez, E. A. Martínez and L. G. Vargas. (2020) “Products and services valuation through unsolicited information from social media”, Soft Computing, vol. 24, pp. 1775–1788.

[5] A. Basant, M. Namita, B. Pooja, and Sonal Garg. (2015) “Sentiment Analysis Using Common-Sense and Context Information”, Hindawi Publishing Corporation Computational Intelligence and Neuroscience.

[6] C. Tawunrat, E. Jeremy. (2015). “Chapter Information Science and Applications, Simple Approaches of Sentiment Analysis via Ensemble Learning”, Volume 339 of the series Lecture Notes in Electrical Engineering, DISCIPLINES Computer Science, Engineering SUBDISCIPLINESAI, Information Systems and Applications-Computational Intelligence and Complexity.

[7] J. K. Matthew, G. Spencer, and Z. Andrea. (2015). “Potential applications of sentiment analysis in educational research and practice – Is SITE the friendliest conference?”, D. Slykhuis, G. Marks (Eds.), Proceedings of Society for Information Technology & Teacher Education International Conference 2015, Association for the Advancement of Computing in Education (AACE), Chesapeake, VA.

[8] J. I. Peláez, F. E Cabrera, and L. G. Vargas. (2018) “Estimating the importance of consumer purchasing criteria in digital ecosystems, Knowledge-Based Systems, vol. 162, pp. 252–64.

[9] V. Sebastian. (2014) “New directions in understanding the decision-making process: neuroeconomics and neuromarketing”, Procedia Social and Behavioral Sciences 127, pp. 758–62.

[10] W. Liu, and R. Ji.. (2018) “Examining the role of online reviews in chinese online group buying context: the moderating effect of promotional marketing, Social Sciences, vol. 7:141.

[11] A. Baraybar-Fernández, M. Baños-González, Ó. Barquero-Pérez, R. Goya-Esteban, and A. De-la-Morena-Gómez. (2017) “Evaluation of emotional responses to television advertising through neuromarketing, Comunicar, vol. 25, pp. 19–28.

[12] J. I. Peláez, E. A. Martínez, and L. G. Vargas. (2019) “Decision making in social media with consistent data. knowledge-based systems”, vol. 172, pp. 33–41.

[13] W. Wereda, and J. Woźniak. (2019) “Building relationships with customer 4.0 in the era of marketing 4.0: the case study of innovative enterprises in Poland”, Social Sciences, vol. 8:177.

[14] A. Baron, G. Zaltman, and J. Olson. (2017) “Barriers to advancing the science and practice of marketing”, Journal of Marketing Management, vol. 33, pp. 893–908.

[15] B. Pang, and L. Lee. (2008) “Opinion mining and sentiment analysis”, Foundations and Trends® in Information Retrieval, vol. 2, pp. 1–135.

[16] R. Prabowo, and M. Thelwall. (2009) “Sentiment analysis: a combined approach”, Journal of Informetrics, vol. 3, pp. 143–57.

[17] A. Balahur, R. Mihalcea and A. Montoyo. (2014) "Computational approaches to subjectivity and sentiment analysis: Present and envisaged methods and applications”, Computer Speech & Language, vol. 28(1), pp. 1-6.

[18] Salima Behdenna, Fatiha Barigou, and Ghalem Belalem. (2018) “Document Level Sentiment Analysis: A survey”, EAI Endorsed Transactions on Context-aware Systems and Applications 4(13):154339, DOI:10.4108/eai.14-3-2018.154339.

[19] Vrushali K. Bongirwar. (2015) “A Survey on Sentence Level Sentiment Analysis”, International Journal of Computer Science Trends and Technology (IJCST) – Volume 3 Issue 3, pp. 110–113.

[20] Tomoki Ito, Kota Tsubouchi, Hiroki Sakaji, Tatsuo Yamashita, and Kiyoshi Izumi. (2020) “Word-Level Contextual Sentiment Analysis with Interpretability”, The Thirty-Fourth AAAI Conference on Artificial Intelligence.

[21] Kim Schouten, and Flavius Frasincar. (2015) “Survey on Aspect-Level Sentiment Analysis”, IEEE Transactions on Knowledge and Data Engineering 28(3):1-1, DOI:10.1109/TKDE.2015.2485209.

[22] P. Turney. (2002) “Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews”, Proceedings of 40th Meeting of the Association for Computational Linguistics, pp. 417–424.

[23] A. D’Andrea, F. Ferri, P. Grifoni and T. Guzzo. (2015) “Approaches, tools and applications for sentiment analysis implementation”, International Journal of Computer Applications, vol. 125, No.3.

[24] J. Qian and C. Gao. (2011) “The application of data mining in CRM”, in 2nd International Conference on Artificial Intelligence, Management Science and Electronic Commerce (AIMSEC), Deng Leng, pp. 5202-5206.

[25] A. Ansari and A. Riasi. (2016) “Customer clustering using a combination of fuzzy c-means and genetic algorithms”, International Journal of Business and Management, vol. 11, no. 7, pp. 59-66.

[26] Q. Zhao and P. Franti. (2014) “Centroid Ratio for a Pairwise Random Swap Clustering Algorithm”, IEEE Transactions on Knowledge and Data Engineering, vol. 26, no. 5, pp. 1090-1101.

[27] T. Kanungo, et al.. (2002) “An efficient k-means clustering algorithm: analysis and implementation”,IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 24, no. 7, pp. 881-892.

[28] D. MacKay. (1967) “An example inference task: Clustering”, in Information theory, inference, and learning algorithms. Cambridge, UK:

[29] J. MacQueen, “Some methods for classification and analysis of multivariate observations”, in Fifth Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, pp. 281-297.

[30] O. Maimon and L. Rokach. (2005) “Clustering methods”, in Data Mining and Knowledge Discovery Handbook. Boston: Springer US, pp. 321-352.

[31] S. Jardim, N. Madeira, and N. Cardoso. (2018) “Cuscarias: a cultural social information system based on co-creation”, Proceedings of the 13th Iberian Conference on Information Systems and Technologies, pp. 1-5.