A Multilingual Lexicon-based Approach for Sentiment Analysis in Social and Cultural Information System Data

Sandra V.B. Jardim, Carlos Mora
Smart Cities Research Center
Polytechnic Institute of Tomar
Tomar, Portugal

Tiago Santana,
Undergraduate Student
Polytechnic Institute of Tomar
Tomar, Portugal

Abstract — **Sentiment analysis, also known as opinion mining, has become ubiquitous in our society, with applications in online searching, computer vision, image understanding, artificial intelligence and marketing communications. In this paper, is described an unsupervised automatic method of multilingual lexicon-based sentiment analysis algorithm, with a dictionary-based approach. The developed algorithm was tested in the sentiment analysis of users' 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.**

Keywords – sentiment analysis; automatic sentiment classification; lexicon-based approach; social media.

I. INTRODUCTION

In decision-making, the human being tends to give great importance to what those who relates to him think and feel, both in the personal and professional fields [1, 2]. Other opinions not only help people make informed decisions, but also increases organizations knowledge, helping them in the decision-making process [3]. Knowing customer opinions, attitudes and emotions regarding to products and services offered, helps organizations to understand the degree of their customers satisfaction, which is of great importance for the decision-making process, allowing them to anticipate or change their commercial strategies, as well as adapt to the evolution of the market and their needs. This fact, combined with the increasing use of digital technologies in the relationship between customers and companies, has stimulated a great amount of research aimed the automated classification of the sentiments subjectively expressed in customers review and opinion texts. For the organizations, sentiment analysis is of great relevance in different areas, such us: analysis of consumer buying patterns [4, 5]; collecting customer feedback on social media, websites or online forms [6]; obtaining knowledge about the stimuli that create the greatest impact on people [7]; understanding the factors that motivate people to like a product or service [8]; conducting research market [9]; categorizing customer service requests; predicting consumer behavior, among others [10].

A. Sentiment Analysis

Sentiment analysis has become ubiquitous in our society, with applications in online searching, computer vision, image understanding, artificial intelligence and marketing communications [11]. Also known as opinion mining, sentiment analysis, is a textual and visual information classification automated process of detecting, extracting and classifying opinions, according to the data polarity (positive, negative and neutral) [12, 13] but also on sentiments and emotions (angry, happy, sad, etc.), urgency (urgent, not urgent) and even intentions (interested v. not interested).

It is used to determine the emotional value expressed in a set of words or in a text, obtaining an understanding of opinions and emotions for subjective texts, mainly related to consumer's reviews on products and services. Sentiments are classified into positive, and negative sentiments, and neutral, in situations with no sentiments involved. Sentimental analysis is a research field in machine learning (ML), natural language processing (NLP) and computational linguistics, which involves three major levels, which determines the tasks required for the process – word level, sentence level, and document level, where the word level is the most complex one given the difficulty in carrying out the analysis, whereas the analysis is simpler at the sentence and document levels [14].

The most popular types of sentiment analysis are:

- Fine-grained sentiment analysis: when the polarity precision is important, it is considered the expansion of the polarity categories to include, for instance, very positive, positive, neutral, negative, and very negative.
- Emotion detection: aims to detect emotions, like happiness, frustration, anger, and sadness.
- Aspect-based sentiment analysis: when it is important to know the aspects or characteristics on which an assessment is made or a positive, neutral or negative opinion is issued.
- Multilingual sentiment analysis: usually more difficult to implement, since it involves a lot of preprocessing and resources, some available online (e.g. sentiment lexicons), while others need to be created (e.g. translated corpora or noise detection algorithms),

A sentiment analysis process can be structured in five main procedures:

- Data extraction: Sometimes the data record accessed is very large, disorganized and disintegrated. The

opinions and sentiments expressed in the texts are expressed in different ways, varying in detail, in the type of vocabulary and language, and may also present slang terms. Manual analysis becomes not only a time-consuming but also a tedious task and can also lead to errors inherent to human intervention. This procedure aims to extract data from one or more sources, creating a dataset that can be computationally analyzed.

- Pre-processing: Several data cleaning techniques are applied to prepare the dataset for efficient computational analysis, such as: identification and elimination of the non-text content and irrelevant information.
- Sentiment detection: Carried out at different levels (word, sentence level, and document) with commonly used techniques, such as: Unigrams, N-Grams, Lemmas, Negation and Opinion words [15].
- Sentiment classification: Through the segmentation of subjective information into classification groups, the respective polarity is classified. Classification groups are generally represented at two extreme points on a continuum and may also involve several categories (refined sentiment analysis).
- Polarity report: The results of a sentiment analysis approach can be displayed in several conventional ways, being the most common the graphical representation, using colors, frequencies, percentages, and size to segment polarity.

B. Sentiment Analysis Applications

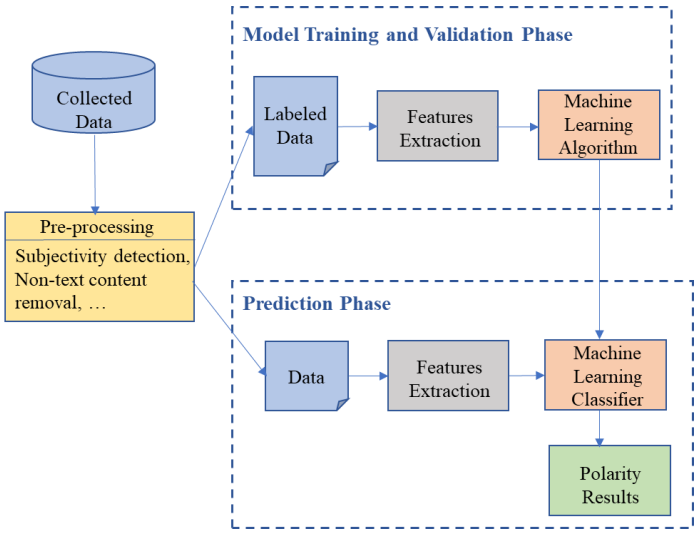
The interest in sentiment analysis research has been increasing tremendously in recent years, due to a wide range of business and social applications in human-centric environments. Text categorization according to affective relevance and opinion exploration for business and social behavior analysis are some of the application areas of sentiment analysis [16, 17]. In the business area, brand monitoring and reputation management, market research, product analytics and customer voice analysis, are some of applications of sentiment analysis.

Sentiment analysis systems also have an important potential role as enabling technologies for other systems [18], such as recommendation systems [19, 20], since it might endow him with the ability to not recommend products or services that receive a lot of negative opinions.

C. Machine Learning Sentiment Analysis

In a sentiment analysis process, the main applied techniques are machine learning and lexicon-based approaches, where the classification task is made using, respectively, supervised machine learning or semantic-based unsupervised methods.

Machine learning is a subset of artificial intelligence, which uses computerized techniques to solve problems based on historical data and information without having to modify the core process [21]. Machine Learning approach can be categorized into supervised and unsupervised learning [22].



While supervised learning can be defined as the process of learning from already known/labeled data to generate initially a model and further predict target class for the particular data [22, 23], unsupervised learning can be defined as the process of learning from unlabeled to discriminate the provided input data [22].

Fig. 1 illustrates a sentiment analysis solution based on a supervised machine learning approach.

Figure 1. Supervised machine learning sentiment analysis

For the extraction of features, the text is converted into the feature vector with the help of the data-driven approach. Each sentence is examined for its subjectivity, maintaining only those that present subjective expressions, while those that convey facts and objective communication being discarded. One way to avoid the high-dimensional input spaces is to assume that only a few features are relevant and necessary for the task. Feature selection tries to exclude all irrelevant features. However, in text categorization this can easily lead to a loss of information since there are often many relevant features [24].

There are several techniques and machine learning algorithms used in training models to carry out sentiment analysis, among which are:

1) Naïve Bayes Classifiers

A family of "probabilistic classifiers" based on the application of Bayes' theorem with strong assumptions of independence between features. Despite belonging to the group of the simplest Bayesian network models [22], when together with the estimate of the kernel density, they can reach higher levels of accuracy [22, 23]. Naive Bayes sentiment classification technique uses the Naive Bayes theorem, defined by

$$P(C|X) = \frac{P(X|C)\,.\,P(C)}{P(X)}\,, \qquad (1)$$

where X denotes de document to analyze and C the class (positive or negative).

In a Naïve Bayes classifier is assumed that the fact that the probability of a word in the document is in a certain category is not related to the probability of the other words being in the same category. In this technique, the classification of a document is carried out in four stages. While in the first step, the data set is converted into a frequency table, in the second the PRIOR is calculated, using the formula $P(C) = N_c/N$, where $P(C)$ is the class probability, N_c is the total count of a particular class in the training dataset and N is the total class count in the training dataset. In the third step, the conditional probability, i.e. the likelihood, of each attribute of the word is calculated. In the last step, is calculated the posteriori probability, using the formula:

$$C_{map} = \text{argmax}\, P(X1, X2, X3, \dots, Xn)\,.\,P(C). \qquad (2)$$

2) Logistic Regression

Logistic Regression, similar to linear regression, can be applied to classification problems assuming that the objective variable is a discrete value. The Logistic Regression model uses a sigmoid flattening function that describes a model prediction as the probability σ that a given entry x belongs to one of the classes of y. The logistical function is given by:

$$\sigma = \frac{1}{1 + e^{-\lambda x}}\ , \qquad (3)$$

where λ is a weighting constant.

The general principle of a logistic classifier is the minimization of an error function E, over a maximum number of iterations or until the function's convergence is reached. The error is given by:

$$E(w, b) = \frac{1}{N}\sum_{i=1}^{N} L\left(y_{(i)}, \sigma(x_i)\right) + \alpha . R(w) \qquad (4)$$

where N is the total number of records, L is the classifier loss function, $\sigma(x)$ is the probability function given by $\sigma(x) = w^T . x + b$ (where w is the vector of parameters and b the linear coefficient), α is a non-negative constant and R is the regularization term defined by the penalty function.

3) Support Vector Machine Classifier

Considered the most accurate text classifier, the Support Vector Machines classifier finds a hyperplane which separates the data into two categories, i.e. positive or negative, with the maximized margin [24]. The name of this technique comes from the fact that it is used a support vector, which is an array of data points, used to find out the boundary of each plane. The classifier classifies a new input, predicting in which side of the margin it belongs.

D. Semantic-based Sentiment Analysis

The semantic-based sentiment analysis approach can be divided into two types of techniques, designated corpus-based, and dictionary/lexicon/knowledge-based.

1) Corpus-based Semantic Analysis

The corpus-based semantic orientation approach requires a large set of data to detect the polarity of terms and, consequently, the overall sentiment translated in the text. The main problem with this approach is its dependence on the polarity of terms in the training corpus [25]. Despite this limitation, its simplicity has enhanced its use in several proposals for the analysis of feelings [26, 27, 28].

In a first phase, the corpus-based semantic orientation approach extracts the feeling terms from the unstructured text, calculating the respective polarities. Most terms that convey sentiment are multi-word features, instead a bag-of-words, which limits their use.

2) Lexicon-based Semantic Analysis

The lexicon-based semantic orientation approach, also known as dictionary or knowledge-based, aims calculating orientation for a document from the semantic orientation of words or phrases in the document [29]. Being an unsupervised approach, it does not require prior training to explore the data. It uses a predefined list of words, where each word is associated with a specific sentiment, based on positive and negative word count [30]. Lexicon-based strategies are very simple and efficient methods, either in the use of computational resources or in the ability to predict. 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 dictionary, which can be created manually [31] or automatically, using seed words to expand the list of words [29, 32]. One of the main challenges of this approach is the construction of a comprehensive lexical dictionary, which is not a simple task since it is necessary to consider the application contexts, which makes difficult to use a unique and board lexicon. On the other hand, and considering the analysis of publications or comments on Social Networks as an objective, the large volume of textual content produced on the Web daily, combined with the variability of the language used, ranging from extremely formal to too informal, often containing slang, represents an added difficulty.

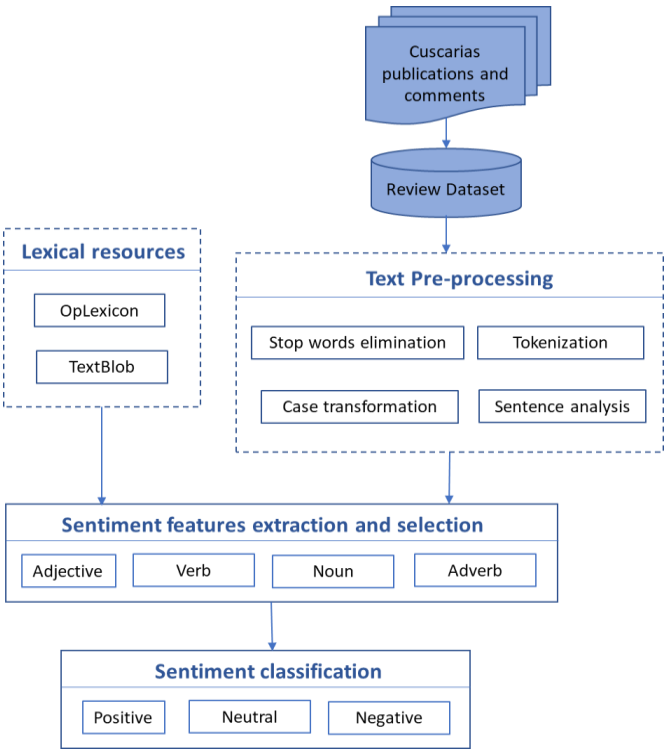
In a lexicon-based approach, the first step is to create a dictionary by compiling a list of adjectives and verbs and their respective values of semantic orientation (positive, neutral and negative). Subsequently, from the document from which it is intended to classify the associated sentiment, and according to the dictionary built in the previous step, all adjectives and verbs in it are extracted and classified. The overall sentiment of the text is obtained by aggregating the classifications obtained for each adjective and verb.

It is important to note that, in the case of lexical approaches, a pre-processing step on the textual content treated is essential. As mentioned before, the extraction of relevant information about the text is essential 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 in a sentiment analysis approach is text segmentation (tokenization), whose goal is to transform the text into a set of terms extracted from the texts (tokens), both at the 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.

II. METHODOLOGY

In this paper, is described an unsupervised automatic method of multilingual lexicon-based sentiment analysis algorithm, where is used a dictionary-based approach. 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 to be possible to score the global sentiment of each text in the data set.



The developed algorithm aims to classify the sentiments or opinions of digital platforms users, by analyzing their publications and comments. To test the developed algorithm, 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 [29]. 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. 2.

Figure 2. Architecture of the proposed solution

III. 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).

A. *Data Set Description*

As previously mentioned, in this work we are focus in the data stored in the Cuscarias cultural and social information system, so the data set it's built from the comments recorded for each of the existing Cuscarias.

For the connection to the database of the information system and collection of the publications of its users, was imported the module *mysql.connector*, enabling the invocation of the *connector* method to configure the connection to the MySQL database, as well as the use of the *cursor* object, which stores the return of a query to the database.

B. *Multilingual Lexicon-Based Sentiment Analysis*

The proposed approach considers publications (or comments) in different languages. Thus, a detection of the language of the text to be processed is carried out by applying the respective sentiment classification method. Language detection is performed using the *detect* method, defined in the language-detection Python library *langdetect*.

Sentiment Analysis for Portuguese texts

For the sentiment classification of a text written in Portuguese, 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.

(1) Text pre-Processing

After extracting publications from users of the Cuscarias platform, pre-processing techniques are applied to prepare de texts for the sentiment features extraction and selection, such as the elimination of non-relevant and non-text information and the case transformation.

(2) Sentiment features extraction and selection

For sentiment identification and extraction, nouns, adjectives, verbs and adverbs are identified, for later classification.

(3) Sentiment classification

The classification of the sentiment of the text, corresponding to a publication, 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 created. 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 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 created algorithm considers the negation “não” (“no”), performing a verification of its existence after the sum of the text's polarity has been performed. This verification aims to

change the value of the polarity of the adjective, verb, noun or adverb to which the negation refers, and if the polarity of the sentiment feature is positive, it changes its polarity to negative, if it is negative, it changes to neutral polarity.

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's good), changes the polarity to positive.

Sentiment Analysis for English texts

In calculating the polarity of texts in English, the TextBlob text processing library was used, 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.

Sentiment Analysis for texts in another languages

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*.

C. *Multilingual Lexicon-Based Sentiment Analysis Automatization*

To perform the sentiment analysis of user's publications and comments 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 has as main objective to eliminate the 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 [33]). For this, the Python *schedule* module was imported, which allows running periodicity functions for one or more processes. In the case of the proposed solution, a function is invoked every minute that, through a query to the information system database, collects comments and publications without associated sentiments. With the results obtained, each publication is analyzed, starting with the identification of the language in which it was written. Once the language is known, the function responsible for the sentiment analysis of the text is invoked, returning its polarity. Once the sentiment polarity is known, the sentiment field in the database is updated.

IV. RESULTS AND DISCUSSION

This section presents and discusses the results obtained by the developed algorithm. Examples of the results obtained in the sentiment analysis of publications in Portuguese and English languages are presented.

Table 1 shows three comments in Portuguese, 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 the of publication.

Tables 3 and 4 presents three comments in English, the pre-processing text results, and the sentiment classification the of each 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 the algorithm does not have the capacity to classify the term "don't agree" as positive. This is because the algorithm 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 I. Pre-processing of Portuguese text.

Original text	Pre-processed text
Este restaurante tem uma vista muito bonita e uma comida maravilhosa mas é muito caro *This restaurant has a very beautiful view and a wonderful food but is very expensive*	Este restaurante tem uma vista muito bonita e uma comida maravilhosa mas é muito caro *This restaurant has a very beautiful view and a wonderful food but is very expensive*
Não, não penso que seja de voltar a este lugar *No, I do not think we should return to this place*	Não, não penso que seja de voltar a este lugar *No, I do not think we should return to this place*
Não, este monumento é muito bonito *No, this monument is very pretty*	Não, este monumento é muito bonito *No, this monument is very pretty*

TABLE II. Sentiment classification for texts in TABLE I.

Pre-processed text	Polarity Value
restaurante vista muito bonita (1) comida maravilhosa (1) muito caro (-1) *restaurant very beautiful (1) view wonderful (1) food very expensive (-1)*	1
não penso (-1) voltar lugar *do not think (-1) return place*	-1
monumento muito bonito (1) *monument very pretty (1)*	1

TABLE III. Pre-processing of English text.

Original text	Pre-processed text
This is a great place, with a wonderful view	This is a great place, with a wonderful view
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. Paris is a beautiful city	No, I don't agree. Paris is a beautiful city

TABLE IV. Sentiment classification for texts in TABLE II.

Pre-processed text	Polarity Value
great (1) place wonderful (1) view	2
Not nice (-1) hotel bedroom tiny (-1) breakfast good (1)	-1
don't agree (-1) Paris beautiful (1) city	0

The performance of the algorithm was evaluated on a dataset of 2,100 publications in Portuguese, 1,450 in English, and 820 in different languages (Spanish and French), with an accuracy in the sentiment classification of 87% in the case of publications in Portuguese, 92 % for publications in English, and 72% and 79% for publications in Spanish and French, respectively.

V. CONCLUSIONS

In this paper, was described an unsupervised automatic method of multilingual lexicon-based sentiment analysis algorithm, with a dictionary-based approach. The architecture of the proposed solution consists of three 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, and the third for the classification of the identified sentiments and the calculation of the global polarity of the expressed sentiment. The implemented solution was tested in the analysis of user's publications in a digital tourism platform, written in four different languages, showing a high accuracy for sentiment classification.

The proposed method, however, is heavily dependent on the syntactic structure and semantic context of the text analyzed and may thus be inaccurate in its conclusions, especially due to possible plural significates in the lexicon used, so the authors propose that further work should be done with an approach based on deep-learning methods pre-trained with datasets that should be context specific for the location of the Cuscarias text collection points.

ACKNOWLEDGMENT

This work has been funded by national funds through FCT - Fundação para a Ciência e a Tecnologia, I.P., under the Project UIDB/05567/2020.

REFERENCES

[1] E. Bericat, "The sociology of emotions: four decades of progress. Current Sociology, vol. 64, 2015, pp. 491–513.

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

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

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

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

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

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

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

[9] W. Wereda, and J. Woźniak, "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, 2019.

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

[11] P. Sánchez-Núñez, C. Heras-Pedrosa, and J. I. Peláez, " Opinion mining and sentiment analysis in marketing communications: a Science mapping analysis in Web of Science (1998–2018)", Social Sciences, vol. 9(3), 2020.

[12] B. Pang, and L. Lee, "Opinion mining and sentiment analysis", Foundations and Trends® in Information Retrieval, vol. 2, 2008, pp. 1–135.

[13] R. Prabowo, and M. Thelwall, "Sentiment analysis: a combined approach", Journal of Informetrics, vol. 3, 2009, pp. 143–57.

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

[15] Y. Mejova, and P. Srinivasan, "Exploring feature definition and selection for sentiment classifiers", Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media, 2011.

[16] M. M. Altawaier, and S. Tiun, "Comparison of machine learning approaches on arabic Twitter sentiment analysis". International Journal on Advanced Science, Engineering and Information Technology, Vol. 6, No. 6, 2016, pp. 1067-1073.

[17] K. Dave, S. Lawrence and D. M. Pennock. "Mining the peanut gallery: Opinion extraction and semantic classification of product reviews". In Proceedings of the 12th international conference on World Wide Web, pp. 519-528.

[18] B. Pang and L. Lee, "Opinion mining and sentiment analysis", Foundations and Trends in Information Retrieval, Vol. 2, 2008, pp. 1–135.

[19] J. Tatemura, "Virtual reviewers for collaborative exploration of movie reviews", Proceedings of Intelligent User Interfaces, 2000, pp. 272–275.

[20] L. Terveen, W. Hill, B. Amento, D. McDonald, and J. Creter, "PHOAKS: a system for sharing recommendations," Communications of the Association for Computing Machinery, vol. 40, 1997, pp. 59–62.

[21] S. B. Kotsiantis, "Supervised machine learning: a review of classification techniques", Informatica, vol. 31, 2007, pp. 249–268.

[22] T. Hastie, R. Tibshirani and J. Friedman, The Elements of Statistical Learning-Data Mining, Inference and Prediction, 2nd Edition, vol. II, Stanford: Springer, 2008, pp. 465-576.

[23] S. Piryonesi, and E. El-Diraby, "Role of data analytics in infrastructure asset management: overcoming data size and quality problems". Journal of Transportation Engineering, vol. 146 (2), 2020.

[24] T. Joachims, Learning to Classify Text Using Support Vector Machines, Kluwer Academic Publishers, 2001.

[25] B. Agarwal, and N. Mittal, "Semantic orientation-based approach for sentiment analysis. B. Agarwal, N. Mittal (eds.) Prominent Feature Extraction for Sentiment Analysis. SC, Springer, 2016, pp. 77–88.

[26] K. Stuart, A. Botella, and I. Ferri, "A corpus-driven approach to sentiment analysis of patient narratives", EPiC Series in Language and Linguistics, vol. 1, 2016, pp. 381–395.

[27] A. Moreno-Ortiza and J. Fernández-Cruzb, "Identifying polarity in financial texts for sentiment analysis: a corpus-based approach", Procedia - Social and Behavioral Sciences, vol. 198, 2015, pp. 330 – 338.

[28] R. Douglas, and C. Zorn, "Corpus-based dictionaries for sentiment analysis of specialized vocabularies", Political Science Research and Methods, vol. 9, 2021, pp. 20–35.

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

[30] A. D'Andrea, F. Ferri, P. Grifoni and T. Guzzo, "Approaches, tools and applications for sentiment analysis implementation", International Journal of Computer Applications, vol. 125, No.3, 2015.

[31] M. R. Tong, "An operational system for detecting and tracking opinions in on-line discussions", Working Notes of the ACM SIGIR 2001 Workshop on Operational Text Classification, 2001, pp. 1–6.

[32] P. Turney, and M. Littman "Measuring praise and criticism: Inference of semantic orientation from association", ACM Transactions on Information Systems, vol. 21(4), 2003, pp. 315–346.

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