Techkatl: A Sentiment Analysis Model to Identify the Polarity of Mexican's Tourism Opinions

Eduardo Roldán Reyes[1][0000-0002-4212-1586]

[1] TecNM / Instituto Tecnológico de Orizaba, Orizaba, CP 94300, Mexico
eroldanr@orizaba.tecnm.mx

Abstract. This article describes the model used for the Sentiment Analysis Task framed within the REST-MEX 2021: Recommendation System for Text Mexican Tourism. The sentiment analysis model, called Techkatl, implemented a generic five-step text mining process for the identification of the polarity of opinions issued by tourism visitors in Mexico. For the polarity detection, Techkatl utilized a supervised learning approach with cross-validation to train and test classification algorithms. For the development, the data analytic RapidMiner platform was used for the rapid prototyping and the performance evaluation of the classification task. The deployment of the model showed a performance above the baseline for fast identification of the polarity with a low computation cost.

Keywords: Sentiment Analysis, Machine Learning, Supervised Algorithms, Polarity Detection, RapidMiner

1 Introduction

Sentiment analysis (SA) is a novel approach to determine the sentiment, emotion, or polarity implicitly or explicitly expressed in an opinion [1]. It is mainly applied on the Internet's social media and e-commerce websites to analyze the comments of users and customers' reviews. Under this approach, the polarity of an opinion is the degree of positiveness, negativity, or neutrality towards a certain topic.

Nowadays, the SA has been successfully applied to understand customers' opinions and to propose marketing strategies to enhance the quality of the products and services of the companies as can be observed in [2–4]. Although several studies have been carried out in the tourism context [5–9], they are almost focused on the English language and very few have been addressed for the Spanish language, specifically on tourism in Mexico. This has motivated the REST-MEX 2021 Recommendation System for Text Mexican Tourism [10]. In this edition, a contest on SA was proposed to challenge researchers and SA practitioners to participate with systems predicting the polarity of a database of opinions issued by tourists who have already traveled to attraction spots of Guanajuato in Mexico.

In this paper a description of the Techkatl team from the TecNM-Instituto Tecnológico de Orizaba – MIA is presented. The Techkatl model, which name comes from

IberLEF 2021, September 2021, Málaga, Spain.

the Nahuatl language meaning “Sentiment”, was developed under the machine learning approach. It performs a five-step generic model, inspired by a text mining model previously developed by our team [11].

The rest of the paper is structured as follows. Section 2 explains the rules for the SA classification task. Section 3 describes the characteristics of the proposed system. Section 4 highlights the experimental evaluation and the attaint results. Finally, the general conclusions are mentioned in Section 5.

2 Task description

The SA task consists of the classification of textual opinions, expressed by tourists visiting interesting spots of Guanajuato in Mexico, to identify the polarity. All the opinions were obtained from the TripAdvisor platform and provided by the contest organizers to participating teams in a .csv file for training. The opinions were registered on the platform between 2002 and 2020. The polarity of each opinion ranged between 1 and 5, where 1 stands for the most negative polarity and 5 the most positive. An excerpt of the released database is shown in Table 1.

Table 1. Example of the training database.

index	Title	Opinion	Place	Gender	Age	Country	Date	Label
1	¡Momias…	Las mom...	Museo…	Male	53	México	22/10/2016	1
…	…	…	…	…	…	…	…	…
5197	Muy bo…	No te…	Monum...	Female	31	México	26/03/2016	5

The entire collection of the comments consist of 7,632 opinions where 5,784 are from Mexican tourists and 1,848 come from other Iberoamerican country’s tourists. For the SA track, the database was split into two partitions: 5,197 were provided as the training set (labeled) and 2,435 were later released as the test set (unlabeled).

To evaluate the participant methods and to determine the winner of the challenge, the Mean Absolute Error (MAE) metric was used (Eq. 1). Thus, the system with the lowest MAE value was considered the winner.

$$MAE = \frac{1}{n}\sum_{t=1}^{n}|e_t| \quad (1)$$

3 Model description

In this part, the SA model developed to deal with the challenge is described. The model has been built under the RapidMiner Studio 9.9 version. The experiments and the evaluation were also performed on this platform. Among several advantages for using the RapidMiner [12], the main reason that motivated us to use RapidMiner is that it allows the rapid development of data analysis processes by chaining operators in a user-friendly graphical environment. The model is composed of five main steps

to perform the SA process: 1) Acquisition, 2) Pre-processing, 3) Processing, 4) Evaluation, and 5) Results. The first and second stages are both applied for Training (A) and Testing (B).

In the next subsections, each step is described in detail. A complete view of the five-step model is shown in Fig. 1.

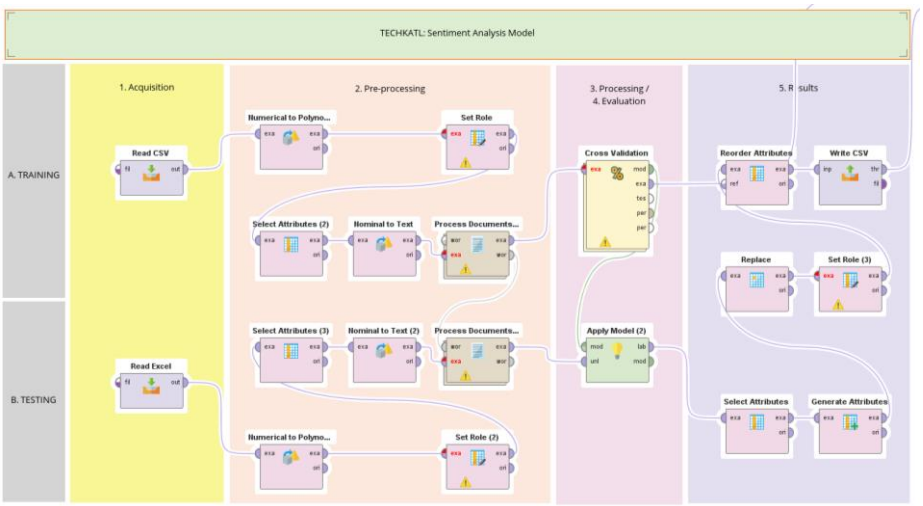

Fig. 1. The SA model on the RapidMiner platform.

3.1 Acquisition

The information acquisition is performed through two operators that read the .csv file for the training set, and the .xls file for the testing one. The parameters of this operator were configured to recognize the encoding of the text file (UTF-8) since the Spanish language has accentuated characters, and to identify the character for column separator (,). The output of these operators is the Example Set, which is a database internally created and displayed as a table in the results view panel of the program interface.

3.2 Pre-processing

The pre-processing step involves two different groups of operators. The first one is composed of four operators which purpose is formatting the data in order to be recognizable for the classification algorithms. These operators are: “Numerical to Polynomial” for changing the type of attributes to a polynomial type; “Set Role” to indicate the index, the regular attributes, and the class (label); “Select Attributes” for dismissing attributes according to its importance or irrelevance (e.g. the index attribute) hence, only the Title and Opinion attributes were kept for further analysis; and “Nominal to Text” to set up the text attributes into string attributes.

The other group of operators is enclosed in the “Process Documents from Data” operator which generates word vectors from the string attributes. The objective of this

group of operators is to reduce the information volume and to increase the efficiency of the classification algorithm. Within this operator the following operators are concatenated to perform the next five sub-stages of text pre-processing:

- Tokenize: this operator fragments the text into syntactic units (i.e. words).
- Transform Cases: usually, the opinions are a mix of uppercase and lowercase words which may be difficult to further processing. With this operator, all the uppercase letters are converter to their lowercase forms.
- Replace Tokens: this operator is used to replace: a) misspelled words, and b) accentuated characters with non accentuated characters. This helps to reduce the volume of the text by identifying duplicate or misspelled words.
- Filter Stopwords (Dictionary): this operator removes the most trivial words such as pronouns, prepositions, and articles by comparing each token to a stop-word list. Since Rapidminer does not have a Spanish stopword list, a custom list with 722 stopwords was created and loaded through this operator. The list of stopwords can be requested to the author on demand. This sub-process helped to reduce by 30% the text volume.
- Stem (Snowball): this operator applies several stemming algorithms for the Snowball language [13]. This operator supports the Spanish language.

Fig. 2 shows the number of removed tokens after the pre-processing task with the training set. The amount of reduced information is up to 60%. A similar result was obtained with the Test set.

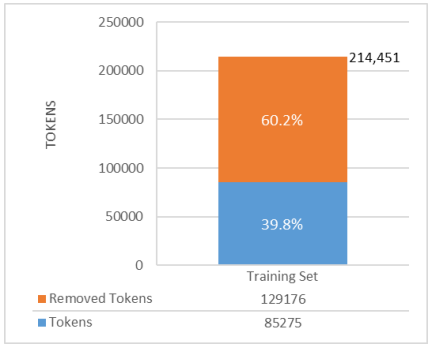

Fig. 2. The valid tokens after pre-processing tasks.

The remaining tokens are used to create a word vector through the Term Frequency - Inverse Document Frequency (TF-IDF) method [14]. The TF-IDF (Eq. 2) computes the relative frequency of a word (t) in a specific document (d) through an inverse proportion of the word over the entire collection of documents (D). The IF-TFD was selected because it provides a simple, reliable, and fast schema to evaluate the relevance of each token within a large collection of opinions.

$$TF - IFD(t, d, D) = \log\big(1 + freq(t, d)\big) \cdot \log\left(\frac{N}{count(d \epsilon D : t \epsilon d)}\right) \quad (2)$$

3.3 Processing

Within this step, a classification task is performed through the application of different supervised machine learning algorithms. This step aims to classify the opinions into five different classes (1 to 5), representing the different degrees of polarity. The algorithms applied were the following:

- k-Nearest Neighbors (k-NN): this algorithm classifies a new opinion based on the majority class of its k neighbor opinions. A similarity metric (the mixed Euclidean distance) is used to measure the distances between the unclassified opinion and its neighbors. For the SA task, a distance of 0 is taken if both opinions are closest, otherwise, the distance is equal to 1. For the experimentation, different k values were selected (k = 1, 3, and 5).
- Trees: with this operator, a decision tree (DT) model is generated. Each leaf of the model represents the class and the nodes represent a splitting rule for one specific attribute. The criterion used to construct and prune the trees was the information gain. Another two algorithm variants were also tested, such as the Gradient Boosted Trees (GBT) and Random Forest (RF).
- Support Vector Machine (SVM): this learning method applies the mySVM algorithm [15] and supports various kernel types. For the experimentation, the Linear kernel type was chosen since the number of attributes is large and the relation between the class labels is linear.
- Bayesian Methods: two variants of these methods were applied: the simple Naïve Bayes (NBS) and the kernel one (NBK). For the second one, a greedy kernel was set with a minimum bandwidth of 0.1 and 10 kernels.
- Artificial Neural Networks (ANN): finally, two of the most representative neural network algorithms were applied. The Neural Net (NN) algorithm built a model using a feed-forward neural network trained by a backpropagation algorithm (i.e. a multi-layer perceptron) and the Deep Learning (DP) algorithm which performed a multi-layer feed-forward artificial neural network trained with stochastic gradient descent using back-propagation [16].

3.4 Evaluation

For the evaluation step, the Cross-Validation operator was applied to estimate the performance of the classification algorithms. This procedure encloses two subprocesses: training and evaluation. First, the input Example Set is split into n=10 subsets of equal dimensions (i.e. number of folds), and one of the subsets is kept as the test dataset. The rest of the subsets are used as the training dataset and processed by the classification algorithm. The procedure is repeated $n-1$ times, with all of the subsets. The performance metrics and results from the n iterations are finally averaged to output a single estimation.

The performance evaluation of the classification model for each test set produces an acceptable estimation of the model performance on unlabeled datasets. Nevertheless, it does not guarantees the same performance on new unlabeled data.

The experiments for the classifier's evaluation were performed on a Dell XPS-9370 PC with a Core i7 Intel microprocessor.

3.5 Results

The last step of the SA model displays two outputs: the performance evaluation of the algorithms and the classifications of the opinions of the Test Set (i.e. polarity). Several operators were applied to meet the requirements of the output submission (Fig. 1). Table 2 summarizes the performance of the classification algorithms for the Training Set.

Table 2. The performance metrics of the evaluated classification algorithms.

Classification Algorithm	Accuracy	MAE	Processing time (CPU-time) in min.
k-NN (*k* = 1, 3, 5)	64.8%, 56.73%, 52.91%	0.324, 0.653, 0.682	43.53, 41.16, 40.69
DT, GBT, RF	51.82%, 57.84%, 53.22%	0.619, 0.542, 0.601	1.88, 121.3,
SVM	51.76%	0.482	1.71
NVS, NBK	**80.3%, 81.74%**	**0.197, 0.234**	**1.4, 3.67**
NN, DP	69.84%, 72.58%	0.387, 0.351	408.32, 1567.85

Along with the MAE and the accuracy metric, the CPU processing time for each algorithm was also measured through the “Log” operator. As a result, the algorithms that were chosen for submission of the files were the Naive Bayes ones.

3.6 Discussion

As can be seen in Table 2, the NVS and the NBK were able to produce the lower values of the MAE (which is the metric that was chosen to rank the team’s results in the contest) and the best accuracy rates. Even if other algorithms may perform higher rates of accuracy, the NVS and NBK algorithms also showed low processing times for the classification task. This can be an important issue since the rapidity of identifying the polarity of opinions could be crucial to producing short-term and low-cost improvement strategies. These were the main reasons why these two methods were selected over the other algorithms tested.

4 Conclusions

In this article, the Techkatl model for the SA track of the REST-MEX 2021 was described. The model development and experimentations were carried out on the RapidMiner platform. It was chosen for its relative ease of use and because it offers a very user-friendly interface to develop machine learning models. Also, it is supported by a large community of practitioners, researchers, and data scientists. It is recom-

mended for rapid prototyping, and it can also be used by decision-makers in crucial areas of industry, management, tourism, or marketing to perform machine learning, text mining, or sentiment analysis tasks. As an example of the practicality provided by this tool, it can be highlighted that the model hereby presented was developed in a very short time (less than a couple of hours).

On the other hand, the Techkatl proposed model showed that even if several algorithms have been developed for SA, many of them are complexes, time-consumers and even performs low rates of efficiency in comparison with other simplest algorithms such as the Naïve Bayes methods which still performing well and fast at a low computing cost. Regarding other complexes and more recent methods, such as the Deep Learning approach, they present the main disadvantages that the training time may be considerable. This is a problem especially when it is required to obtain correct classifications in the short term and with limited computing power.

Acknowledgments

The author of this paper would like to express his gratitude to Conacyt, the Tecnológico Nacional de México, and to recognize the support of colleagues and students of the Instituto Tecnológico de Orizaba.

References

1. Ravi, K., Ravi, V.: A survey on opinion mining and sentiment analysis: Tasks, approaches and applications. Knowledge-Based Systems. 89, 14–46 (2015). https://doi.org/10.1016/j.knosys.2015.06.015.
2. Balazs, J.A., Velásquez, J.D.: Opinion Mining and Information Fusion: A survey. Information Fusion. 27, 95–110 (2016). https://doi.org/10.1016/j.inffus.2015.06.002.
3. Gull, R., Shoaib, U., Rasheed, S., Abid, W., Zahoor, B.: Pre Processing of Twitter's Data for Opinion Mining in Political Context. Procedia Computer Science. 96, 1560–1570 (2016). https://doi.org/10.1016/j.procs.2016.08.203.
4. Moussa, M.E., Mohamed, E.H., Haggag, M.H.: A survey on opinion summarization techniques for social media. Future Computing and Informatics Journal. 3, 82–109 (2018). https://doi.org/10.1016/j.fcij.2017.12.002.
5. Park, J., Lee, B.K.: An opinion-driven decision-support framework for benchmarking hotel service. Omega. 103, 102415 (2021). https://doi.org/10.1016/j.omega.2021.102415.
6. Li, W., Guo, K., Shi, Y., Zhu, L., Zheng, Y.: DWWP: Domain-specific new words detection and word propagation system for sentiment analysis in the tourism domain. Knowledge-Based Systems. 146, 203–214 (2018). https://doi.org/10.1016/j.knosys.2018.02.004.
7. Sann, R., Lai, P.-C.: Understanding homophily of service failure within the hotel guest cycle: Applying NLP-aspect-based sentiment analysis to the hospitality industry. International Journal of Hospitality Management. 91, 102678 (2020). https://doi.org/10.1016/j.ijhm.2020.102678.

8. Mehraliyev, F., Kirilenko, A.P., Choi, Y.: From measurement scale to sentiment scale: Examining the effect of sensory experiences on online review rating behavior. Tourism Management. 79, 104096 (2020). https://doi.org/10.1016/j.tourman.2020.104096.
9. Li, S., Li, G., Law, R., Paradies, Y.: Racism in tourism reviews. Tourism Management. 80, 104100 (2020). https://doi.org/10.1016/j.tourman.2020.104100.
10. Álvarez-Carmona, M.Á., Aranda, R., Arce-Cárdenas, S., Fajardo-Delgado, D., Guerrero-Rodríguez, R., López-Monroy, A.P., Martínez-Miranda, J., Pérez-Espinosa, H., Rodríguez-González, A.: Overview of Rest-Mex at IberLEF 2021: Recommendation System for Text Mexican Tourism. Procesamiento del Lenguaje Natural. 67, (2021).
11. Vásquez Rojas, C., Roldán Reyes, E., Aguirre y Hernández, F., Cortés Robles, G.: Integration of a text mining approach in the strategic planning process of small and medium-sized enterprises. Industr Mngmnt & Data Systems. 118, 745–764 (2018). https://doi.org/10.1108/IMDS-01-2017-0029.
12. Kotu, V., Deshpande, B.: Chapter 15 - Getting Started with RapidMiner. In: Kotu, V. and Deshpande, B. (eds.) Data Science (Second Edition). pp. 491–521. Morgan Kaufmann (2019). https://doi.org/10.1016/B978-0-12-814761-0.00015-0.
13. Snowball: A language for stemming algorithms, http://snowball.tartarus.org/texts/introduction.html, last accessed 2021/06/05.
14. Salton, G., Buckley, C.: Term-weighting approaches in automatic text retrieval. Information Processing & Management. 24, 513–523 (1988). https://doi.org/10.1016/0306-4573(88)90021-0.
15. Ruping, S.: Incremental learning with support vector machines. In: Proceedings 2001 IEEE International Conference on Data Mining. pp. 641–642 (2001). https://doi.org/10.1109/ICDM.2001.989589.
16. Neural Net - RapidMiner Documentation, https://docs.rapidminer.com/latest/studio/operators/modeling/predictive/neural_nets/neural_net.html, last accessed 2021/06/05.