Sentiment Classification for Mexican Tourist Reviews based on K-NN and Jaccard Distance

Alejandra Romero-Cantón[1] and Ramon Aranda[2,3][0000−0001−8269−3944]

[1] Tecnologíco Nacinal de México, Campus Mérida, 97118, Yucatán, México
[2] Centro de Investigación Científica y de Educación Superior de Ensenada, Unidad de Transferencia Tecnológica Tepic (CICESE-UT3), 63173, Nayarit, Mexico
[3] Consejo Nacional de Ciencia y Tecnología (Conacyt), 03940, CDMX, Mexico

Abstract. In this paper is presented a propose solution to the Sentiment Analysis challenge presents in the Recommendation System for Text Mexican Tourism task during the Iberian Languages Evaluation Forum 2021. The task consists of predicting the polarity of an opinion issued by a tourist who traveled to the most representative places of Guanajuato, Mexico. Thus, our approach is based K-Nearest Neighbors by using a distance based on the Jaccard coefficient concept. In the training stage, by using the training data, our approach first clusters every word from every opinion (review) by the respective class. Then, the stop words from each cluster are deleted. After, the normalized frequency of each word in a class is computed. In this way, the set of words (*trained* words) with their normalized frequency (NF) are used as class feature vector. In the classification stage, when a new opinion is given, each word from it is intersect with the *trained* words for each class and the NF of the intersected words are summed (dissimilarity value). The predicted class is assigned to the class with the most high dissimilarity value. The performance on the testing data were of 1.26 MAE and 0.22 of F-measure. We think that the obtained results are because the data is unbalanced and our approach does not deal with that issue.

Keywords: K-NN · Jaccard Distance · Sentiment analysis · Mexican tourist texts.

1 Introduction

In 2018, the World Economic Forum reports that the travel & tourism industry generated 10.4% of the world GDP and supported over 319 million jobs [6]. In the last year, global tourism has been impacted strongly due to COVID-19 pandemic and in the last decade tourism has also been influenced by numerous technological advances and tools such as digitization, information and communication technology, machine learning, robotics, and artificial intelligence (AI) [12, 9, 10, 3].

IberLEF 2021, September 2021, Málaga, Spain.

Most of international travelers plan their trips by digital means, and a big part of their decisions rely on other travelers shared online information, e.g. online touristic reviews [5]. To synthesize large amounts of reviews, it is essential to use algorithms from the Artificial Intelligence field, specifically the area of the Natural Language Processing (NLP). his sub-field of the artificial intelligence aims to achieve human-like processing capabilities of the language for diverse scopes [4, 8]. NLP intersects artificial intelligence and linguistics [11] and covers a wide range of methods to analyze and represent naturally occurring text at one or more linguistic examination levels.

One task of the Recommendation System for Text Mexican Tourism task during the Iberian Languages Evaluation Forum 2021 [1] is classified the polarity (positive/negative) of an opinion issued by a tourist who traveled to the most representative places of Guanajuato, Mexico. This task is based on a sub–field of the PLN known as Sentiment analysis (SA) [2]. Thus, in this work, we propose a method to classify the polarity for mexican tourist reviews based on K-Nearest Neighbour (K-NN) and Jaccard Distance (JD).

This work is organized as follows:

- Section 2 describes the task to solve.
- Section 3 shows in details the proposal followed in this work.
- In section 4 the results are presented.
- Finally, section 5 presents the conclusions and limitations of our proposal.

2 Task Description

The subtask is a classification task where the participating system can predict the polarity of an opinion issued by a tourist who traveled to the most representative places of Guanajuato, Mexico. Guanajuato city is a well-known destination for domestic tourists and it has gained a progressive notoriety in the international arena since the last quarter of the previous century. Apart from famous international destinations within the Mexican territory such as the cases of Cancun and Mexico City, Guanajuato ranks number six among the most visited cities for tourism purposes[1]. Thus, this Sentiment Analysis problem is defined as follows:

- *"Given an opinion about a Mexican tourist place, the goal is to determine the polarity, between 1 and 5, of the text."* Where 1 indicates most negative and 5 most positive.

2.1 Data set

This collection was obtained from the tourists who shared their opinion on TripAdvisor between 2002 and 2020. Each opinion's class (review polarity) is an integer between [1, 5], where 1 represents the most negative polarity and 5 the

[1] https://www.datatur.sectur.gob.mx/SitePages/CompendioEstadistico.aspx

Table 1: Distribution of the polarity training data set.

Class (polarity)	Number of Instances (rows)	Percentage
1	80	1.54%
2	145	2.80%
3	686	13.20%
4	1596	30.71%
5	2690	51.76%
Total	5197	100%

most positive. Each tourist has information about nationality and gender. Rest-Mex organizers available two data sets[2] one for training and one for evaluation. Each instance (row) in the training and testing datasets contain the information as described below:

- **Index**: the index of each opinion.
- **Title**: The title that the tourist himself gave to his opinion.
- **Opinion**: The opinion expressed by the tourist.
- **Place**: Place that the tourist visited and to which the opinion is directed.
- **Gender**: the gender of the tourist.
- **Age**: The age of the tourist at the time of issuing the opinion.
- **Country**: The country of origin of the tourist.
- **Date**: the date when the review was issued.
- **Label**: it represents the polarity of the review, labels goes from 1 to 5. Note that for the testing data set, the labels values are unknown.

Training data set consists of 5197 instances. Table 1 shows the distribution of the review polarities for the training data set. It is important to mention that the training data set the classes are unbalanced. The test data set contained 2216 instances (the distribution polarity is unknown).

3 Proposed Approach

Our proposal consists in two main stages: training stage and classification stage. We describe each stage below.

3.1 Training stage

In this stage, we use the using the training data to extract features of each class. Thus, our approach first clusters every word from every opinion (review) by the respective class. Then, the stop words from each cluster are deleted. We call to the result set of words fro class c, *trained* words Ω_c. After, the normalized frequency, $\omega_{i,c}$, of i-th word in the class, c, is computed. In this way, the sets Ω_c with their normalized frequency $\omega_{i,c}$ (for $c \in 1, 2, 3, 4, 5$ and $i = 1, 2, ..., N_c$ where $N_c = |\Omega_c|$) are used as class feature vector.

[2]https://sites.google.com/cicese.edu.mx/rest-mex-2021

3.2 classification stage

In the classification stage, when a new opinion/review is given, the stop words from it are deleted. Then the resulting set of words for that opinion is called Θ. After, each word in Θ is intersect with the set Ω_c (*trained* words of class c). Then, the NF of the intersected words are added. This can be represented by equation 1:

$$S_c = \sum_{k \in \Omega_c \cap \Theta} \omega_{k,c} \tag{1}$$

Note that equation 1 is based on the concept of the Jaccard Distance [7]. Thus, the predicted class for the review Θ, $C(\Theta)$, is assigned to the class with the most high similarity value S_c:

$$C(\Theta) = \arg\max_c \{S_c\} \, \forall \, c \in \{1, 2, 3, 4, 5\} \tag{2}$$

Equation 2 is the representation of the K-NN method, when K= 1.

4 Results

Figure 1 show the wordclouds of the sets Ω_c weighed with their corresponding $\omega_{k,c}$ for all classes. Note that although there are many words repeated in all classes words as "museo" and "momias" are more frequent in classes 1 and 2, and word as "guanajuato" and "historia" are more frequent in classes 3 to 5.

The official results for our proposal are as follows:

- Accuracy: 36.95
- F-measure: 0.22
- MAE: 1.27

In this sense our approach obtained the last place according to MAE value (15th place). According to accuracy, we obtained the 12th place. Finally, we obtained the 9th place with accordance to F-measure.

5 Conclusions

In this work, we presented a simple solution based on the concept of the Jaccar Distance to classify the sentiment analysis problem presented on Recommendation System for Text Mexican Tourism task during the Iberian Languages Evaluation Forum 2021. Although, our proposal is based in a simple idea it showed potential. The most significant disadvantage of our approach was the unbalance training data set. Additionally, our proposal could be improved by removing the representative words as subjects and only work with qualifying adverbs.

(a) Label/Class 1

(b) Label/Class 2

(c) Label/Class 3

(d) Label/Class 4

(e) Label/Class 5

Fig. 1: Wordclouds for Ω_c weighed with their corresponding $\omega_{k,c}$ for all the classes

References

1. 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)
2. Anis, S., Saad, S., Aref, M.: A survey on sentiment analysis in tourism. International Journal of Intelligent Computing and Information Sciences **20**(1), 1–15 (2020). https://doi.org/10.21608/IJICIS.2020.106309
3. Buhalis, D.: Technology in tourism-from information communication technologies to eTourism and smart tourism towards ambient intelligence tourism: a perspective article. Tourism Review **75**(1), 267–272 (Jan 2020). https://doi.org/10.1108/TR-06-2019-0258, https://doi.org/10.1108/TR-06-2019-0258, publisher: Emerald Publishing Limited
4. Cai, T., Giannopoulos, A.A., Yu, S., Kelil, T., Ripley, B., Kumamaru, K.K., Rybicki, F.J., Mitsouras, D.: Natural language processing technologies in radiology research and clinical applications. Radiographics **36**(1), 176–191 (2016)

5. Calderón, F.A.C., Blanco, M.V.V.: Impacto de internet en el sector turístico. Revista UNIANDES Episteme **4**(4), 477–490 (2017)
6. Calderwood, L.U., Soshkin, M.: The travel and tourism competitiveness report 2019 (Sep 2019)
7. Álvarez Carmona, M.A., Franco-Salvador, M., Villatoro-Tello, E., Montes-y Gómez, M., Rosso, P., Villaseñor-Pineda, L.: Semantically-informed distance and similarity measures for paraphrase plagiarism identification. Journal of Intelligent & Fuzzy Systems **34**(5), 2983–2990 (2018). https://doi.org/10.3233/JIFS-169483, publisher: IOS Press
8. Chowdhury, G.G.: Natural language processing. Annual review of information science and technology **37**(1), 51–89 (2003)
9. Gossling, S., Scott, D., Hall, C.M.: Pandemics, tourism and global change: a rapid assessment of covid-19. Journal of Sustainable Tourism **29**(1), 1–20 (2021). https://doi.org/10.1080/09669582.2020.1758708, https://doi.org/10.1080/09669582.2020.1758708
10. Guerra-Montenegro, J., Sanchez-Medina, J., Lana, I., Sanchez-Rodriguez, D., Alonso-Gonzalez, I., Del Ser, J.: Computational intelligence in the hospitality industry: A systematic literature review and a prospect of challenges. Applied Soft Computing **102**, 107082 (2021). https://doi.org/https://doi.org/10.1016/j.asoc.2021.107082, https://www.sciencedirect.com/science/article/pii/S1568494621000053
11. Nadkarni, P.M., Ohno-Machado, L., Chapman, W.W.: Natural language processing: an introduction. Journal of the American Medical Informatics Association **18**(5), 544–551 (2011)
12. Qiu, R.T., Park, J., Li, S., Song, H.: Social costs of tourism during the covid-19 pandemic. Annals of Tourism Research **84**, 102994 (2020). https://doi.org/https://doi.org/10.1016/j.annals.2020.102994, https://www.sciencedirect.com/science/article/pii/S0160738320301389