Classifying Tourist Text Reviews by Means of Mutual Information Features

Alvaro Zaid Gallardo-Hernández[1,*], Ramón Aranda[2,3] and Angel Diaz-Pacheco[4]

[1]*Departamento de Ciencias e Ingenieras, Universidad Iberoamericana Puebla, San Andrés Cholula, Puebla, México.*

[2]*Centro de Investigación en Matemáticas, Sede Mérida, Mérida, Yucatán, México.*

[3]*Consejo Nacional de Humanidades, Ciencias y Tecnologías, Ciudad de México, México*

[4]*Departamento de Ingeniería Electrónica, División de Ingenierías, Universidad de Guanajuato – Campus Irapuato-Salamanca, Yuriria, Mexico*

Abstract

This paper introduces a proposed solution for the classification of tourist text reviews. The problem was initially presented at Rest-Mex 2023: Research on Sentiment Analysis Task for Mexican Tourist Texts. The objective of this task is to determine the polarity (1 and 5), the type of opinion (hotel, restaurant, or attraction), and the country (Mexico, Cuba, Colombia) associated with a given set of reviews. Our approach is primarily based on the Mutual Information (MI) measure. During the training stage, our approach involves clustering each word from the provided training data according to their respective classes. Subsequently, we compute the MI value of each word within each class. Additionally, we generate synonyms for each word and incorporate them into a set, associating them with the same MI value as their respective word. This set of words, referred to as "trained" words, along with their normalized MI values, is utilized as class features. In the classification stage, when a new instance is provided, each word is compared with the "trained" words belonging to each class. The MI values of the intersected words are then summed. The predicted class is assigned based on the class with the highest sum value.

Keywords

Mutual Information, Sentiment Analysis, Rest-Mex

1. Introduction

In the 2019 edition of the "Travel & Tourism Competitiveness Report" (TTCR) [1], published by the World Economic Forum, it was reported that the Travel & Tourism (T&T) sector was experiencing remarkable growth. The World Tourism Organization (UNWTO) stated that international tourist arrivals worldwide reached 1.4 billion in 2018, surpassing earlier predictions by two years. However, the findings of the TTCR also raised concerns about a potential tipping point where the relentless pursuit of growth and competitiveness in the sector could undermine the very assets on which it depends.

Fast forward two years, and the T&T sector looks drastically different. The COVID-19 pandemic had a devastating impact on the demand for travel, hitting the sector particularly

IberLEF 2023, September 2023, Jaén, Spain

*Corresponding author.

192163@iberopuebla.mx (A. Z. Gallardo-Hernández); arac@cimat.mx (R. Aranda); angel.diaz@ugto.mx (A. Diaz-Pacheco)

0000-0001-8269-3944 (R. Aranda)

hard. Shutdowns, travel restrictions, and the disappearance of international travel not only severely affected companies but also tourism-dependent national economies. Fortunately, there are now positive indications of recovery, although the pace and progress vary across different regions and market segments. Additionally, the complexities of this uneven recovery are further compounded by factors like the war in Ukraine [2].

As a result, the T&T sector and its customers have likely undergone permanent changes. Travelers have become more discerning, especially regarding the health and hygiene conditions of potential destinations. They are also cautious about the potential impact of future COVID variants, as well as challenges arising from government policies, border closures, and travel disruptions. Furthermore, the pause in international travel allowed leisure and business travelers to reflect on the environmental consequences of their choices. Consequently, governments and T&T businesses have had to reassess their investments, develop strategies to mitigate risk and volatility in demand, and adapt to the changing expectations of their customers . Additional to the COVID-19 pandemic impacts, 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) [3, 4, 5, 6, 7, 8]. Thus, most of international travelers plan their trips by digital means, and a part of their decisions are based on information online [9].

One task of the *Rest-Mex 2023: Research on Sentiment Analysis Task for Mexican Tourist Texts*[10] is to determine the polarity (1 and 5), the type of opinion (hotel, restaurant, or attraction), and the country (Mexico, Cuba, Colombia) associated with a given set of reviews. For this reason, it is essential to use algorithms from the Artificial Intelligence field, specifically the area of the Natural Language Processing (NLP) to achieve human-like processing capabilities of the language for diverse scopes [11, 12]. NLP intersects artificial intelligence and linguistics [13] and covers a wide range of methods to analyze and represent naturally occurring text at one or more linguistic examination levels, for example see [14, 15, 16, 17, 18]. Thus, in this work, we propose a method to predict the classes based on the Mutual Information measure [19].

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; and finally, section 5 presents the conclusions and limitations of our proposal.

2. Task Description: Sentiment Analysis

The task at hand involves a classification problem where participating systems are tasked with predicting the polarity, type and country of opinions expressed by tourists who have visited attractions, restaurants, and hotels in Mexico, Cuba, and Colombia. The dataset used for this task was compiled from opinions shared by tourists on TripAdvisor between 2002 and 2022. Each opinion belongs to a specific class represented by an integer ranging from 1 to 5, where 1 indicates the most negative polarity and 5 denotes the most positive. Additionally, each opinion is labeled with its corresponding type (hotel, restaurant, or attraction). For example:

- "Un callejón donde tienes que besar a tu amante por años de felicidad, en el amor es parte de un mito en esta ciudad especial. El callejón estrecho con escalones no es muy especial en sí mismo. Lo que lo hace especial es toda la historia a su alrededor."

- **Polarity**: 5 (Very positive)
- **Type**: Attractive
- **Location**: Mexico

To evaluate the results of the polarity task, the organizer proposed to give more weight to minority classes. For the sentiment analysis collection of the Rest-Mex, the minority classes are the ones with the most negative polarities. Therefore, for this edition, to evaluate the result of the polarity classification, it is as follows:

$$Res_P(k) = \frac{\sum_{i=1}^{|C|}((1 - \frac{T_{c_i}}{T_c}) * F_i(k))}{\sum_{i=1}^{|C|} 1 - \frac{T_{c_i}}{T_c}}, \tag{1}$$

where k is a forum participant system, $C = 1, 2, 3, 4, 5$, T_c is the total instances in the collection, T_{c_i} is the total is instances in the class i. Finally, $F_i(K)$ is the F-measure value for the class i obtained by the system k. Thus, this formulation gives more weight to the classes with less instances. For the type ($Res_A(k)$) and country ($Res_A(C)$) classification, the organizer proposed only to average the F-measure values corresponding their respective classes. Finally, the final score of the whole task for the k system is given by:

$$Sentiment(k) = \frac{2 * Res_P(k) + Res_A(k) + Res_C(k)}{4} \tag{2}$$

3. Proposed Approach

To attack the sentiments analysis task in tourist data, it is proposed to use simple features that can capture important information to determine the polarity of an opinion in such a way that it is quick to calculate and represent. Especially to offer an option for restricted applications in time or memory (such as IoT solutions) and that cannot use approaches that, although they have outstanding effectiveness results, can be slow or use much computational power, in addition to having the advantage of being language-independent features.

Our proposal is a improve of a previous work [20], it consists in three main stages: prepossessing, training and classification. For the prepossessing stage, we applied to the text next steps:

- Uppercase was converted to lowercase.
- Stop-words were removed.
- Punctuation marks were removed.
- The digits were replaced by the letter 'd'.
- Stemming was applied to the tokens in the texts.
- Removed tokens that appear less than 50 times in the data set.

3.1. Training stage

The main different from [20] is in the this stage. In this stage, we use the using the training data to extract features of each subtask (polarity, type and country). Thus, to analyze the information

from the dataset, similar to [21], we propose to use the well-known Mutual Information (MI) measure. The MI measure was applied to all training data to extract the features for each epidemiological color (red, orange, yellow and green) [22]. This measure basically computes the mutual dependence between two variables X and Y (information that X and Y share). MI is computed by the following equation:

$$MI(X,Y) = P(X,Y)Log(P(X,Y)/P(X)P(Y)), \tag{3}$$

where $P(X,Y)$ is the joint probability between the variables X and Y. For example, if X and Y are independent, then X is not important and does not exert any influence over Y and vice versa; then MI would be close to zero. Conversely, if X is describe in terms of Y (or Y is in terms of X), then all information conveyed by X is shared with Y [23]. In our case, MI measures the influence of a word $X = b$ with $b \in B =$ {all the words in the collections} in a class $Y = c$ with $c \in C =$ {classes in subtask}:

- If a word b appears in all classes, then it is not relevant in any way, resulting in $MI(b,c) \approx 0$. The intuitive idea is that such word b does not help to discriminate among different classes (epidemiological colors).
- If the word b is almost exclusive to a class c, then this word is considered valuable for c, and the expected result would be $MI(b,c) > 0$. The intuition is that the higher the MI score, the more representative the word is to the class (epidemiological colors)..
- If a word appears repeatedly in other classes but not in class c, the result would be $MI(b,c) < 0$. The idea is that the lower the MI score, the less useful is the word to represent the class.

MI potentially reveals representative words for each class. Thus, it is possible to detect exclusive words describing the reviews on each class [24]. However, to set of words obtained by MI values, we add up to 5 synonymous to give a better representation of the classes. Thus, we call to the result set of words and and MI measures for class c, *trained* feature set Ω_c. The i-th element, $\omega_{i,c} \in \Omega_c$ is a tuple of values, $(\omega^w_{i,c}, \omega^{MI}_{i,c})$, where $\omega^w_{i,c}$ represents the i-th word and $\omega^{MI}_{i,c}$ represents the normalized MI measure for $\omega^w_{i,c}$. Note that the MI value for each synonym is the same as its word of origin.

3.2. Classification stage

The classification stage is the same as [20], when a new instance is given, first the prepossessing steps are applied. Then the resulting set of words for the instance is called Θ. After, Θ is intersected with the words in set Ω^w_c (set of *trained* words, $\omega^w_{i,c}$, for class c). Then, we compute the sum of the values $\omega^{MI}_{k,c}$ for $k \in \Theta \cap \Omega^w_c$. This can be represented by equation 4:

$$S_c = \sum_{k \in \Omega^w_c \cap \Theta} \omega^{MI}_{k,c} \tag{4}$$

Thus, the predicted class for a instance Θ, $C(\Theta)$, is assigned to the class with the most high similarity value S_c:

$$C(\Theta) = \arg\max_c \{S_c\} \tag{5}$$

with $c \in C$ represents the possibles class values for each subtask (polarity, type and country). For example, for Type, C ={hotel, restaurant, atractive}

4. Results

The official results for our proposal shows that we obtained a sentiment score (equation 2) of 0.229. In this sense our approach obtained the last place in the task. The obtained macro F-measures were of 0.183, 0.301 and 0.250 for the subtasks polarity, Type and Country respectively. Although, our proposal uses a simple idea, we aim to have a accuracy of 55.67 for polarity, 44.39 for Type and 47.25 for Country.

5. Conclusions

In this study, we introduced a straightforward solution for the Sentiment Analysis Task of Rest-Mex 2023, utilizing the Mutual Information measure. Despite its simplicity, our approach demonstrated promising potential. However, we observed a significant drawback in our methodology, namely the imbalance within the training dataset. Furthermore, we discovered that numerous meaningless words (e.g., *queretarcdm*, *metrocdmx*, etc.) exhibited high MI values, yet these words were essentially noise present in the dataset. Therefore, to enhance our proposal, it is imperative to address this issue by removing such meaningless words from consideration.

References

[1] L. U. Calderwood, M. Soshkin, The travel and tourism competitiveness report 2019, 2019.

[2] Travel & tourism development index 2021, rebuilding for a sustainable and resilient future, 2022.

[3] R. T. Qiu, J. Park, S. Li, H. Song, Social costs of tourism during the covid-19 pandemic, Annals of Tourism Research 84 (2020) 102994. URL: https://www.sciencedirect.com/science/article/pii/S0160738320301389. doi:https://doi.org/10.1016/j.annals.2020.102994.

[4] S. Gossling, D. Scott, C. M. Hall, Pandemics, tourism and global change: a rapid assessment of covid-19, Journal of Sustainable Tourism 29 (2021) 1–20. URL: https://doi.org/10.1080/09669582.2020.1758708. doi:10.1080/09669582.2020.1758708. arXiv:https://doi.org/10.1080/09669582.2020.1758708.

[5] J. Guerra-Montenegro, J. Sanchez-Medina, I. Lana, D. Sanchez-Rodriguez, I. Alonso-Gonzalez, J. Del Ser, Computational intelligence in the hospitality industry: A systematic literature review and a prospect of challenges, Applied Soft Computing 102 (2021) 107082. URL: https://www.sciencedirect.com/science/article/pii/S1568494621000053. doi:https://doi.org/10.1016/j.asoc.2021.107082.

[6] D. Buhalis, Technology in tourism-from information communication technologies to eTourism and smart tourism towards ambient intelligence tourism: a perspective article, Tourism Review 75 (2020) 267–272. URL: https://doi.org/10.1108/TR-06-2019-0258. doi:10.1108/TR-06-2019-0258, publisher: Emerald Publishing Limited.

[7] A. Diaz-Pacheco, M. A. Álvarez-Carmona, R. Guerrero-Rodríguez, L. A. C. Chávez, A. Y. Rodríguez-González, J. P. Ramírez-Silva, R. Aranda, Artificial intelligence methods to support the research of destination image in tourism. a systematic review, Journal of Experimental & Theoretical Artificial Intelligence 0 (2022) 1–31. doi:10.1080/0952813X.2022.2153276.

[8] M. A. Álvarez-Carmona, R. Aranda, A. Y. Rodríguez-Gonzalez, D. Fajardo-Delgado, M. G. Sánchez, H. Pérez-Espinosa, J. Martínez-Miranda, R. Guerrero-Rodríguez, L. Bustio-Martínez, Ángel Díaz-Pacheco, Natural language processing applied to tourism research: A systematic review and future research directions, Journal of King Saud University - Computer and Information Sciences 34 (2022) 10125–10144. URL: https://www.sciencedirect.com/science/article/pii/S1319157822003615. doi:https://doi.org/10.1016/j.jksuci.2022.10.010.

[9] F. A. C. Calderón, M. V. V. Blanco, Impacto de internet en el sector turístico, Revista UNIANDES Episteme 4 (2017) 477–490.

[10] M. Á. Álvarez-Carmona, Á. Díaz-Pacheco, R. Aranda, A. Y. Rodríguez-González, L. Bustio-Martínez, V. Muñis-Sánchez, A. P. Pastor-López, F. Sánchez-Vega, Overview of rest-mex at iberlef 2023: Research on sentiment analysis task for mexican tourist texts, Procesamiento del Lenguaje Natural 71 (2023).

[11] T. Cai, A. A. Giannopoulos, S. Yu, T. Kelil, B. Ripley, K. K. Kumamaru, F. J. Rybicki, D. Mitsouras, Natural language processing technologies in radiology research and clinical applications, Radiographics 36 (2016) 176–191.

[12] G. G. Chowdhury, Natural language processing, Annual review of information science and technology 37 (2003) 51–89.

[13] P. M. Nadkarni, L. Ohno-Machado, W. W. Chapman, Natural language processing: an introduction, Journal of the American Medical Informatics Association 18 (2011) 544–551.

[14] M. A. Álvarez-Carmona, A. P. López-Monroy, M. Montes-y Gómez, L. Villasenor-Pineda, H. Jair-Escalante, Inaoe's participation at pan'15: Author profiling task, Working Notes Papers of the CLEF 103 (2015).

[15] M. E. Aragón, M. A. Álvarez-Carmona, M. Montes-y Gómez, H. J. Escalante, L. V. Pineda, D. Moctezuma, Overview of mex-a3t at iberlef 2019: Authorship and aggressiveness analysis in mexican spanish tweets., in: IberLEF@ SEPLN, 2019, pp. 478–494.

[16] M. Á. Álvarez-Carmona, R. Aranda, S. Arce-Cárdenas, D. Fajardo-Delgado, R. Guerrero-Rodríguez, A. P. López-Monroy, J. Martínez-Miranda, H. Pérez-Espinosa, A. Rodríguez-González, Overview of rest-mex at iberlef 2021: Recommendation system for text mexican tourism, Procesamiento del Lenguaje Natural 67 (2021). doi:https://doi.org/10.26342/2021-67-14.

[17] M. Á. Álvarez-Carmona, Á. Díaz-Pacheco, R. Aranda, A. Y. Rodríguez-González, D. Fajardo-Delgado, R. Guerrero-Rodríguez, L. Bustio-Martínez, Overview of rest-mex at iberlef 2022: Recommendation system, sentiment analysis and covid semaphore prediction for mexican tourist texts, Procesamiento del Lenguaje Natural 69 (2022).

[18] M. Á. Álvarez-Carmona, E. Villatoro-Tello, L. Villaseñor-Pineda, M. Montes-y Gómez, Classifying the social media author profile through a multimodal representation, in: Intelligent Technologies: Concepts, Applications, and Future Directions, Springer, 2022, pp. 57–81.

[19] C. E. Shannon, A mathematical theory of communication, The Bell System Technical Journal 27 (1948) 379–423. doi:10.1002/j.1538-7305.1948.tb01338.x.

[20] A. Romero-Cantón, R. Aranda, AngelDiaz-Pacheco, J. P. Ramírez-Silva, Mexican epidemiological semaphore color prediction by means of mutual information features, in: CEUR Workshop Proceedings, Coruña, Spain, 2022.

[21] R. Guerrero-Rodriguez, M. Á. Álvarez-Carmona, R. Aranda, A. P. López-Monroy, Studying online travel reviews related to tourist attractions using nlp methods: the case of guanajuato, mexico, Current Issues in Tourism (2021) 1–16. doi:https://doi.org/10.1080/13683500.2021.2007227.

[22] R. D. Hjelm, A. Fedorov, S. Lavoie-Marchildon, K. Grewal, P. Bachman, A. Trischler, Y. Bengio, Learning deep representations by mutual information estimation and maximization, arXiv preprint arXiv:1808.06670 (2018).

[23] M. Ravanelli, Y. Bengio, Learning speaker representations with mutual information, arXiv preprint arXiv:1812.00271 (2018).

[24] M. Á. Álvarez-Carmona, M. Franco-Salvador, E. Villatoro-Tello, M. Montes-y Gómez, P. Rosso, L. Villaseñor-Pineda, Semantically-informed distance and similarity measures for paraphrase plagiarism identification, Journal of Intelligent & Fuzzy Systems 34 (2018) 2983–2990. doi:10.3233/JIFS-169483, publisher: IOS Press.