Thematic Unsupervised Classification of Tourist Texts using Latent Semantic Analysis and K-Means

Julio Madera-Quintana[1*], Anibal Hernández-Gónzalez[1] and Yoan Martínez-López [1]

[1] *University of Camagüey, Camagüey, Cuba.*

Abstract

Sentiment analysis and thematic unsupervised classification of tourist texts have gained importance in recent years. Rest-Mex 2023 proposes two tasks: sentiment analysis and thematic unsupervised classification of Mexican tourist texts. The thematic unsupervised classification task aims to group texts related to Mexican tourism into four distinct topics. In this paper, we propose a procedure based on the TF-IDF and LSA algorithms to convert texts into vectors and execute de K-Means method for clustering of the text. The results show that the proposed method is comparable to others in the same competition.

Keywords

Thematic Unsupervised Classification, Mexican Tourism Texts, Rest-Mex 2023, Clustering Algorithms, Text Vectorization

1. Introduction

Sentiment Analysis and thematic unsupervised classification are important tasks in Natural Language Processing, which have been applied in various domains, including tourism. These tasks aim to extract meaningful information from texts, such as opinions, sentiments, and topics, among others. In the context of tourism, the analysis of tourist texts can provide valuable insights for the improvement of tourist services and destinations[18-20]. However, most studies in this area have focused on English texts, leaving a gap in research for other languages, such as Spanish [21-23]. In this context, Rest-Mex 2023 proposes two tasks: sentiment analysis and thematic unsupervised classification of Mexican tourist texts [1].

Text Clustering is a popular technique in natural language processing that involves grouping similar documents based on their content. The goal of Text Clustering is to discover patterns and relationships in unstructured text data, which can be used for various applications such as Information Retrieval, Recommendation Systems, and Sentiment Analysis. In recent years, there has been a growing interest in applying Text Clustering to Spanish language text given the increasing amount of digital content available in this language [2].

One of the most important challenges in Text Clustering is to find an appropriate representation of the text data that captures its semantic content. In the case of Spanish language text, this challenge is compounded by the complexity of the language, which has a rich vocabulary and complex grammar. To address this challenge, researchers have explored various techniques such as word embeddings, topic models, and graph-based representations [2, 3].

Another important aspect of Text Clustering is the choice of the clustering algorithm. There are various methods available, such as K-Means [4, 5], Hierarchical Clustering [6], and Spectral Clustering [7], each with its strengths and weaknesses. In the context of Spanish language text, researchers have explored the effectiveness of different clustering algorithms, as well as hybrid approaches that combine multiple algorithms [8, 9].

IberLEF 2023, September 2023, Jaén, Spain

EMAIL: julio.madera@reduc.edu.cu (J. Madera-Quintana); anibal.hernandez@reduc.edu.cu (A. Hernández-González); yoan.martinez@reduc.edu.cu (Y. Martínez-López)

ORCID: 0000-0001-5551-690X (J. Madera-Quintana); 0009-0006-5337-7954 (A. Hernández-González); 0000-0002-1950-567X (Y. Martínez-López)

CEUR Workshop Proceedings (CEUR-WS.org)

Evaluation of Text Clustering results is another important aspect of this field. It is essential to measure the quality of the clustering results and compare them against ground truth or human-labelled data. In the case of Spanish language text, several evaluation metrics have been proposed, such as the Normalized Mutual Information (NMI), F-measure, and Adjusted Rand Index (ARI) [10 - 12].

The application of Text Clustering to real-world problems in Spanish language text is an area of active research. There is still much to be explored in this field, and researchers are actively working to develop more effective methods and applications for Text Clustering in Spanish. The code implementation is available at: https://github.com/ajhglez99/rest_mex_2023_clustering_task.git.

The paper is organized as follows. In the next section, we present the proposed methodology. Next, we discuss the result of the proposed procedure. In the last section, we present the conclusions and future work.

2. Methodology

In the Rest-Mex 2023, for the thematic unsupervised classification task, the same dataset of 150000 news items related to Mexican tourism was used. The news items were carefully downloaded and tagged based on four different topics related to tourism. The goal of the task was to group the news items into four distinct topics using unsupervised clustering techniques. Several clustering algorithms were applied, including K-Means, and hierarchical clustering. The performance of the algorithms was evaluated using internal validation metrics such as silhouette score and external validation metrics such as purity and entropy.

2.1. Proposed procedure

In this section, we present the general procedure for text data preprocessing and clustering. Algorithm 1 shows the method executed over the Rest-Mex dataset on a thematic unsupervised task. First, it opens the dataset and then preprocesses it by removing links, special characters, numbers, and stop words and converts it to lowercase and strips white spaces [13]. In step 3, the algorithm applies the TF-IDF algorithm to convert the preprocessed text into a matrix of vectors [14]. Step 4 performs dimensionality reduction on the matrix of vectors by using the Latent Semantic Analysis (LSA) technique [15]. In step 5, we initialize the K-Means clustering algorithm to obtain four clusters, fit the model and apply this to get the labels of each text [16]. Finally, it returns the cluster labels ready to export in the Rest-Mex output format. The complexity of the algorithm depends on the size of the dataset and the number of clusters specified for K-Means [17].

ALGORITHM 1: PROCEDURE TO CLUSTERING REST-MEX TOURISM TEXTS

```
    Input: Rest-Mex thematic unsupervised classification of tourist texts dataset
    Output: Cluster label for each text in the dataset
1   dataset = load_dataset()
2   cleaned_dataset = preprocess(dataset)
3   vectors = tfidf(cleaned_dataset)
4   vectors = dimensionality_reduction_lsa(vectors)
5   labels = kmeans(vectors)
6   return labels
```

For the TF-IDF algorithm, we configure this to take into account only words that occur at least five times; for each text, it returns a vector with 300 components. After that, the LSA algorithm is executed to reduce each vector to a new vector with 100 components. We execute the LSA several times with a different number of components to return and the better results were with 100 components. Finally, the K-Means algorithm is executed with the parameter number of clusters equal to four. We output the two better runs of all the experiments executed moving some of these parameters.

The TF-IDF Vectorizer

TF-IDF stands for Term Frequency - Inverse Document Frequency, it is a common algorithm used in Text Mining and Natural Language Processing. It weights the importance of words based on how frequently they appear in a document (Term Frequency) and how unique they are across all documents (Inverse Document Frequency) [16]. This helps distinguish common but irrelevant words from important discriminative words. Reduces the impact of very common words that appear frequently in most documents but do not add much meaning. This makes the model more robust and focused on descriptive content words.

After calculating TF-IDF scores for all words in all documents, it represents each document as a vector of TF-IDF weights. This vector representation allows the use of clustering algorithms like K-Means or Hierarchical Clustering on the documents. The resulting clusters tend to group documents with similar topics because they share words with higher TF-IDF scores (i.e. relevant and distinctive terms for that topic).

In summary, TF-IDF helps extract meaningful features (words) from text data that are more suitable for clustering algorithms. It ignores common words and focuses on words that can better distinguish between documents and clusters. This often results in more coherent and interpretable clusters of text documents [16].

Latent Semantic Analysis

Latent Semantic Analysis (LSA) is a text mining technique that is useful for Text Clustering. LSA uses a technique called Singular Value Decomposition (SVD) to reduce the dimensionality of the word-document matrix. This reduces noise and isolates the main semantic themes in the text data. The reduced dimensional space captures the "latent semantics" of the text, meaning the underlying concepts and topics that the words are referring to. This more semantic representation of the text is better for clustering [15, 17].

After applying LSA, semantically similar documents (i.e. about the same topic) will be represented by similar vectors in the reduced dimensional space. This makes it easier for clustering algorithms to group these documents. The reduced dimensional space helps eliminate some of the problems in languages. Words with similar meanings (synonyms) will be placed close together, and ambiguous words (polysemes) will be assigned more distinct representations based on their different uses.

By experimenting with the number of dimensions to reduce, we can control the level of specificity vs generality in the document representations and resulting clusters. More dimensions tend to produce more fine-grained clusters. LSA produces a more semantic representation of text data by identifying the major factors (topics) underlying word usage. This semantic representation aligns similar documents together and separates unrelated documents, resulting in more coherent text clusters [15, 17].

K-Means Algorithm

K-Means is a simple and efficient algorithm that can handle large amounts of text data and produce clusters quickly. This makes it suitable for applications that require clustering large corpora of documents like the Rest-Mex dataset. It requires the number of clusters (k) as input. This enables us to control how many clusters/topics we want the algorithm to discover in the text data. Without prescribing specific topics, K-Means finds the most suitable clusters based on the text representations [16, 17].

K-Means works by iteratively assigning documents to clusters based on distance from cluster centroids and re-calculating the centroids based on the assigned documents. This optimization procedure tends to converge on coherent text clusters. This algorithm works well when combined with text feature extraction or dimensionality reduction techniques like TF-IDF, LSA and Word2Vec, the first two of these applied in our proposal. These techniques produce vector representations of texts that

K-Means can easily work with. The resulting clusters represent the dominant topics present in the text corpus. We can inspect the cluster centroids (average vector of assigned documents) to determine the words and documents that best represent each cluster/topic [16].

In summary, K-Means offers a computationally efficient algorithm for discovering a specified number of topics or clusters in text data. When combined with appropriate text feature extraction techniques, it tends to produce reasonably good quality text clusters with interpretable topics. Also, the resulting clusters are not necessarily the optimal partitioning of the data - the algorithm can get stuck in local optima. However, in practice, K-Means often produce useful and meaningful text clusters for exploration.

In the case of the Rest-Mex thematic task, we know the number of clusters to detect, K-Means is an excellent algorithm to detect these. As we described in the preceding paragraphs, we performed a pre-processing of the texts by converting incorrectly encoded characters, and removing URLs and stop words. Then we converted the text into numerical vectors using the TF-IDF technique, built the groups based on a distance metric, applied a dimensionality reduction process with the LSA algorithm, and finally applied evaluation metrics to identify how relevant the groups are.

3. Results and discussion

The results showed that the thematic unsupervised classification task can be performed with reasonable accuracy on Mexican tourist texts. To evaluate each system in the unsupervised classification task, an alignment must first be done. Given the Gold Standard, the output of each k system must be renumbered so that the themes correspond. This is because the only restriction that the participating teams have is that they must make 4 groups with the news shared in the competition. This means that the labels do not necessarily coincide for the same groups expected in the Gold Standard. For this reason, a re-labelling will be done for each system using the Gold Standard label that shares the most instances with each of the groups resulting from the k system. Once the alignment is done, it will be evaluated with a macro-F-measure as shown in the following equation.

$$Thematic(k) = \frac{1}{|L|}\sum_{i=1}^{|L|} F_i(k)$$

General ranking of teams

The results of the run of the following algorithms for each team are shown in Table 1. Our team JCMQ-Team_run_5 used the LSA algorithm proposal for the competition.

Table 1 shows the performance of four teams in the task measured by two metrics: Macro F1 and Accuracy. Macro F1 is a measure of the overall quality of the predictions made by the teams, while Accuracy is a measure of how many of the predictions were correct. The team that achieved the best performance was "Javilonso-Team_javier_alonso_thematic_spacy_kmeans10_5", with a Macro F1 score of 0.282 and an Accuracy score of 44.8. The team that came in second place was "CIMAT-Team_run3_thematic", with a Macro F1 score of 0.240 and an Accuracy score of 35.6. The team that came in third place was "JCMQ-Team_run_5", with a Macro F1 score of 0.218 and an Accuracy score of 35.3. Although this team did not perform as well as the top two teams, they still achieved a respectable score and were able to beat out the fourth team.

Finally, the fourth team was "MCE-Team_2ndIterKmeans", with a Macro F1 score of 0.203 and an Accuracy score of 34.3. This team achieved the lowest score of the four teams and was given an Honorable Mention (HM). In summary, the results show that "Javilonso-Team_javier_alonso_thematic_spacy_kmeans10_5" was the winner, with "CIMAT-Team_run3_thematic" coming in second and "JCMQ-Team_run_5" in third place.

The third-place team was "JCMQ-Team_run_5". They achieved a Macro F1 score of 0.218 and an Accuracy score of 35.3. Compared to the team in second place, they had a slightly lower Macro F1 score but a slightly higher Accuracy score. Based on these metrics, it seems that the team performed reasonably well in the task, but there was room for improvement. It's possible that they could have achieved a higher score if they had used a different approach or if they had more time to refine their model. Overall, coming in third place is still a good achievement, as it shows that the team was able to compete effectively against other teams.

Table 1
General result for each team in the thematic unsupervised track at Rest-Mex 2023.

Ranking	Run	Macro F1	Accuracy
1st	Javilonso-Team_javier_alonso_thematic_spacy_kmeans10_5	0.282753376	44.81711043
2nd	CIMAT-Team_run3_thematic	0.240045779	35.62898298
3rd	JCMQ-Team_run_5	0.218224315	35.27660652
HM	MCE-Team_2ndIterKmeans	0.203086076	34.30292449

Evaluating Precision and Recall

Now, we evaluate the performance of each team by category (Insecurity, Prices, Gastronomy and Landscape) using precision and recall metrics. For this, we use two metrics, Precision and Recall.

Precision is defined as:

$$precision = \frac{TP}{TP + FP}$$

Recall is defined as:

$$Recall = \frac{TP}{TP + FN}$$

Where:

- TP: True Positive
- FP: False Positive
- TN: True Negative
- FN: False Negative

Precision results

Table 2 shows the results of the Precision metric. Avg Precision is the average precision score across all target categories for that team's model run. Higher is better, indicating a higher proportion of the model's retrieved examples were relevant. Precision for each target category shows the precision score specifically for that category. This indicates the proportion of retrieved examples that were relevant for that category.

Team3 had the highest average precision at 0.4208, indicating a higher proportion of its overall retrieved examples were relevant on average. Team1 had the highest precision for the Insecurity and Gastronomy categories, while Team2 had the highest for Prices. This suggests these teams' models retrieved a higher proportion of relevant examples for those specific categories. Team3 had by far the highest precision for Landscape at 0.5937, indicating a much higher proportion of its retrieved Landscape examples were relevant compared to the other teams.

Precision scores vary more widely between categories within each team compared to the average precision scores. This suggests the proportion of relevant examples retrieved differed more significantly between categories for each model.

In summary, while Team3's model achieved the highest average precision, indicating a higher overall proportion of relevant examples retrieved, the different teams demonstrated large variations in precision for the specific target categories. This underscores the importance of evaluating precision separately by category in addition to average precision.

Table 2
Precision for each team in the thematic unsupervised track at Rest-Mex 2023.

Run	Avg Precision	Precision (Insecurity)	Precision (Prices)	Precision (Gastronomy)	Precision (Landscape)
Team1	0.4141	**0.4508**	0.4725	**0.4842**	0.2488
Team2	0.3648	0.3520	0.4652	0.3742	0.2676
Team3	**0.4208**	0.3534	**0.6811**	0.0550	**0.5937**
Team4	0.2968	0.3583	0.2340	0.2320	0.3627

Recall results

Table 3 shows the results of the Recall metric. Run Avg Recall is the average recall score across all target categories for that team's model run. Higher is better, indicating the model retrieved more relevant examples on average. Recall for each of the target categories (Insecurity, Prices, Gastronomy, Landscape) shows the recall score specifically for that category. This indicates how well the model retrieved relevant examples for that category.

Table 3
Recall for each team in the thematic unsupervised track at Rest-Mex 2023.

Run	Avg Recall	Recall (Insecurity)	Recall (Prices)	Recall (Gastronomy)	Recall (Landscape)
Team1	**0.3250**	**0.9686**	0.1547	**0.1367**	0.0402
Team2	0.3047	0.9309	**0.1625**	0.0864	0.0389
Team3	0.3033	0.9592	0.0705	0.0910	**0.0923**
Team4	0.2649	0.8637	0.0767	0.0794	0.0398

Team1 had the highest average recall at 0.3250, indicating it performed best overall at retrieving relevant examples. Team1 and Team3 had the highest recall for the Insecurity category, around 0.96 - 0.95, meaning they performed best at retrieving relevant examples related to insecurity. Team2 had the highest recall for the Prices category at 0.1625, performing best for that target. Team3 had the highest recall for the Gastronomy category at 0.0923, performing best for retrieving gastronomy-related examples. The recall scores across teams and categories vary, indicating different levels of performance for different target categories. This is common and suggests the models struggle more with some categories than others.

In summary, these results show that while Team1 had the best average recall overall, the different teams demonstrated varying levels of performance for the specific target categories, with different teams performing best for different targets. This again underscores the importance of evaluating performance separately by category in addition to average recall.

Legend for Table 1 and Table 2:

- Team1: Javilonso-Team_javier_alonso_thematic_spacy_kmeans10_5
- Team2: CIMAT-Team_run3_thematic
- Team3: JCMQ-Team_run_5
- Team4: MCE-Team_2ndIterKmeans

So in conclusion, looking at the average scores alone can give a high-level picture of overall performance, but breaking the results down by specific category reveals more nuanced differences in how well the models performed for particular targets. Both types of analysis provide useful insight.

4. Conclusions and further work

In this paper, we presented a general procedure to resolve the task related to the thematic unsupervised classification of Mexican tourist texts in the context of Rest-Mex 2023. The proposed tasks provide valuable insights for the analysis of tourist texts in Spanish, which can be useful for the improvement of tourist services and destinations. Our proposal is based on the use of TF-IDF, LSA and K-Means algorithms and the result shows the promissory application of these algorithms in unsupervised classification of text problems. Future work can explore the use of deep learning models for these tasks and the integration of other features, such as demographic and contextual information.

5. References

[1] Álvarez-Carmona, M. Á., Díaz-Pacheco, Á., Aranda, R., Rodríguez-González, AY, Bustio-Martínez, L., Muñis-Sánchez, V., Sánchez-Vega, F. (2023). Overview of Rest-Mex at IberLEF 2023: Research on Sentiment Analysis Task for Mexican Tourist Texts. Procesamiento Del Lenguaje Natural, 71.

[2] Nakamura, T., Shirakawa, M., Hara, T., & Nishio, S. (2018). Wikipedia-based relatedness measurements for multilingual short text clustering. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, *18*(2), 1-25.

[3] Núñez-Reyes, A., Villatoro-Tello, E., Ramírez-de-la-Rosa, G., & Sánchez-Sánchez, C. (2017). A compact representation for cross-domain short text clustering. In *Advances in Computational Intelligence: 15th Mexican International Conference on Artificial Intelligence, MICAI 2016, Cancún, Mexico, October 23–28, 2016, Proceedings, Part I 15* (pp. 16-26). Springer International Publishing.

[4] Ahmed, M., Seraj, R., & Islam, S. M. S. (2020). The K-Means algorithm: A comprehensive survey and performance evaluation. *Electronics*, *9*(8), 1295.

[5] Sinaga, K. P., & Yang, M. S. (2020). Unsupervised K-Means clustering algorithm. *IEEE access*, *8*, 80716-80727.

[6] Rios, R. A., Nogueira, T., Coimbra, D. B., Lopes, T. J., Abraham, A., & Mello, R. F. D. (2021). Country transition index based on hierarchical clustering to predict next COVID-19 waves. *Scientific reports*, *11*(1), 15271.

[7] Duan, L., Aggarwal, C., Ma, S., & Sathe, S. (2019, November). Improving spectral clustering with deep embedding and cluster estimation. In *2019 IEEE International Conference on Data Mining (ICDM)* (pp. 170-179). IEEE.

[8] de Campos, L. M., Fernández-Luna, J. M., Huete, J. F., & Redondo-Expósito, L. (2020). Automatic construction of multi-faceted user profiles using text clustering and its application to expert recommendation and filtering problems. *Knowledge-Based Systems*, *190*, 105337.

[9] Li, Q., Li, S., Zhang, S., Hu, J., & Hu, J. (2019). A review of text corpus-based tourism big data mining. *Applied Sciences*, *9*(16), 3300.

[10] CHEN, Y., & ZHAO, X. (2022). Varied density clustering algorithm based on border point detection. *Journal of Computer Applications*, *42*(8), 2450.

[11] de Souto, M. C., Coelho, A. L., Faceli, K., Sakata, T. C., Bonadia, V., & Costa, I. G. (2012, October). A comparison of external clustering evaluation indices in the context of imbalanced data sets. In *2012 Brazilian Symposium on Neural Networks* (pp. 49-54). IEEE.

[12] Wagner, S., & Wagner, D. (2007). *Comparing clusterings: an overview* (pp. 1-19). Karlsruhe: Universität Karlsruhe, Fakultät für Informatik.

[13] Orellana, G., Arias, B., Orellana, M., Saquicela, V., Baculima, F., & Piedra, N. (2018, November). A study on the impact of pre-processing techniques in Spanish and english text classification over short and large text documents. In *2018 international conference on information systems and computer science (INCISCOS)* (pp. 277-283). IEEE.

[14] Orellana, G., Arias, B., Orellana, M., Saquicela, V., Baculima, F., & Piedra, N. (2018, November). A study on the impact of pre-processing techniques in Spanish and english text classification over short and large text documents. In *2018 international conference on information systems and computer science (INCISCOS)* (pp. 277-283). IEEE.

[15] Dumais, S. T. (2004). Latent semantic analysis. *Annu. Rev. Inf. Sci. Technol.*, *38*(1), 188-230.

[16] Álvarez-Carmona, M. A., Aranda, R., Rodríguez-González, A. Y., Pellegrin, L., & Carlos, H. (2022). Classifying the mexican epidemiological semaphore colour from the covid-19 text spanish news. *Journal of Information Science*, 01655515221100952.

[17] Kumbhar, R., Mhamane, S., Patil, H., Patil, S., & Kale, S. (2020, June). Text document clustering using K-Means algorithm with dimension reduction techniques. In *2020 5th International Conference on Communication and Electronics Systems (ICCES)* (pp. 1222-1228). IEEE.

[18] A. Diaz-Pacheco, M. Á. 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 (2022) 1–31.

[19] M. A. Alvarez-Carmona, R. Aranda, A. Rodriguez-Gonzalez, D. Fajardo-Delgado, M. G. A. Sanchez, H. Perez-Espinosa, J. Martinez-Miranda, R. Guerrero-Rodriguez, L. Bustio-Martinez, A. D. Pacheco, Natural language processing applied to tourism research: A systematic review and future research directions, Journal of King Saud University-Computer and Information Sciences (2022).

[20] E. Olmos-Martínez, M. Á. Álvarez-Carmona, R. Aranda, A. Díaz-Pacheco, What does the media tell us about a destination? the cancun case, seen from the usa, canada, and mexico, International Journal of Tourism Cities (2023).

[21] M. Á. Álvarez-Carmona, R. Aranda, S. Arce-Cardenas, D. Fajardo-Delgado, R. Guerrero-Rodríguez, A. P. López-Monroy, J. Martínez-Miranda, H. Pérez-Espinosa, A. Y. Rodríguez-González, Overview of rest-mex at iberlef 2021: recommendation system for text mexican tourism, Procesamiento del Lenguaje Natural 67 (2021).

[22] 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).

[23] M. Á. Álvarez-Carmona, R. Aranda, A. Y. Rodríguez-González, L. Pellegrin, C. Hugo, Classifying the mexican epidemiological semaphore colour from the covid-19 text spanish news, Journal of Information Science (2022).