

An Ensemble Based Clustering Approach to Group Mexican News

Jorge Ramos-Zavaleta^{1,2,*}, Adrian Rodríguez^{1,2}, Lizbeth Rodríguez^{1,2} and Jean Arreola^{1,2}

¹Monterrey Institute of Technology and Higher Education, México

²Center for Research in Mathematics, México

Abstract

This paper presents the findings and insights gained through our participation in the Rest-Mex Thematic Unsupervised Classification task. We discuss an ensemble-based clustering approach for grouping Mexican news related to tourism themes. The primary objective is to identify representative clusters within the dataset. To accomplish this objective, we implemented clustering algorithms (K-means and OPTICS) on a reduced dataset using the UMAP technique. Our aim was to enhance the accuracy of the base clusters by identifying data points that were not originally captured by the base clustering method. This was achieved by incorporating a supplementary layer of K-means clusterings on top of the initial base cluster results. The results showed that the ensemble clustering approach improved the accuracy and macro F1-measure of the base clustering method. Overall, the paper presents a promising approach to clustering Mexican news related to tourism topics using unsupervised learning and ensemble techniques.

Keywords

Natural Language Processing, Clustering, Ensemble, Mexican news

1. Introduction

While machine learning and artificial intelligence have been considered in past editions of REST-MEX [1, 2], the bulk of such tasks have focused on supervised learning.

Applications of AI in Tourism have been very limited due to not having as wide a variety of labeled data sets compared to other topics. Unsupervised learning can unconstrain us from the need for labeled data and manual handcrafted feature engineering, thereby facilitating exploit sources of information (news, reviews of places, blogs, among others) and the use of methods of machine learning[3, 4].

In this work, we present a systematic approach that involves the utilization of a Doc2Vec embedding technique to convert textual data into vector representations. Following this, we employ UMAP representation to enhance computational efficiency when working with base cluster models. These UMAP representations are then passed to a K-means layer, thereby aiming to enhance the accuracy of the base clusters.


IberLEF 2023, September 2023, Jaén, Spain

*Corresponding author.

✉ jorge.ramos@cimat.mx (J. Ramos-Zavaleta)



© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

	lenNews	tokens
count	114,550	114,550
mean	4,391.67	650.16
std	3,652.81	561.37
min	54	1
25%	2,340.00	336
50%	3,457.00	506
75%	5,366.00	798
max	32,767	5,858

Table 1
Basic statistics of the data.

2. The Data

The corpus consists of 114,550 news obtained from google news. The data were collected over the last two years references to 4 tourism themes. Even when the News were carefully tagged, as it is an unsupervised classification task, the data was released in a single unlabeled set, meaning the themes could not be identified.

Each row contains a column including at first the webpage where the news were collected and the text of the news.

In the table 1 we present some basic statistics about Tokens/words and number of characters in the News of the dataset. This shows that there exist a lot of variance between the format and the content of the diferent news.

By applying an exploration on data, we find that more than half of the News provided have more than 512 tokens, this implies that it is impossible to think on use BERT embeddings directly. Furthermore, 572 News has only one token, and in many cases, we only found an URL without the body of the article.

3. Method

A similar dataset but applied to a different task can be found in [5] and a similar cleansing was performed in this dataset. For the contest, we considered different alternatives for clustering methods but usually, most of the clustering algorithms have some issues like not being scalable to large datasets and not having enough capacity to catch the complexity of the data relationships in large dimensions. In Figure [6], we can find a table with some clustering methods and their capability to scale to large datasets.

In Figure 1 we can see that only a few of these algorithms are really capable to deal with large datasets. These algorithms can be separated into distance-based methods (like K-means) or density based (like DBSCAN or OPTICS). In this sense, the former is usually restricted to a flat geometry of the data and the latter have the issue that they can be very sensitive to their

parameters [7].

Apart from the issues concerned to the selection of the right clustering method, because of the nature of the data we need to create a vectorial representation of the data. For the sake of simplicity we consider a Word2Vec approach[8] so we can perform different trials even for the dense-based cluster that are computationally intensive compared with a simpler model like k-means that has a really fast execution.

For our approach, we consider stacking clustering methods in a way that can be considered as an ensemble clustering. In this way, our approach allows it to improve the results of a base clustering labeling by finding points that were not considered by the base clustering method and attaching them to the base clustering results.

Method name	Parameters	Scalability	Usecase	Geometry (metric used)
K-Means	number of clusters	Very large $n_{samples}$, medium $n_{clusters}$ with MiniBatch code	General-purpose, even cluster size, flat geometry, not too many clusters, inductive	Distances between points
Affinity propagation	damping, sample preference	Not scalable with $n_{samples}$	Many clusters, uneven cluster size, non-flat geometry, inductive	Graph distance (e.g. nearest-neighbor graph)
Mean-shift	bandwidth	Not scalable with $n_{samples}$	Many clusters, uneven cluster size, non-flat geometry, inductive	Distances between points
Spectral clustering	number of clusters	Medium $n_{samples}$, small $n_{clusters}$	Few clusters, even cluster size, non-flat geometry, transductive	Graph distance (e.g. nearest-neighbor graph)
Ward hierarchical clustering	number of clusters or distance threshold	Large $n_{samples}$ and $n_{clusters}$	Many clusters, possibly connectivity constraints, transductive	Distances between points
Agglomerative clustering	number of clusters or distance threshold, linkage type, distance	Large $n_{samples}$ and $n_{clusters}$	Many clusters, possibly connectivity constraints, non Euclidean distances, transductive	Any pairwise distance
DBSCAN	neighborhood size	Very large $n_{samples}$, medium $n_{clusters}$	Non-flat geometry, uneven cluster sizes, outlier removal, transductive	Distances between nearest points
OPTICS	minimum cluster membership	Very large $n_{samples}$, large $n_{clusters}$	Non-flat geometry, uneven cluster sizes, variable cluster density, outlier removal, transductive	Distances between points
Gaussian mixtures	many	Not scalable	Flat geometry, good for density estimation, inductive	Mahalanobis distances to centers
BIRCH	branching factor, threshold, optional global clusterer.	Large $n_{clusters}$ and $n_{samples}$	Large dataset, outlier removal, data reduction, inductive	Euclidean distance between points
Bisecting K-Means	number of clusters	Very large $n_{samples}$, medium $n_{clusters}$	General-purpose, even cluster size, flat geometry, no empty clusters, inductive, hierarchical	Distances between points

Figure 1: This table provides some details for different clustering algorithms, their scalability and the possible usecase for each.

3.1. UMAP

UMAP (Uniform Manifold Approximation and Projection) algorithm [9] is a dimensionality reduction technique used for data visualization and clustering tasks. Unlike traditional methods such as PCA and t-SNE, UMAP preserves both the local and global structure of the data, making it highly effective for revealing meaningful patterns and relationships in high-dimensional datasets.

The UMAP algorithm can be divided into several key steps:

- **Constructing the Fuzzy Topological Structure:** The algorithm identifies the nearest neighbors for each data point, considering their distances and a user-defined parameter called the `n_neighbors`. Fuzzy connectivities are computed by comparing the distances between data points, giving higher weights to closer neighbors. The connectivities are then transformed into a fuzzy simplicial set.
- **Optimizing the Low-Dimensional Embedding:** The algorithm initializes random low-dimensional embeddings for the data points. It employs stochastic gradient descent to optimize the embeddings, minimizing the discrepancy between the fuzzy topological structure in the high-dimensional space and the low-dimensional embedding.
- **Handling Different Data Types and Distance Metrics:** For numerical data, UMAP employs the Euclidean distance metric by default. For categorical data, UMAP uses a distance metric called the "Gower distance," which considers the dissimilarity between categorical values.
- **Controlling Parameters.** Key parameters include the number of neighbors (`n_neighbors`), the target dimensionality of the embedding (`n_components`), and the learning rate for stochastic gradient descent.

UMAP has gained popularity due to its ability to generate high-quality visualizations that faithfully represent the complex relationships in high-dimensional data. It excels in preserving local and global structures, allowing for more comprehensive data exploration. Furthermore, UMAP is computationally efficient, making it suitable for large-scale datasets.

Overall, UMAP has emerged as a state-of-the-art technique for dimensionality reduction, providing a powerful tool for visualizing complex datasets and extracting meaningful insights and has been widely applied in several applications in different fields like in genomics [10] or sociology [11].

3.2. OPTICS

OPTICS (Ordering Points To Identify the Clustering Structure) is a density-based clustering algorithm that was introduced by [12]. It is similar to DBSCAN, but it has a number of advantages, including:

- It can identify clusters of varying densities
- It can identify clusters of arbitrary shapes.
- It is more scalable than DBSCAN.

OPTICS works by first ordering the data points according to their density. Points that are most densely connected are placed at the beginning of the ordering, and points that are least

densely connected are placed at the end. Once the data points have been ordered, OPTICS identifies clusters by finding connected sequences of points that have a high density.

OPTICS has been shown to be effective for a variety of clustering tasks . It has been used to cluster text documents, gene expression data, and social network data, some use cases can be found in [13] and [14].

3.3. K-means

The K-Means algorithm [15] is one of the most used unsupervised machine learning algorithms for clustering. It aims to partition a dataset into K distinct clusters based on similarity. The algorithm iteratively assigns data points to the nearest centroid and updates the centroids by computing the mean position of the assigned points. This process continues until convergence is reached.

One key aspect of the K-Means algorithm is the initialization of centroids. Proper initialization is crucial as it can impact the convergence speed and the quality of the resulting clusters. The choice of distance metric is another critical consideration in K-Means. The most commonly used distance metric is the Euclidean distance, however alternative metrics like Manhattan distance or cosine similarity can be employed depending on the characteristics of the data.

K-Means is known for its simplicity and efficiency, making it suitable for large-scale datasets. However, it is sensitive to the initial configuration and may converge to local optima.

Overall, the K-Means algorithm provides a practical and effective means of clustering data, with applications in various domains such as image segmentation, customer segmentation, and anomaly detection.

3.4. Ensemble Clustering

Ensemble Clustering has become an active area for research in recent years because of the success of the supervised learning methods that applied the ensemble techniques. Some works like the one on Nguyen [16] or the one from Topchy [17] can give a good introduction to how these methods work.

In particular, in [18] a survey about the methods and the problems that ensemble clusterings can face is widely presented. The basic idea of ensemble clusterings is the same that the one used for the supervised learning ensemble methods, but because of the nature of the problem, some measures have to be taken so the different models can be comparable. In figure 2 can be seen that the basic structure of the ensemble clustering method resembles completely to the ensembles in the supervised learning setting.

The main problem with these methods is determining how to compare the groups that are assigned since the labels or the centers can vary from one model to another. To solve this issue

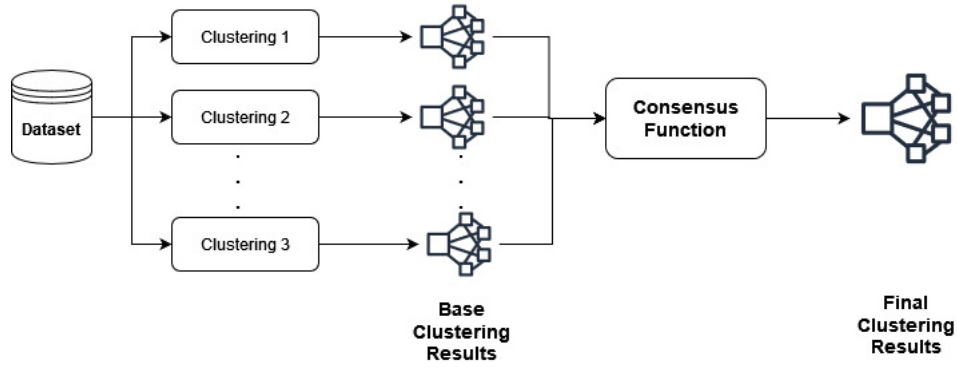


Figure 2: The basic ensemble clustering follows the same idea that the supervised learning ensembles.

it is possible to consider 2 solutions: Relabeling the results, usually by creating a similarity matrix indicating when two points fall in the same cluster. On the other side, another possible approach is to modify the consensus function so it considers different possibilities as the looking for a median partition that fits most of the data partitions of the different models, this specific method can be found in [17].

The goal of this type of ensemble methods is to get the best base clustering, however, we were looking for how we could improve the results of a given base clustering method. This allows us to solve the problem in a simpler way than other ensemble clusterings and also could allow us to use our approach with other ensemble clustering methods.

In our approach, the base clustering results are used for a layer of clustering methods as labels so the relabeling is based in a density sense. In this way, the layer of clustering methods is considered with a larger number of clusters to be found so it can allow to find relationship between points that the base cluster did not consider to belong to the same cluster. Once the results of each method of the layer have been calculated, a simple consensus voting is generated to improve the accuracy of the base cluster by finding points that really belong to a cluster but were not considered in the original base cluster.

An important characteristic of our approach is that the results of the consensus layer can be passed to a new layer of models that considers a larger number of clusters than the past layer and keeps improving the classification of the points from the original clusters.

3.5. Ensemble Pipeline

For this work, we consider a complete pipeline to test our approach with two different base cluster methods (K-means and OPTICS). In Figure 3 is displayed the basic pipeline that was used.

We have carried out a simple process of cleansing the database, which mainly consists of

removing content that does not correspond to the news per se, and in those cases where the content is irrelevant to the news, we replace the text with an extract of the text present in the news URL, this part usually contains relevant information of the evaluated text.

Once we have the data prepared, we vectorize the texts using a Doc2Vec embedding to vectorize the data and later we perform a UMAP representation, which allowed us a faster computation with the base cluster models. With the UMAP representation, the base cluster is initiated and the results are stored and passed to a K-means layer to evaluate them. We did a test with one layer of k-means with 30 clusters and another one with two layers of k-means, with 30 and 80 clusters respectively.

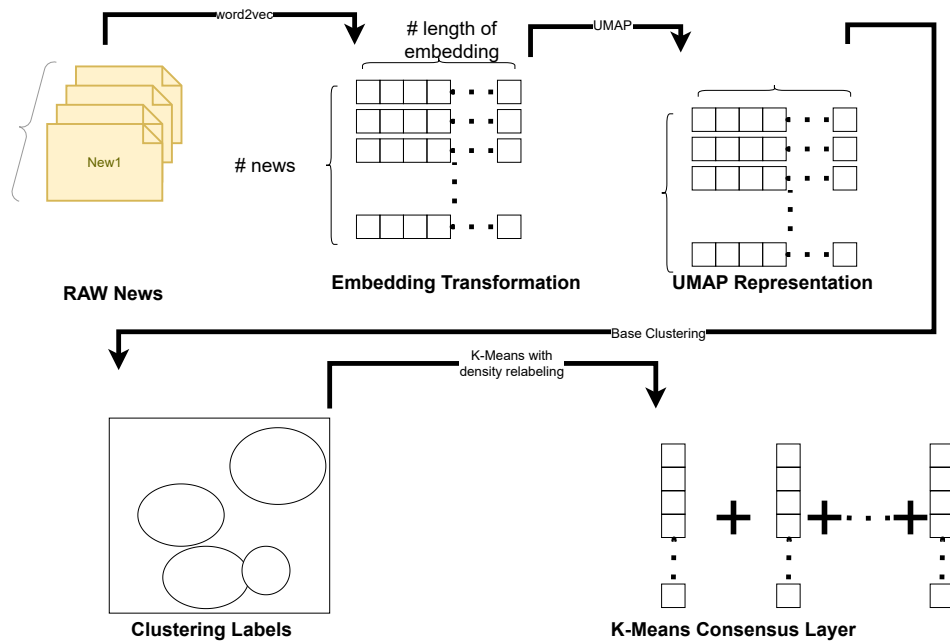


Figure 3: Basic pipeline used to perform the method for the competition.

4. Results and Discussion

We introduced a method to improve the performance of a base cluster by stacking a k-means layer with more segments than the base cluster. This method allows us to find some smaller structures that the base cluster does not consider, thus seeking to improve the segmentation accuracy.

Table 2 presents the accuracy and Global F1-measure for six cases, 3 for each base cluster, including the results of adding a first layer of k-means and the results of adding a second layer. In Figure 4 UMAP representation of the base clusters results can be seen, is shown that both

	Base Cluster		1st layer		2nd layer	
Base Cluster	Accuracy	F1 Measure	Accuracy	F1 Measure	Accuracy	F1 Measure
Kmeans	33.34	0.2	33.77	0.201	34.30	0.203
OPTICS	11.39	0.098	24.067	0.154	24.074	0.154

Table 2

Accuracy and Macro F1-Measure for the differenten trials. Can be seen that each time that a layer of k-means is used the results is improved for both the accuracy and Macro F1-Measure for both base clusters.

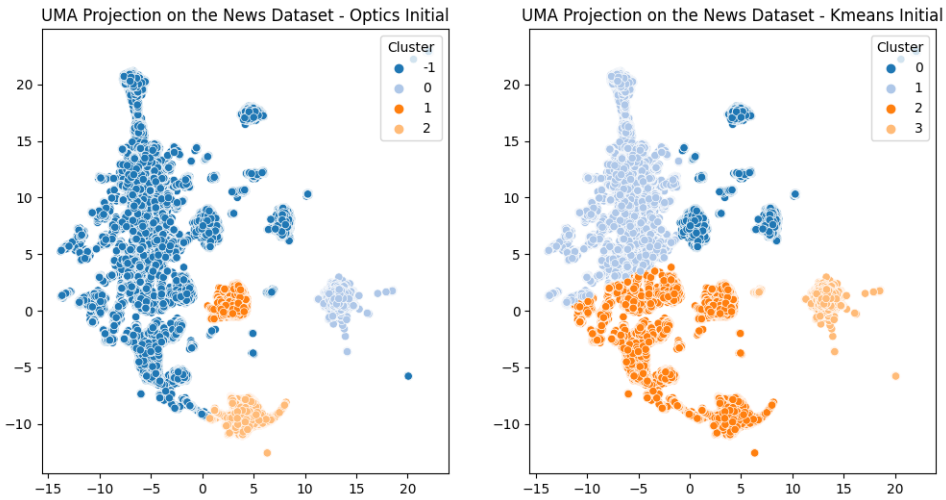


Figure 4: UMAP representation of the base clusters results.

base cluster presents different sizes in the clusters.

In general, there is an improvement in evaluation metrics as we add layers to the base cluster. In the OPTICS clustering method, we have the highest rate of accuracy growth as we add layers. However, the best accuracy was obtained by adding new layers of the k-means method. This suggests that the idea of incorporating layers to a base method can help to improve the segmentation of the groups, but that this improvement depends on the selected method.

As it is an unsupervised task, it is difficult to compare the results of the assigned labels versus the original labels to determine some steps to improve the performance of our solution. However, still there is room for what can be done in order to improve the results: to begin with, we could consider a different embedding representation like BERT, we also can drop the UMAP representation and run the clusters directly on the data this would allow us to not multiply the error of the data that is already carried by the embedding representation.

Other improvements can be carried out, like considering samplings of the data for the k-means

layers to avoid the troubles of possible outliers and also to do a systematic review about the number of clusters that has to be considered for each layer like said before a review of the different methods for the base cluster is needed to verify if a different method that the ones in the layers are used our method could provide a better improvement than considering the same algorithm for the base cluster and the layers.

5. Conclusions

We present in this work an ensemble-based clustering approach for grouping Mexican news related to tourism themes by using unsupervised learning and an ensemble approach. The primary objective is to identify representative clusters within the dataset, for this our approach involves utilizing a Doc2Vec embedding technique to convert textual data into vector representations, we also employed an UMAP representation to enhance computational efficiency when working with base cluster models, and finally we incorporate at least one supplementary layer of K-means clusterings on top of the initial base cluster results that allows to improve accuracy. The results showed that the ensemble clustering approach improved the accuracy and macro F1-measure of the base clustering method.

Overall, this paper presents a systematic approach that involves the utilization of different base clustering algorithms (K-means and OPTICS) on a reduced dataset using the UMAP technique. The proposed ensemble-based clustering approach enhances the accuracy of the base clusters by identifying data points that were not originally captured by the base clustering method by using the base cluster results to relabel the results of several kmeans with a density based approach and updating the results with a majority consensus function.

Even when for this specific work we consider an UMAP representation if not necessary is the method wants to be applied to other datasets and only was considered to reduce the computational time for this work. Some further research can be carried out to improve the results for this method and this particular dataset, including considering different embedding representations like BERT and doing a systematic review about the number of clusters that has to be considered for each new layer of kmeans.

References

- [1] 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).
- [2] 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).

- [3] 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>.
- [4] 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. URL: <https://doi.org/10.1080/0952813X.2022.2153276>. doi:10.1080/0952813X.2022.2153276. arXiv:<https://doi.org/10.1080/0952813X.2022.2153276>.
- [5] J. Ramos-Zavaleta, A. Rodríguez, A mexico's covid traffic light color prediction system based on mexican news, *Proceedings of the Fourth Workshop for Iberian Languages Evaluation Forum*. (2022).
- [6] Clustering, <https://scikit-learn.org/stable/modules/clustering.html>, Accessed: 2023-05-25.
- [7] N. Ohadi, A. Kamandi, M. Shabankhah, S. M. Fatemi, S. M. Hosseini, A. Mahmoudi, Sw-dbscan: A grid-based dbscan algorithm for large datasets, in: *2020 6th International Conference on Web Research (ICWR)*, IEEE, 2020, pp. 139–145.
- [8] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient estimation of word representations in vector space, *arXiv preprint arXiv:1301.3781* (2013).
- [9] L. McInnes, J. Healy, N. Saul, L. Grossberger, Umap: Uniform manifold approximation and projection, *Journal of Open Source Software* 3 (2018) 861. doi:10.21105/joss.00861.
- [10] J. S. Packer, Q. Zhu, C. Huynh, P. Sivaramakrishnan, E. Preston, H. Dueck, D. Stefanik, K. Tan, C. Trapnell, J. Kim, et al., A lineage-resolved molecular atlas of *c. elegans* embryogenesis at single-cell resolution, *Science* 365 (2019) eaax1971.
- [11] C. Lin, C. Griffith, K. Zhu, V. Mathur, Understanding vulnerability of children in surrey (2018).
- [12] M. Ankerst, M. M. Breunig, H.-P. Kriegel, J. Sander, Optics: Ordering points to identify the clustering structure, *ACM Sigmod record* 28 (1999) 49–60.
- [13] S. Babichev, B. Durnyak, V. Zhydetskyy, I. Pikh, V. Senkivskyy, Application of optics density-based clustering algorithm using inductive methods of complex system analysis, in: *2019 IEEE 14th International Conference on Computer Sciences and Information Technologies (CSIT)*, volume 1, IEEE, 2019, pp. 169–172.
- [14] B. Shen, Y.-S. Zhao, Optimization and application of optics algorithm on text clustering, *Journal of Convergence Information Technology* 8 (2013) 375.
- [15] J. B. MacQueen, A k-means clustering algorithm, in: *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, 1967, pp. 281–297. URL: <https://projecteuclid.org/euclid.bsm/1200512992>.
- [16] N. Nguyen, R. Caruana, Consensus clusterings, in: *Seventh IEEE international conference on data mining (ICDM 2007)*, IEEE, 2007, pp. 607–612.
- [17] A. Topchy, A. K. Jain, W. Punch, Clustering ensembles: Models of consensus and weak

partitions, *IEEE transactions on pattern analysis and machine intelligence* 27 (2005) 1866–1881.

- [18] T. Boongoen, N. Iam-On, Cluster ensembles: A survey of approaches with recent extensions and applications, *Computer Science Review* 28 (2018) 1–25.
- [19] 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).