Leveraging LDA Topic Modeling and BERT Embeddings for Thematic Unsupervised Classification of Tourism News in Rest-Mex Competition

Erika Rivadeneira-Pérez[1,*], Cipriano Callejas-Hernández[1]

[1]Mathematics Research Center (CIMAT), Guanajuato, Mexico.

Abstract

In this work we present a solution to the *Thematic Unsupervised Classification*, a track presented in this year Rest-Mex competition, in which four topic-based groups are to be found from a unlabeled text item news related to tourism in Mexico. Our approach includes LDA topic modeling on a term-document representation, as well as BERT embeddings. Our BERT-based approach achieved a second place among all participating teams in the competition, demonstrating the effectiveness of mixing pre-trained models with traditional machine learning techniques.

Keywords

Sentiment Analysis, Deep Learning, Transformer Models

1. Introduction

The tourism industry is highly complex industry, characterized by a multitude of service providers, intermediaries and customers. To provide personalized recommendations, it requires experiences based on the comprehensive understanding of needs and wishes of individual travelers. This facilitates decision-making and better planning, and it is part of what is called in the literature as "smart tourism" [7, 11, 9, 8, 10]. On the other hand, document classification is a standard task in Machine Learning, but the process usually relies on supervised or semi-supervised approaches. Particularly in tourism Machine Learning has been used for different purposes, such as market segmentation (clustering) in Hudson. Recently in 2020, Egger [6] identified travelers at two different points in time based on their perceived risk of COVID-19 during the pandemic, in order to segment the travelers as being anxious, nervous or reserved.

The pipeline across all our approaches that we shall detail below, consists of getting a numeric vector representation of each news item and then using common clustering techniques to group them. To this end, centroids are used as the estimated tourism-related topic or tags.

IberLEF 2023, September 2023, Jaén, Spain

[*]Corresponding author.

erika.rivadeneira@cimat.mx (E. Rivadeneira-Pérez); cipriano.callejas@cimat.mx (C. Callejas-Hernández)

CEUR Workshop Proceedings (CEUR-WS.org)

1.1. Unsupervised Text Classification

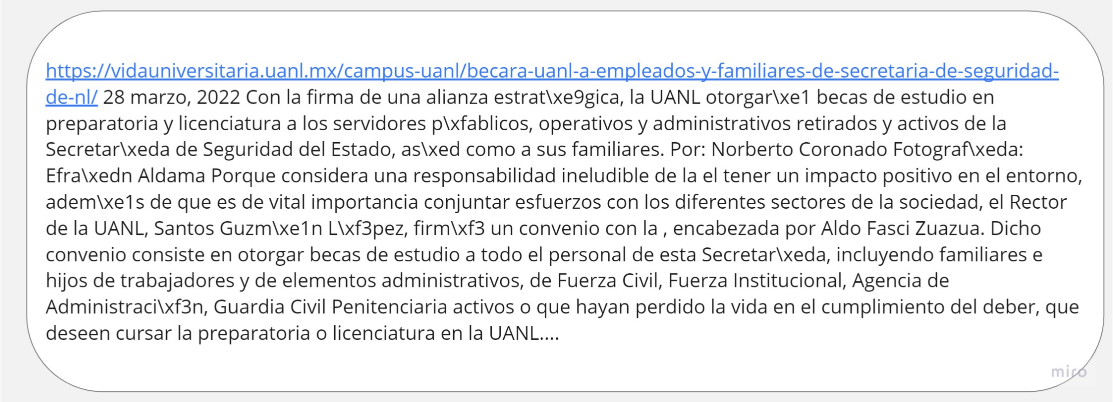
A very common method of unsupervised learning is clustering, which aims to identify distinct groups in data, that is, we seek to learn something about the structure and patterns inherited by the data. On other hand, unsupervised text classification aims to perform categorization without using annotated data during training and therefore offer the potential to reduce annotation costs. In this direction little research has been conducted on unsupervised text classification, see for instance the survey [2] . Moreover, Braun et al [2] show that similarity-based approaches are the most popular technique, however for our approaches we followed a segmentation-based one, where either we used clustering methods (or topic modeling) considering that the number of clusters (or topics) is known in advance.

2. Thematic Unsupervised Classification Track

For this task, 50,000 news items were collected on 4 different topics related to tourism. The challenge is to identify the groups in an automatic way. We call our approach a segmentation-based one, as detailed below.

2.1. Corpus Description

All data was obtained from google news over the last two years regarding 4 unrevealed touristic topics, downloaded and tagged. Figure 1 shows a news example of the corpus.



https://vidauniversitaria.uanl.mx/campus-uanl/becara-uanl-a-empleados-y-familiares-de-secretaria-de-seguridad-de-nl/ 28 marzo, 2022 Con la firma de una alianza estrat\xe9gica, la UANL otorgar\xe1 becas de estudio en preparatoria y licenciatura a los servidores p\xfablicos, operativos y administrativos retirados y activos de la Secretar\xeda de Seguridad del Estado, as\xed como a sus familiares. Por: Norberto Coronado Fotograf\xeda: Efra\xedn Aldama Porque considera una responsabilidad ineludible de la el tener un impacto positivo en el entorno, adem\xe1s de que es de vital importancia conjuntar esfuerzos con los diferentes sectores de la sociedad, el Rector de la UANL, Santos Guzm\xe1n L\xf3pez, firm\xf3 un convenio con la , encabezada por Aldo Fasci Zuazua. Dicho convenio consiste en otorgar becas de estudio a todo el personal de esta Secretar\xeda, incluyendo familiares e hijos de trabajadores y de elementos administrativos, de Fuerza Civil, Fuerza Institucional, Agencia de Administraci\xf3n, Guardia Civil Penitenciaria activos o que hayan perdido la vida en el cumplimiento del deber, que deseen cursar la preparatoria o licenciatura en la UANL....

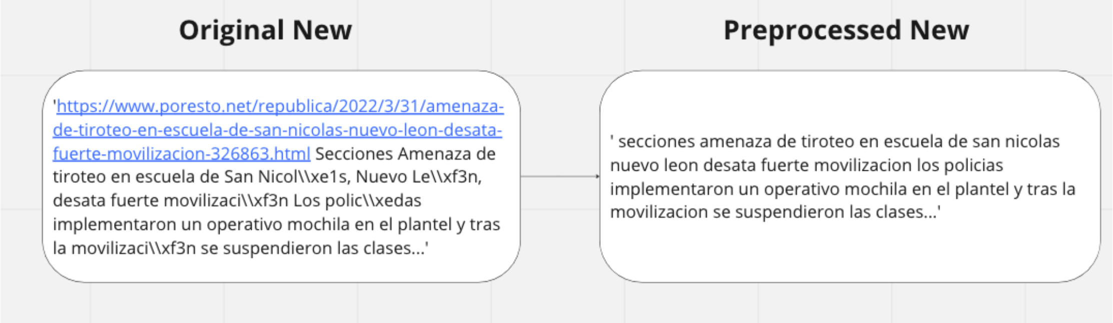
Figure 1: Example of an element in the dataset.

2.2. Data Preprocessing

As illustrated in Figure 2 we decided to homogenize the text in the following way:

- Removing hyperlinks.
- Removing numerical and special characters due to the large amount of numeric data that does not provide relevance to the topics, such as dates or monetary amounts, for example.

- Converting the entire text to lowercase.



'https://www.poresto.net/republica/2022/3/31/amenaza-de-tiroteo-en-escuela-de-san-nicolas-nuevo-leon-desata-fuerte-movilizacion-326863.html Secciones Amenaza de tiroteo en escuela de San Nicol\\xe1s, Nuevo Le\\xf3n, desata fuerte movilizaci\\xf3n Los polic\\xedas implementaron un operativo mochila en el plantel y tras la movilizaci\\xf3n se suspendieron las clases...'

' secciones amenaza de tiroteo en escuela de san nicolas nuevo leon desata fuerte movilizacion los policias implementaron un operativo mochila en el plantel y tras la movilizacion se suspendieron las clases...'

Figure 2: Example of a preprocessed new

2.3. Approach 1. Topic Modelling

The first proposal is to use of Latent Dirichlet Allocation (LDA), a widely used statistical model in NLP for unsupervised text clustering. LDA aims to discover the underlying topics in a set of text documents and group them into coherent categories [3]. This model is based on the premise that each document is a combination of various topics, and each topic is characterized by its distribution of words, this is known as a word frequency method. Our pipeline in this approach is the following:

- *Text preprocessing:* Text is preprocessed according to the specifications in section 2.2.
- *Text Vector Representation:* Text documents are map into a vector where each element represents the count of a specific word in the document.
- *Topic inference:* Topic inference involves calculating the probability of each document belonging to each topic based on the words it contains. In other words, topics are assigned to documents based on the probability of belonging to each of them. Table 1 shows the Top-10 most frequent words per Topic among all news.
- *Interpretation of results:* Finally, we analyze the assignment of topics to documents and understanding the discovered patterns.

Topic	**Palabras más frecuentes**
Topic 1	autoridades, fiscalia, noticias, leon, personas, elementos, seguridad, policia, dos, mas
Topic 2	parte, asi, ciudad, ser, si, notificaciones, tambien, hace, mexico, mas
Topic 3	lectura, anuncio, min, articulo, seguridad, foto, tiempo, mas, mexico, nacional
Topic 4	personas, millones, tambien, gobierno, mexico, guanajuato, yucatan, pesos, mil, mas

Table 1
Top Frequent Words per Topic among all the news

Note that in Table 1 frequent words overlap among different topics identified by LDA, which can make it challenging to differentiate them clearly. This overlap of words can be attributed to various factors, such as thematic similarity between topics or the presence of shared vocabulary. Hence, accurately discerning the boundaries and distinctive characteristics of each topic can be difficult. In our particular case, we have observed that the LDA model faces difficulties in accurately and efficiently differentiating more specific topics due to the intersections of words among the identified topics, see Figure 3 where the overlap between clusters is shown.

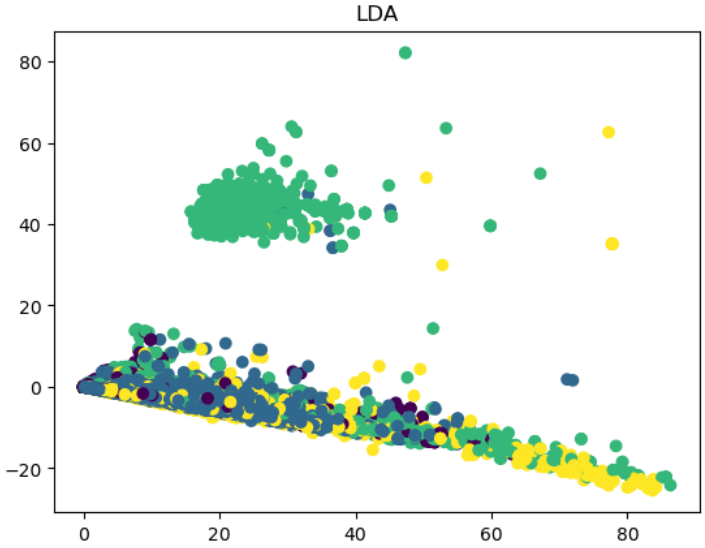


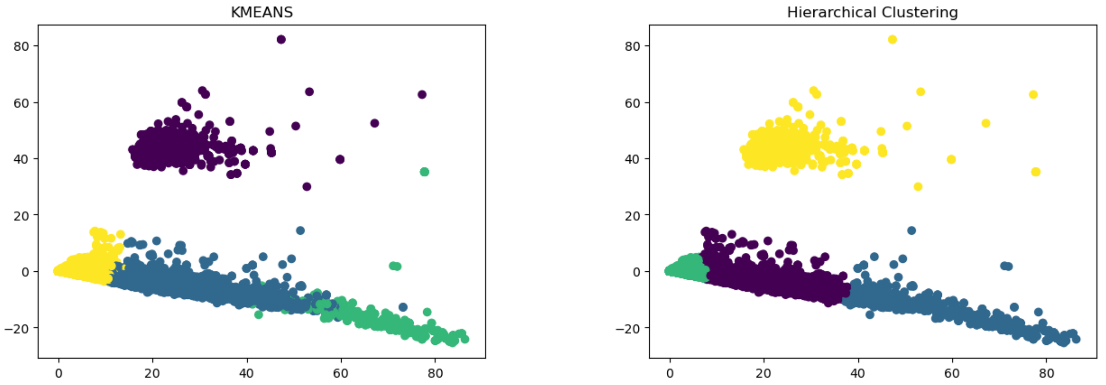
Figure 3: Corpus visualization and groups obtained through LDA topic modelling

2.4. Approach 2. Machine Learning Techniques

For the second approach, we considered classical clustering methods: k-means and hierarchical clustering, using the same word frequency representation of Section 2.3. The resulting topic groups of k-means and hierarchical clustering can be seen in Figure 4a and Figure 4b respectively. The figures show the two-dimensional SVD representation of the text and the corresponding topics to which each news belongs. We can observe that the observations of the obtained topics, for the most part, do not overlap with each other. However this traditional approach shows a more, at least visually speaking, clear segmentation of our text data.

2.5. Approach 3 ML techniques on Bert's corpus representation

Contextualized word and sentence embeddings produced by pre-trained models such as BERT have demonstrated the state-of-the-art results in NLP tasks, recently Selva et al. [15] showed that using BERT versus BOW as document representation surpass in topic coherence, moreover, the use of term frequency to select topic words fails to capture the semantics of clusters precisely because of words with high frequency may be common across different clusters, as seen in



(a) Visualization of corpus representation and clustering results obtained with K-means algorithm.

(b) Visualization of corpus representation and clustering results obtained with Hierarchical Clustering

Figure 4: ML unsupervised classification techniques on term frequency representation of the corpus

Figure 3. We believe that high quality document embeddings are critical for clustering-based topic modelling. In this approach, we used pre-trained model embeddings [16], and afterwards we used Kmeans clustering for topic modelling because of its efficiency shown in Section 2.4. As a consequence that most of the news are related to tourism, hence using the same language, as some overlap words previously mentioned.

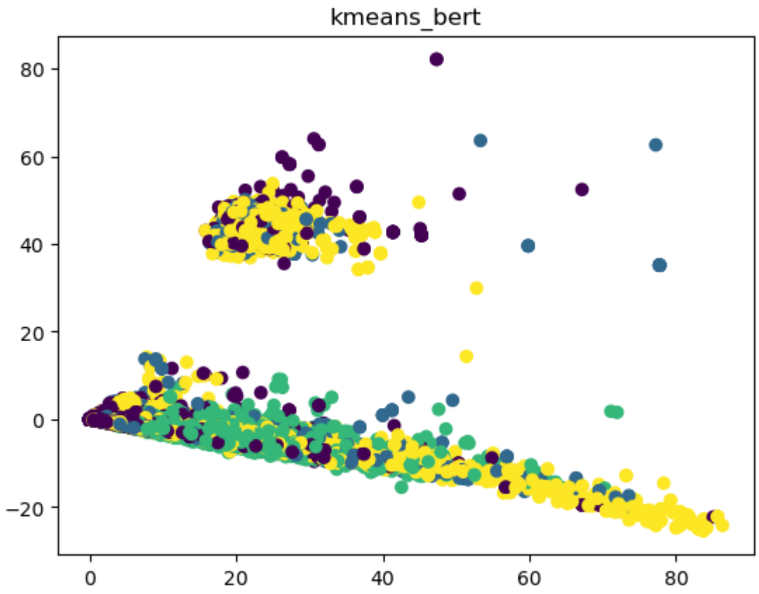


Figure 5: Visualization of SVD corpus representation and clustering results obtained with K-means algorithm on Bert's corpus representation

Table 2
Evaluation results of the thematic track runs. T-F represents Term-Frequency and HC the hierarchical clustering method.

Run	Approach	Macro F1	Accuracy	Avg Precision	Avg Recall
Run 3	BERT + KMeans (Sec. 2.5)	**0.2400**	35.6289	**0.3648**	**0.3047**
Run 4	T-F BOW + HC (Sec. 2.4)	0.2385	**63.0179**	0.3102	0.2554
Run 1	LDA + KMeans (Sec. 2.3)	0.1933	30.1920	0.3541	0.2915
Run 2	T-F BOW + KMeans (Sec.2.4)	0.0836	9.5862	0.2871	0.2888

3. Results

3.1. Evaluation Metrics

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 identify 4 groups with the news shared in the competition [1]. This means that the labels do not necessarily coincide for the same groups expected in the Gold Standard. For this reason, a re-labeling 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 Equation 1.

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

Table 2 shows each run and the approach used. We can see that the best approach, with respecto to the Macro F1 metric, is to use ML techniques on a BERT representation of the corpus, while word frequency representation and Hierarchical Clustering showed a pretty good approximation. Finally, both the LDA and term-frequency BOW and Kmeans showed a poorly behaviour.

In Table 3, the ranking results among the participating teams are displayed where we achieve the second place with the performance of our run 3.

Table 3
Thematic Evaluation Track Ranking

Ranking	Run	Macro F1	Accuracy	Avg Precision	Avg Recall
1st	Javilonso-Team	0.2827	44.8171	0.4141	0.3251
2nd	**CIMAT-Team_run3**	**0.2400**	**35.6289**	**0.3648**	**0.3047**
3rd	JCMQ-Team_run_5	0.2182	35.2766	0.4209	0.3033
HM	MCE-Team_2ndIterKmeans	0.2031	34.3029	0.2968	0.2649

4. Conclusions

As showed in Figures 4 and 4a, a term-frequency representation plus traditional clustering methods achieves a good segmentation, at least visually speaking. However as noted in Table 2, pre-trained contextualized embeddings surpass the simple term-frequency representation nonetheless in Figure 5 this segmentation is not clear enough, at least visually. We hypothesize that this might be related to the overlap between news items in terms of tourism-related keywords, such as COVID, since this created a big change in the market. We believe that more features could have been included in this approach, such is the case of better handling the scope of bert-embeddings, straightforward we passed news items through the pre-trained model all-MiniLM-L6-v2 sentence transformer [16] cutting of up to 128 items, resulting in a 384 dimensional dense vector space. Because of the little known results on this topic and being this the first time of this track in the Rest-Mex 2023 competition we wanted to try something simpler.

5. Acknowledgments

The authors thank Dr. Johan Johan VanHorebeek from Mathematics Research Center (CIMAT) for his support and guidance in creating this project.

References

[1] 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).

[2] Schopf, T., Braun, D., & Matthes, F. (2022). Evaluating unsupervised text classification: zero-shot and similarity-based approaches. arXiv preprint arXiv:2211.16285.

[3] Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. Journal of machine Learning research, 3(Jan), 993-1022.

[4] Author, F., Author, S., Author, T.: Book title. 2nd edn. Publisher, Location (1999)

[5] Aschauer, W., & Egger, R. (2023). Transformations in tourism following COVID-19? A longitudinal study on the perceptions of tourists. Journal of Tourism Futures.

[6] Hudson, Simon, and Brent Ritchie. "Understanding the domestic market using cluster analysis: A case study of the marketing efforts of Travel Alberta." Journal of Vacation Marketing 8.3 (2002): 263-276.

[7] Egger, Roman. Applied Data Science in Tourism: Interdisciplinary Approaches, Methodologies, and Applications. Zeitschrift für Tourismuswissenschaft, 2021

[8] 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.

[9] 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).

[10] 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)

[11] 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).

[12] Author, A.-B.: Contribution title. In: 9th International Proceedings on Proceedings, pp. 1–2. Publisher, Location (2010)

[13] Haj-Yahia, Z., Sieg, A., & Deleris, L. A. (2019, July). Towards unsupervised text classification leveraging experts and word embeddings. In Proceedings of the 57th annual meeting of the Association for Computational Linguistics (pp. 371-379).

[14] LNCS Homepage, http://www.springer.com/lncs. Last accessed 4 Oct 2017

[15] Selva Birunda, S., & Kanniga Devi, R. (2021). A review on word embedding techniques for text classification. Innovative Data Communication Technologies and Application: Proceedings of ICIDCA 2020, 267-281.

[16] Reimers, N., & Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. arXiv preprint arXiv:1908.10084.