

Seeking Clustering Excellence: Unleashing the Power of Sentence Transformers and Preprocessing Techniques

Javier Alonso-Mencía^{1,*}

¹University Carlos III of Madrid (UC3M), Madrid, Spain

Abstract

The article presents the participation of the Javier Alonso-Mencía team in the REST-MEX@IberLef 2023 text clustering competition. The solution proposed by the Javier Alonso-Mencía team is based on clustering with K-means and Gaussian Mixture using representations of texts obtained through Sentence Transformer (multilingual). Prior to the clustering, several transformations are applied, including the elimination of stopwords and punctuation, as well as lemmatization of words. After obtaining vectors from the Sentence Transformer, dimensionality reduction is performed using Uniform Manifold Approximation and Projection (UMAP). The best clusters, along with the selection of certain hyperparameters, was determined through visual analysis of the groups in a two-dimensional UMAP plane.

Keywords

Clustering, REST-MEX, 2023, Transformers, Sentence-transformers, Spacy, NLP, Spanish,

1. Introduction

Within the Rest-Mex framework, an unprecedented unsupervised classification task was embarked upon in the 2023 competition [1], signifying a notable departure from previous approaches [2, 3]. The task centered around clustering a substantial dataset comprising approximately 100,000 news items related to four distinct tourism topics. The overarching objective was to generate four cohesive groups by employing advanced text analysis techniques. These news items, meticulously collected from Google News over a span of two years, formed the foundation of the competition.

The extensive corpus of approximately 100,000 labeled news articles served as the backbone of this research endeavor. The participants were expected to navigate the vast dataset, applying their clustering algorithms to identify underlying patterns and generate distinct groups of news items. The competition demanded both technical expertise in natural language processing and a

IberLEF 2023, September 2023, Jaén, Spain


*Corresponding author.

✉ javilonso9@gmail.com (J. Alonso-Mencía)

🌐 <https://javieralonso.io/> (J. Alonso-Mencía)



© 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)

keen understanding of the tourism domain. By engaging with this unique challenge, researchers sought to unlock valuable insights into the organization and categorization of tourism-related news content.

2. Methodology

2.1. Data processing

| News - Token analysis | |
|-----------------------|---------|
| Count | 114550 |
| Mean | 3999.20 |
| Std | 3425.66 |
| Min | 0 |
| 25% | 2075 |
| 50% | 3119 |
| 75% | 4914 |
| Max | 32732 |

Table 1
Analysis of tokens in News field

2.1.1. Text decoding

For the decoding step in the news clustering NLP competition, a function was employed to process the text data. The function aimed to handle potential encoding issues and remove unnecessary components from the text.

Within the function, an attempt was made to encode the text using the UTF-8 encoding and subsequently decode it using the "unicode_escape" encoding. This step aimed to handle any special characters or escape sequences present in the text. In case an exception occurred during the encoding and decoding process, the function gracefully handled it by reverting to the original text and applying the same step of removing the initial link and trailing space.

2.1.2. Removing links

A crucial step involved removing links from the text data. A function was designed to eliminate the initial link present in the text of every news. By applying this function to the "News" column of the dataset, the links were successfully removed, ensuring a cleaner and more focused representation of the news content for subsequent analysis.

2.1.3. Removing punctuation

A procedure was implemented to remove punctuation from the textual data. This involved utilizing the string module to eliminate ASCII punctuation. Additionally, regular expressions

were employed to eliminate any remaining punctuation, including newline characters, resulting in a refined dataset ready for further analysis.

2.1.4. Removing stop words

A function was implemented with the aim of improving the quality of the data by performing two key operations: converting the text to lowercase and removing stopwords.

To accomplish this, the function utilized the NLTK library to download the stopwords for the Spanish language. These stopwords, commonly occurring words that typically do not contribute much to the meaning of the text.

Within the function, the input text was split into individual words, and each word was converted to lowercase. The function then checked if the lowercase version of each word existed in the "stop_words" set. If a word was not present in the set, it was considered significant and added to the filtered_text list.

Finally, the filtered_text list was joined back into a string, with each word separated by a space, and returned as the preprocessed version of the original text. By applying this function to the data, the text was effectively transformed to lowercase and devoid of stopwords.

2.1.5. Spacy lemmatization

In order to further refine the text data for the news clustering NLP competition, an additional preprocessing step was applied, namely lemmatization. This technique involved transforming words to their base or canonical form, enhancing the accuracy and interpretability of the text analysis. By leveraging the Spanish language model "es_core_news_md" from the Spacy library, the words were effectively lemmatized, capturing their essential meaning and standardizing their representation.

This lemmatization process improved the accuracy and meaningfulness of the analysis, making the subsequent clustering tasks in the competition more effective. It is important to mention that it was decided to keep a copy of the dataset without lemmatization applied in order to analyze afterwards its performance.

2.2. Generating text embeddings

2.2.1. Sentence transformers

The usage of Sentence Transformers [4] instead of a conventional transformer like BERT [5] offers several advantages. Firstly, Sentence Transformers are specifically designed to generate sentence-level embeddings, capturing the semantic meaning of entire sentences rather than just individual words or tokens. This ability allows for a more comprehensive representation of text data and facilitates downstream tasks such as clustering and similarity comparisons.

Secondly, Sentence Transformers are trained on a diverse range of multilingual data, enabling them to handle text from different languages effectively. This multilingual capability is

particularly advantageous in scenarios involving multilingual or cross-lingual analysis, where BERT models may not be as effective.

Additionally, Sentence Transformers are pre-trained on large-scale datasets, incorporating a wealth of linguistic knowledge and contextual information. This pre-training enhances their ability to understand the nuances and intricacies of natural language, resulting in improved performance on various NLP tasks.

Moreover, Sentence Transformers offer efficient inference and computation times, making them suitable for real-time or large-scale applications. Their architecture and optimization techniques ensure fast and effective encoding of textual data.

Overall, employing Sentence Transformers over traditional transformers like BERT provides enhanced sentence-level embeddings, multilingual support, comprehensive linguistic knowledge, and efficient computation, thereby improving the performance and versatility of NLP applications, including news clustering in this particular competition.

2.2.2. Applying sentence transformers

In order to generate text embeddings for the news clustering NLP competition, a widely-used library called Sentence Transformers [6] available at HuggingFace platform [7] was utilized. This library incorporates pre-trained models that are capable of encoding textual data into numerical vectors.

Specifically, a pre-trained model named `sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2` available at HuggingFace platform was employed, as it is a multilingual model including Spanish language. This model maps sentences and paragraphs to a 384 dimensional dense vector, important fact that will be addressed in the following section.

The process involved loading the SentenceTransformer model and retrieving the texts from the dataset. These texts were then passed through the model to obtain their corresponding text embeddings. These text embeddings represent the semantic meaning of the texts in the form of numerical vectors. By generating these text embeddings, the textual data was transformed into a suitable format for subsequent analysis.

2.3. Reducing dimensionality

It was necessary to perform dimensionality reduction on the sentence embeddings generated by the Sentence Transformer model. The initial sentence embeddings had a dimensionality of 384, which could potentially hinder the efficiency and effectiveness of subsequent clustering tasks.

To address this, a dimensionality reduction technique called UMAP (Uniform Manifold Approximation and Projection) [8] was employed. UMAP is known for its ability to preserve the local and global structure of the data while reducing its dimensionality. In the implementation used, the UMAP algorithm was applied with the 'cosine' metric, ensuring the consideration of cosine similarity during the reduction process.

By applying UMAP, the sentence embeddings were compressed into a lower-dimensional space, allowing for more efficient processing and analysis. The reduction of dimensionality helped to alleviate computational complexity and potential noise in the data, improving the overall clustering performance.

In order to determine the optimal configuration for the UMAP dimensionality reduction technique, different combinations of `n_neighbors` and `n_components` were tested. The range of values explored for `n_neighbors` was from 4 to 11, while `n_components` ranged from 4 to 11 as well. By systematically varying these parameters, the effects of different neighborhood sizes and reduced dimensions on the clustering performance could be assessed.

After evaluating the results, the combination yielding the best clustering performance was selected. The final choices of `n_neighbors` and `n_components` values, combined with the clustering techniques are shown in the next section.

2.4. Clustering

After the dimensionality reduction step, two clustering methods, namely **K-means** and **Gaussian Mixture**, were applied to the reduced embeddings. These methods were chosen for their effectiveness in clustering tasks and their availability in the widely-used scikit-learn python library [9]. Both K-means and Gaussian Mixture were configured to generate **4 clusters**, as specified by the requirements of the news NLP competition.

While other clustering methods exist, they were not considered for this particular competition due to the requirement of specifying the number of clusters. Certain clustering algorithms, such as DBSCAN or Agglomerative Clustering, do not inherently require the prior knowledge of the number of clusters. However, since the competition guidelines specified four clusters, methods that necessitate specifying the cluster count were employed.

3. Results and discussion

3.1. Evaluation

To select the best submissions in the news NLP competition, a combination of methods was employed. Initially, the UMAP dimensionality reduction technique was used to explore various combinations of `n_neighbors` and `n_components`, ranging from 4 to 11, in order to determine the optimal configuration. By systematically testing different parameter values, the effects of neighborhood sizes and reduced dimensions on clustering performance were assessed.

Following the dimensionality reduction, two clustering methods, K-means and Gaussian Mixture, were applied to the reduced embeddings. Both methods were executed with four clusters, as specified by the competition guidelines. This allowed for the partitioning of news articles into distinct categories or topics.

To evaluate the clustering results, an additional UMAP dimensionality reduction was performed, reducing the embeddings to two dimensions. This reduction enabled the generation

of a plot that visualized the clustering outcomes. The selected submissions were primarily determined based on the parameters chosen, particularly those that visually demonstrated four well-defined clusters on the plot. You can visualize the final submissions in Figure 1 and Figure 2.

While other methods for selecting the best submissions may exist, such as performance metrics or statistical measures, the limited time available for the competition precluded further exploration of alternative approaches. Despite this limitation, the combination of UMAP dimensionality reduction, clustering with K-means and Gaussian Mixture, and the visual analysis of the reduced embeddings proved effective in identifying the most promising submissions.

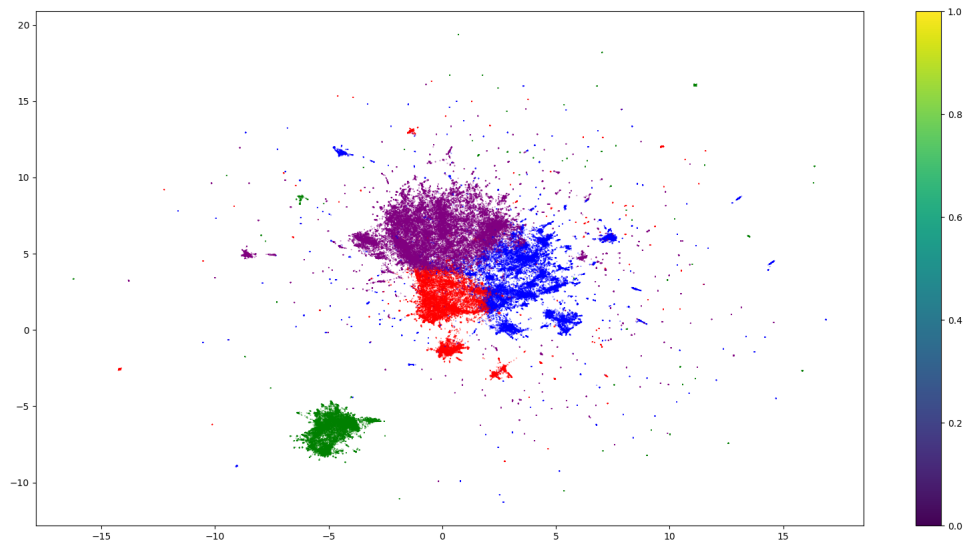


Figure 1: Plot representing in 2-dim instances distributed in 4 clusters, one per color [purple, blue, red, green]. Spacy lemmatization was used in this case. K-means clustering technique. $n_neighbors=10$; $n_components=5$

Table 2 shows the configurations used to submit the possible clusters to the competition.

| Submission | Clustering method | Num. clusters | Num. neighbors | Num. components | Spacy lemmatization | Figure |
|------------|-------------------|---------------|----------------|-----------------|---------------------|--------|
| 1 | K-means | 4 | 10 | 5 | Yes | 1 |
| 2 | K-means | 4 | 10 | 11 | No | 2 |

Table 2
Configurations submitted in the news clusterings competition

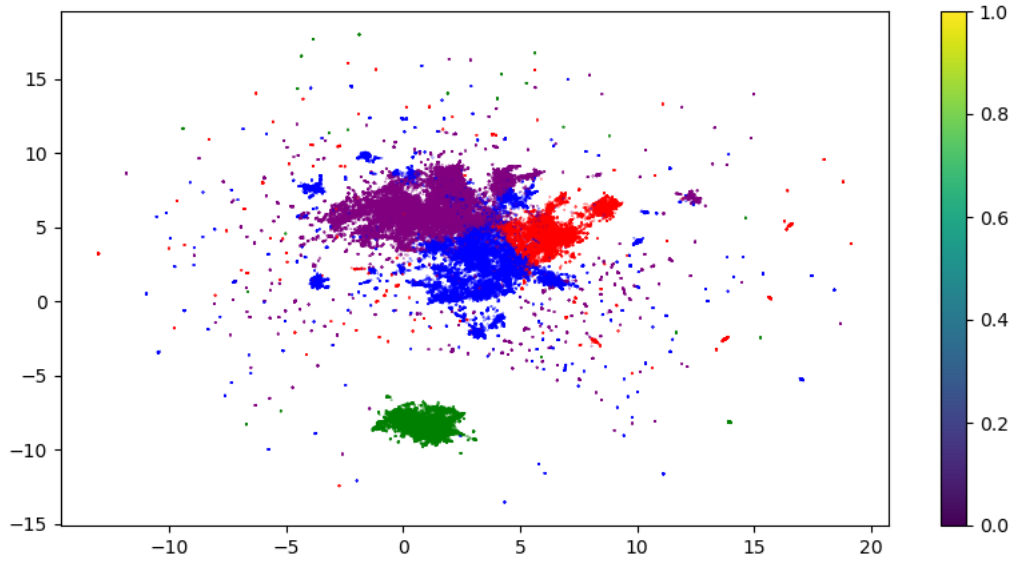


Figure 2: Plot representing in 2-dim instances distributed in 4 clusters, one per color [purple, blue, red, green]. Spacy lemmatization was not used in this case. K-means clustering technique. $n_neighbors=10$; $n_components=11$

3.2. Possible news themes

In **Submission 1** (refer to Figure 1 and Table 2), it was analyzed the predominant words within each cluster using the **tf-idf** technique. The objective was to deduce potential themes associated with the news content of each cluster. The resulting most frequent words for each cluster are presented below:

1. **Cluster 0:** ['escrito', 'por', 'nota', 'publicado', 'notificación', 'actriz', 'cantante', 'instagram', 'compartir', 'video', 'actor', 'minuto', 'debate', 'amor', 'historia', 'yo', 'azteco', 'mundo', 'azteca', 'siempre']
2. **Cluster 1:** ['cooki', 'personalizar', 'utilizamos', 'sol', 'optimizar', 'clic', 'acordar', 'acerca', 'usted', 'mejorar', 'aceptar', 'sitio', 'ofrecer', 'experiencia', 'publicidad', 'visitar', 'uso', 'precio', 'pesos', 'tampico']
3. **Cluster 2:** ['turismo', 'maya', 'turistico', 'salud', 'desarrollo', 'agua', 'proyecto', 'merida', 'educacion', 'roo', 'millón', 'sector', 'covid19', 'empresa', 'ciento', 'economico', 'nivel', 'programa', 'vila', 'internacional']
4. **Cluster 3:** ['violencia', 'fiscalia', 'delito', 'cartel', 'victima', 'homicidio', 'arma', 'guardia', 'detenido', 'policia', '135', 'min', 'reproducción', 'armado', 'criminal', 'investigacion', 'lectura', 'politico', 'vehiculo', 'elemento']

Hereafter, the previous words were used in ChatGPT tool [10] (based on GPT3 generative model [11]) to infer possible topics for each cluster, as follows:

1. **Cluster 0:** Theme related to news about actors, actresses, singers, social media platforms (such as Instagram), debates, and published articles.
2. **Cluster 1:** Theme related to personalization, optimization, cookies, online advertising, website visits, prices, and the city of Tampico.
3. **Cluster 2:** Theme related to tourism, particularly in the Yucatan region of Mexico, including tourist destinations such as the Riviera Maya and Merida, as well as topics of health, economic development, education, and international programs.
4. **Cluster 3:** Theme related to violence, organized crime, crimes, police investigations, victims, homicides, weapons, and political elements associated with security.

Table 3 shows the number of instances and the main topic inferred by ChatGPT for each cluster.

| Cluster | Instances | Possible topic (ChatGPT) |
|---------|-----------|--------------------------------------|
| 0 | 22023 | Entertainment and Social Media |
| 1 | 15995 | Online Customization and Advertising |
| 2 | 31441 | Tourism and Regional Development |
| 3 | 45091 | Crime and Security |

Table 3

Number of instances per cluster in submission 1, as well as the possible topic for each cluster news inferred by ChatGPT

3.3. Competition results

The first place in the ranking was obtained in one of the approaches proposed in this paper (**Submission 1**), achieving an overall F1 score of 0.282 (see Table 4 and Table 5). The system attained an accuracy of 44.81, indicating its overall performance in classification. With respect to specific categories, the system obtained an F1 score of 0.615 for Insecurity, showcasing its ability to accurately cluster news items related to this theme. In terms of Prices, the system achieved an F1 score of 0.233, demonstrating its competence in effectively categorizing news articles discussing price-related aspects. Gastronomy-related news articles were also classified with reasonable accuracy, yielding an F1 score of 0.2133. However, the system's performance in categorizing news items related to Landscapes was relatively lower, with an F1 score of 0.069.

Overall, the submission showcased commendable performance, emerging as the winner in the competition. This achievement highlights the system's robustness and effectiveness in unsupervised text classification, even with variations in performance across different thematic clusters.

3.4. Comparing actual and inferred news themes

In comparing the inferred topics with the actual topics for your paper, there are both similarities and differences.

| Rank | Run | Macro F1 | Accuracy |
|------|--|----------|----------|
| 1st | Javilonso-Team_javier_alonso_thematic_spacy_kmeans10_5 | 0,282 | 44,81 |
| 2nd | JCMQ-2-Team_run3_thematic | 0,240 | 35,62 |
| 3th | JCMQ-Team_run_5 | 0,218 | 35,27 |

Table 4
Overall final results from Clustering competition Rest-Mex 2023

| Rank | Run | F1 (Insecurity) | F1 (Prices) | F1 (Gastronomy) | F1 (Landscapes) |
|------|--|-----------------|-------------|-----------------|-----------------|
| 1st | Javilonso-Team_javier_alonso_thematic_spacy_kmeans10_5 | 0,615 | 0,233 | 0,213 | 0,069 |
| 2nd | JCMQ-2-Team_run3_thematic | 0,510 | 0,240 | 0,140 | 0,068 |
| 3th | JCMQ-Team_run_5 | 0,516 | 0,127 | 0,068 | 0,159 |

Table 5
Specific results per cluster obtained in Clustering competition Rest-Mex 2023

- The inferred topic of **Entertainment and Social Media** partially aligns with the actual topic of **Gastronomy**. While the inferred topic captures the essence of the entertainment industry, focusing on actors, actresses, singers, and social media platforms, it does not directly correspond to the actual topic of Gastronomy.
- The inferred topic of **Online Customization and Advertising** closely matches the actual topic of **Prices**. The inferred topic encompasses personalization, optimization, cookies, online advertising, and website visits, which aligns well with the actual topic of Prices.
- The inferred topic of **Tourism and Regional Development** partially aligns with the actual topic of **Landscapes**. While the inferred topic captures the aspects of tourism, including destinations, economic development, education, and international programs, it does not specifically mention landscapes.
- The inferred topic of **Crime and Security** aligns with the actual topic of **Insecurity**. The inferred topic captures violence, organized crime, police investigations, victims, homicides, weapons, and political elements associated with security, closely resembling the actual topic.

Overall, the inferred topics show both accurate and partial alignment with the actual topics. It is important to note the model's ability to capture the general themes and related concepts, but it may not always precisely identify the specific topics being discussed.

4. Conclusions

In conclusion, this paper presented a comprehensive approach to the news clustering NLP competition, involving preprocessing steps such as punctuation and link removal, lowercasing, stop word removal, and lemmatization. The Sentence Transformers model was utilized to generate high-dimensional sentence embeddings capturing semantic information. The UMAP algorithm was then employed to reduce the dimensionality of the embeddings, enabling more efficient analysis.

Subsequently, K-means and Gaussian Mixture clustering methods were applied to identify distinct clusters representing news topics or themes. The UMAP dimensionality reduction facilitated the selection of the best submissions based on visually well-defined clusters.

The results highlight the significance of preprocessing, dimensionality reduction, and clustering techniques in achieving meaningful outcomes in news clustering tasks, while future research can explore additional methods for evaluation and enhancement of clustering approaches.

While the clustering results obtained in the news competition NLP were satisfactory, they did not reach a level of brilliance. This could be attributed to several factors, including the complexity and diversity of news articles, the limitations of the chosen clustering algorithms, and the inherent subjectivity and ambiguity present in news categorization.

As future work, the application of few-shot learning techniques, such as SetFit, could be explored to train a classification model using a small number of manually extracted examples. This approach would address the challenge of limited labeled data by enabling the model to generalize to new classes with only a few labeled samples. By carefully selecting representative examples and incorporating them into the training process, the model's ability to classify news articles into meaningful categories could be enhanced.

The experimentation can be found in a Github repository [12].

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] 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).
- [3] 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).

- [4] N. Reimers, I. Gurevych, Sentence transformers: Multilingual sentence embeddings using BERT, 2020.
- [5] Nlptown/bert-base-multilingual-uncased-sentiment, 2021. URL: <https://huggingface.co/nlptown/bert-base-multilingual-uncased-sentiment>.
- [6] N. Reimers, I. Gurevych, Sentence transformers, <https://www.sbert.net>, 2021. Accessed: [Insert Date].
- [7] Hugging face – the ai community building the future., 2023. URL: <https://huggingface.co/>.
- [8] L. McInnes, J. Healy, J. Melville, Umap: Uniform manifold approximation and projection, in: 2018 IEEE International Conference on Data Mining Workshops (ICDMW), IEEE, 2018, pp. 230–237. URL: <https://umap-learn.readthedocs.io>. doi:10.1109/ICDMW.2018.00032.
- [9] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, É. Duchesnay, Scikit-learn: Machine learning in Python, *Journal of Machine Learning Research* 12 (2011) 2825–2830.
- [10] OpenAI, Openai, n.d. URL: <https://chat.openai.com/>.
- [11] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei, Language models are few-shot learners, in: H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, H. Lin (Eds.), *Advances in Neural Information Processing Systems*, volume 33, Curran Associates, Inc., 2020, pp. 1877–1901.
- [12] J. Alonso, Restmex23_nlp, https://github.com/javilonso/RestMex23_NLP, 2023. Accessed: May 23, 2023.