**Supplementary material**

**Data analysis**

**Sentiment Analysis with portuguese tweets data**

In the scope of our study, an extensive dataset comprising over 100,000 tweets concerning the 74 Brazilian National Parks was acquired. The primary objective was to discern the sentiment embedded within these tweets and ascertain the principal themes associated with each sentiment classification. The initial phase encompassed the classification of these tweets based on their sentiment through the utilization of a pre-trained transformer model - BERTimbau Base (a.k.a. "bert-base-portuguese-cased") (Souza et al. 2020). This model was subsequently refined to cater specifically to our task, which involves the categorization of tweets into either positive, negative, or neutral sentiments.

Given the focus on understanding and classifying the sentiment expressed within tweets pertaining to Brazilian national parks, the model necessitated fine-tuning utilizing a pre-structured dataset that had already been categorized across a spectrum of sentiment types. Nonetheless, it is noteworthy that pre-trained models such as BERTimbau Base are primarily designed for three fundamental natural language processing (NLP) tasks: Named Entity Recognition, Sentence Textual Similarity, and Recognizing Textual Entailment, and their direct application to text classification is not inherently straightforward (Tunstall L. et al., 2022). Hence, adaptations were implemented to align the model with our distinct classification objectives.

In this endeavor, we harnessed a comprehensive dataset originating from the B2W e-commerce company, in conjunction with the Corpus Buscapé (B2W-Reviews01, 2018; Corpus Buscapé, 2013). This publicly accessible dataset encompasses in excess of 130,000 user reviews encompassing a diverse range of products. Among its attributes, it features binary labels denoting whether users would recommend the product to others and ratings scored on a scale from 1 to 5 stars. It is pertinent to mention that our analysis solely engaged with the user rating component.

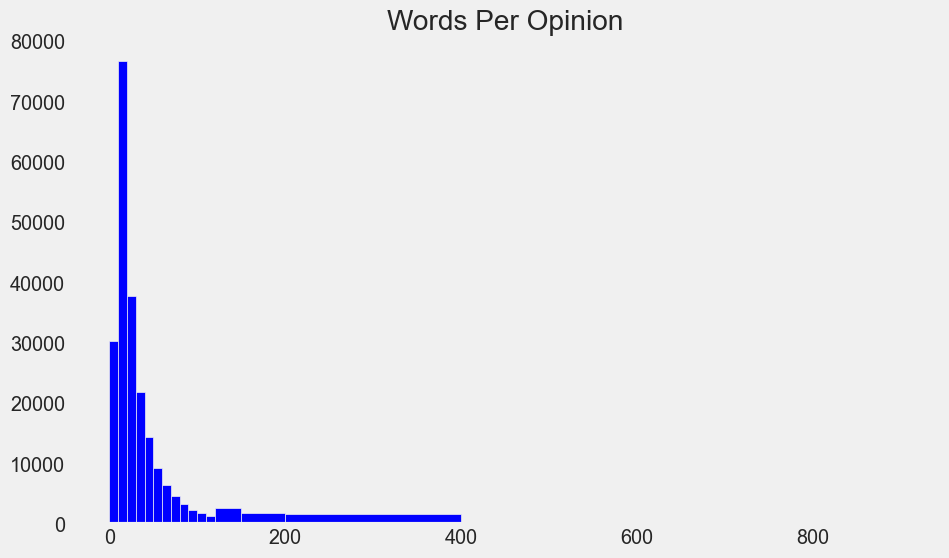
Opinando, an institution specializing in mining opinions from Portuguese textual content and established under the auspices of the Research Office of the University of São Paulo (USP), has contributed significantly to the domain of natural language processing (NLP). One of its notable contributions is the creation of various robust corpora germane to this field. Among these, the Corpus Buscapé emerges as a substantial compilation of Portuguese product reviews, harvested in the year 2013 from the Buscapé website, renowned for its product and price exploration capabilities. Unlike the previously referenced datasets, this corpus employs a rating spectrum spanning from 0 to 5. As a corollary, comments affiliated with a rating of zero were omitted from the analytical considerations.

The decision to select this particular dataset hinged on its capacity to yield precise predictions for the classification of the target tweets. Notably, this opinion-oriented dataset is stratified based on a "rating" feature, encompassing scores ranging from 1 to 5. A rating of 1 signifies a markedly negative sentiment, while a rating of 5 corresponds to a highly positive sentiment. As part of our classification strategy, ratings of 1 and 2 were grouped as "negative," a rating of 3 as "neutral," and ratings of 4 and 5 as "positive." With the model having undergone training and validation using this dataset, its application was extended to a subset of approximately 2,000 tweets culled from our larger collection. This selected subset had been manually categorized in advance to serve as a dedicated test set for evaluation purposes.

**How Long Are Our Tweets?**

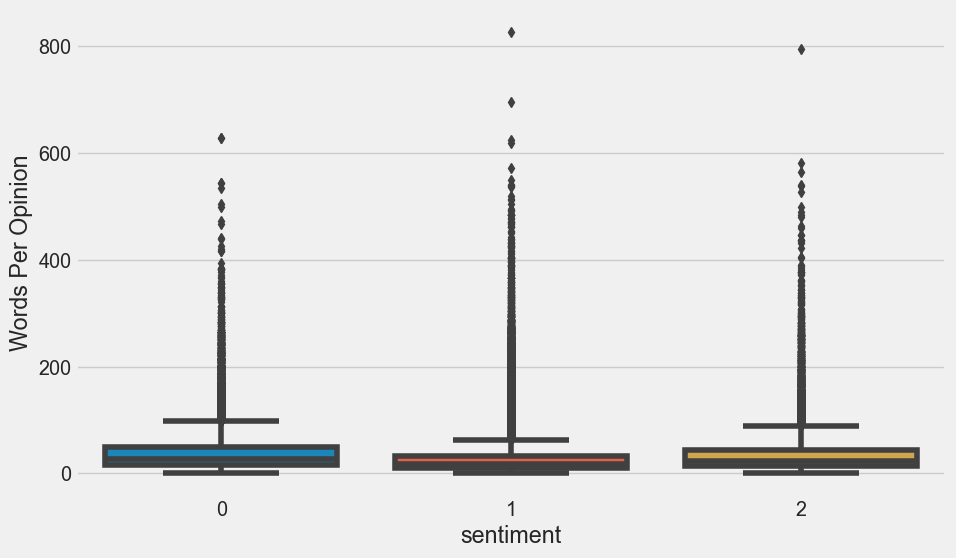
Transformer models possess a defined upper limit for the length of input sequences, commonly denoted as the "maximum context size" or "tokens." Tokens within these models can encompass complete words, subword components, or even singular characters, such as punctuation marks.

For instance, a tweet is constrained to a maximum of 280 characters. Considering the average of approximately 6 characters per word in the Portuguese language, the typical tweet may contain approximately 47 words, contingent on the nature of the content under analysis. The e-commerce sales platform’s dataset on product reviews presents the following word frequency:



**Figure 1:** Word frequency about product reviews on Buscapé.

By type of sentiment:



**Figure 2:** Boxplot illustrating the sentiment per word in the reviews from the Buscapé website. The number 1 represents positive sentiment, 0 represents negative sentiment, and 2 represents neutral sentiment.

Positive sentiments generally exhibit greater word length in their expressions; however, this trend is subject to exceptions, as evidenced by the elongated tails observed in the boxplot representations. Hence, adaptations were implemented to align the model with our distinct classification objectives through Transfer Learning techniques.

**BERTopic analysis**

BERTopic is a Python library for natural language processing topic modelling that combines transformer embeddings with clustering algorithms to identify topics in a corpus of texts (Grootendorst, 2022). The BERTopic model supports over 50 languages and has been compared to other models, such as LDA, for performing topic modelling on short texts from social media platforms and has shown exceptional performance in extracting topic representations (Egger and Yu 2022).

In the first step of the BERTopic algorithm, we obtain embeddings for all documents in the corpus, which are numeric vector representations of the documents. The next step is to perform clustering on the embedded documents, for which dimensionality reduction techniques, such as Uniform Manifold Approximation and Projection (UMAP), are employed to reduce the high dimensionality of the embedding vectors ([McInnes](https://arxiv.org/search/stat?searchtype=author&query=McInnes%2C+L) et al., 2018). The UMAP algorithm is used by default because it preserves both the local and global structure of the data with superior runtime performance, an important factor in representing the semantics of text data. Preprocessing of text data is an optional step in natural language processing. Generally, It is not recommended to remove stop words as a preprocessing step when using the BERTopic model because transformer-based embedding models, which we utilise, require the complete context to generate accurate embeddings.

The default clustering algorithm used by BERTopic is HDBSCAN, which is a density-based model that automatically identifies the number of clusters without requiring a pre-specified number of clusters. HDBSCAN is a hierarchical density-based clustering algorithm proposed by (Campello et al., 2013). In this algorithm, documents with higher similarity are grouped into clusters based on cluster stability. One important characteristic of HDBSCAN is that it does not force the assignment of a data point to a specific cluster. If the data point does not fit into any similarity-based group, it is considered an outlier (Capellaro, 2021). Once the documents are assigned to clusters, the next step is to obtain the topic representation for each cluster using class-based Term Frequency-Inverse Document Frequency (c-TF-IDF).

This method selects the top words with the highest c-TF-IDF scores within a cluster to represent each topic ([Grootendorst](https://arxiv.org/search/cs?searchtype=author&query=Grootendorst%2C+M), 2022). This means that the higher the value of a term, the more representative it is of its topic. Furthermore, following the identification of the values associated with each term within the topics, a comprehensive evaluation and inspection of the topics was conducted to detect any potential content that might be misconstrued as a singular topic (See Table 1 in results). This consideration, as noted by Egger and Yu (2022), highlights a potential limitation of the model, particularly when dealing with extensive amounts of data for analysis. Given that the use of BERTopic also requires significant effort due to the dynamic nature of topic structures, which change when researchers experiment with different numbers of topics, it can be considered a laborious task to access the topics that best represent the database. Although BERTopic offers the advantage of leveraging domain-specific knowledge to search for specific topics, as done in this study, this process can still be considered exhaustive.

For the purpose of our study, two main steps were undertaken: (i) identification of potential negative topics within our corpus, encompassing all Brazilian national parks, and (ii) segregation of tweets specifically related to the six most frequently visited parks (ICMBio, 2021), followed by clustering to discern the prominent negative topics associated with each individual park. To achieve this, we performed the BERTopic model with the following hyperparameters:

* For the UMAP algorithm we set n\_neighbors or the number of samples used during the manifold approximation to 15, n\_components or the dimensionality that holds the most information possible to 5, min\_dist to 0, in order to get more clustered embeddings and selected the cosine metric to compute distances in high dimensional space.
* For HDBSCAN we set the metric to euclidean in order to compute distances in an array and prediction\_data to True to be able to apply to our dataset later, not just to fit the model, for all the datasets, no matter what park the tweets are from. And we set the min\_cluster\_size parameter or minimum size of the clusters depending on the number of observations we have. The purpose is to reach a reasonable number of topics and also that they contain coherent information to know what they are talking about.
* For BERTopic we set the parameter nr\_topics to auto in order to focus on the interpretation of the topics. Besides we use the function CountVectorizer with a list of portuguese stopwords and ngram\_range between 1 and 2 n-gram words to be extracted, and the function ClassTfidfTransformer in order to reduce the impact of the most frequent words, also the MaximalMarginalRelevance function in order to limit the number of duplicate words that we can find in each topic, and finally, the function SentenceTransformer with the bert-base-portuguese-cased model in order to use the same embedding model for the negative tweets selected as in the previous prediction step.

After evaluating the possible topics generated by the model, we identified, based on our knowledge, the topics that remained consistent across all generated models.

**References**

Campello, R., Moulavi, D., Sander, J., 2013. Density-Based Clustering Based on Hierarchical Density Estimates. Gold Coast.

Egger, R., Yu, J., 2022. A Topic Modeling Comparison Between LDA, NMF, Top2Vec, and BERTopic to Demystify Twitter Posts. Frontiers in Sociology 7. <https://doi.org/10.3389/fsoc.2022.886498>

Grootendorst, M., 2022. BERTopic: Neural topic modeling with a class-based TF-IDF procedure.

ICMBio, 2021. Unidades de conservação federais atingem novo recorde de visitação em 2021. Accessed on 23 April 2023. <https://www.gov.br/pt-br/noticias/viagens-e-turismo/2022/04/unidades-de-conservacao-federais-atingem-novo-recorde-de-visitacao-em-2021>

McInnes, L., Healy, J., Melville J., 2018. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. <https://arxiv.org/abs/1802.03426>

Souza, F., Nogueira, R., Lotufo, roberto, 2020. BERTimbau: Pretrained BERT Models for Brazilian Portuguese, in: 9th Brazilian Conference,. pp. 403–417. https://doi.org/https://doi.org/10.1007/978-3-030-61377-8

Tunstall L., von Werra L., Wolf T., 2022. Natural Language Processing with Transformers. O’Reilly Media, Inc.