Weekly update - 2023-04-04 to 2023-04-11

Paula Jeniffer dos Santos Viriato[1][0000−0003−0900−1686]

Institute of Computing, University of Campinas, Campinas - SP, Brazil
p234831@dac.unicamp.br

1 Clustering and Analytics - Facebook Posts of Amazon Tourism

The *Facebook Posts of Amazon Tourism* dataset contains publications by small and medium-sized companies in a specific sector: tourism in the Amazon region of Brazil. Sousa [4] proposed this dataset and the post categories were designated as follows:

1. Uncategorized
2. Announcement of a new product
3. Sweepstakes and contests
4. Sales
5. Consumer Feedback
6. Infotainment
7. Organization Brand
8. Non-agreement between evaluators

During the analysis of the dataset and the attempt at clustering, the high specificity of the data, which caused a similarity of vocabulary, made it difficult to identify possible clusters. Figure 1 presents the ten most used words in each identified cluster.

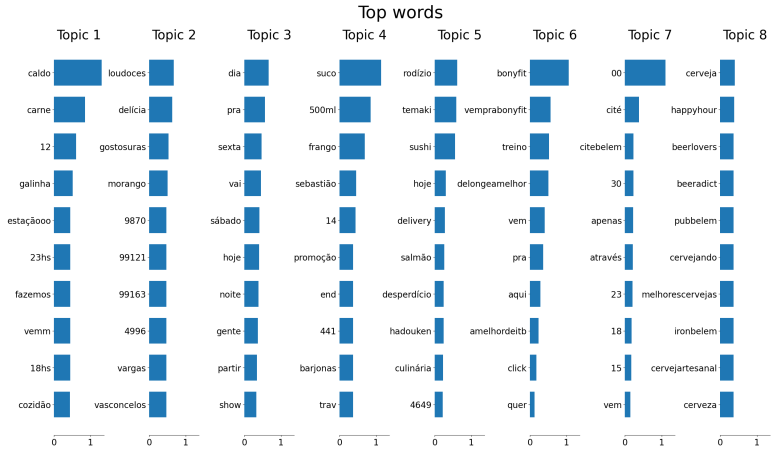

Fig. 1: Most used words in each identified cluster

In Figure 2, comparing the word cloud of topics 1, 2, and 3 identified, we can see that the dataset as a whole is about tourism and is biased toward the intended purposes of this project. We found that topic 1 tends to talk about events, and topics 2 and 3 tend to talk about food.

Fig. 2: Comparing the word cloud of topics 1, 2, and 3

Finally, Figure 3 presents the level of cohesion of the dataset, as all topics are highly correlated, and the size of the difficulty of distinguishing between posts. We can then discard the use of this dataset, as it is too specific and unsuitable for our purposes.

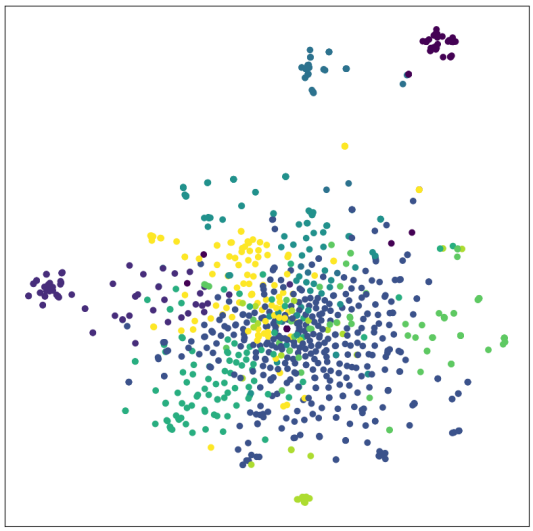

Fig. 3: Dataset Cluster Distribution

The Python Notebook used to perform the clustering and analysis is available at the following link.

2 Analysis - News of the Brazilian Newspaper

The dataset by Santana [2] consists of 167,053 reports from the Brazilian newspaper Folha de São Paulo. This dataset is too vast, extensive, and heavy to be processed in online environments, such as Google Collaboratory, making it difficult to post on GitHub. Such a dataset must be divided into smaller parts for better management and clarity.

Another difficulty regarding this dataset is that it has 123,411 categories, almost the same number of reports present, which would result in a cluster with a very small number of reports. The alternative would be the subcategory field, but this field has a null value for 82% of the reports.

This dataset does not seem attractive for a first clustering because it is practically not categorized, but we can use it in the future for comparison with other clusterings.

3 Next steps

An alternative thought about a dataset in which it is possible to verify more precisely the interests of the users is the extraction of data from Reddit forums. Reddit is the self-proclaimed "front page of the internet" [3], being the world's largest online discussion forum [1]. It is organized into topics such as music, movies, gardening, computing, and more. The extraction of data from these forums focused on specific themes can help in a stronger recommendation for users.

- Extract posts and comments from Reddit;
- Search for a textually descriptive marketing dataset;
- Research on *Invisible Machines*;
- Read the paper *Predicting the Need for Xai from High-Granularity Interaction Data*;
- Project schedule generation;
- Explore courses within the scope of the project:
  - Neuromarketing and mental triggers;
  - Inbound marketing;
  - Digital marketing.
- MC750 Course Planning for Digital Marketing Assessment.

References

1. Medvedev, A.N., Lambiotte, R., Delvenne, J.C.: The anatomy of reddit: An overview of academic research. Dynamics On and Of Complex Networks III: Machine Learning and Statistical Physics Approaches 10 pp. 183–204 (2019)
2. Santana, M.: Kaggle-news of the brazilian newspaper (2021)
3. Singer, P., Flöck, F., Meinhart, C., Zeitfogel, E., Strohmaier, M.: Evolution of reddit: from the front page of the internet to a self-referential community? In: Proceedings of the 23rd international conference on world wide web. pp. 517–522 (2014)

4. de Sousa, G.N., Junior, A.F.L.J., Lobato, F.M.F.: Facebook posts for analyzing Content Strategies for Digital Consumer Engagement: a curated dataset (Jul 2021). https://doi.org/10.5281/zenodo.5113266, https://doi.org/10.5281/zenodo.5113266