Automated Business Classification using Cosine Similarity and Sentence Transformers

Sebastian Cândea

March 2025

1 Introduction

Business classification is a critical task in various industries, particularly for taxation and regulatory compliance. This project aims to classify businesses based on their descriptions, business tags, sectors, categories, and niches by leveraging embeddings and cosine similarity. The dataset consists of 9,500 businesses, and the classification is performed against a predefined taxonomy.

2 Methodology

We utilize a zero-shot classification approach by computing sentence embeddings using `all-MiniLM-L6-v2`, a lightweight transformer model from Sentence Transformers. The embeddings of business descriptions are compared with taxonomy labels using cosine similarity. The workflow involves:

1. Preprocessing textual data: lowercasing, removing special characters, stopword removal, and lemmatization.
2. Generating embeddings for taxonomy labels and business attributes.
3. Computing weighted embeddings based on attribute importance.
4. Assigning labels using cosine similarity between business embeddings and taxonomy embeddings.

2.1 Challenges in Data

The dataset contains inconsistencies, especially in the `business_tags` column. For example:

```
Theme Island Marine Terminal is a ship chandlery company that
specializes in the export and import of cruise ships.  They
are the largest ship chatterer in the North Georgia RFP market
and have licenses to export passenger ships (Class 4 and Class
9) as well as general cargo ships (Class 4 and 9).
['Non-profit Organization', 'Sportsmanship Training', 'Coaching
Services', 'Sportsmanship Association']
```

This demonstrates a clear mismatch between the business description and tags, which can affect classification accuracy.

3 Results

The classification model achieved the following metrics based on cosine similarity scores:

Metric	Value
Mean Cosine Similarity	0.6037
Standard Deviation	0.0927
Min Similarity	0.3030
25th Percentile	0.5420
Median Similarity	0.5978
75th Percentile	0.6583
Max Similarity	0.9102

Table 1: Cosine Similarity Distribution of Assigned Labels

This implies that, on a sample of 10 companies, around 6-7 are classified very close to their correct class (as presumed by the author), or even correctly assigned.

4 Code Structure and Logic

The project consists of two main scripts:

4.1 Classification Script (`cod_clasificare.py`)

- Reads `date.csv` and taxonomy labels.
- Generates embeddings for each company by weighting attributes.
- Computes cosine similarity and assigns the best matching label.
- Saves results back to `date.csv`.

4.2 Preprocessing Script (`preprocesare.py`)

- Cleans text by lowercasing, removing stopwords, and lemmatizing.
- Generates embeddings for preprocessed text.
- Saves cleaned data for further classification.

5 Conclusion

This approach demonstrates a practical zero-shot classification methodology for business taxonomy mapping. Future work could involve fine-tuning embeddings with supervised learning or integrating external knowledge sources to enhance accuracy.