Transforming the Delivery of Public Information in Mexico with Advanced Language Models

Workshop on the Modernisation of Official Statistics
21-22 November 2023
Introduction

• National Statistical Institute of Statistics and Geography of Mexico (INEGI) is always exploring cutting-edge technologies which potentially may add value to the services it provides to society.

• It has actively engaged in developing natural language projects.
  – Improvement of its Search Engine
  – Coding of Economic Statistics
  – Analysis of Gender Oriented Services
Among the multitude of purposes that the Mexican Statistical Institute is pursuing, the dissemination of official statistics stands out as a crucial duty.

- In 2017, 49.1% of the webpage users mentioned having found the information they were looking for.
- In 2022, a search engine was added, but still only 54.4% of search outputs were useful for the users.
- The bounce rate after making a query was 68.8%
Search Engine Improvement

Search Engine Comparison

<table>
<thead>
<tr>
<th>Category</th>
<th>WITHOUT Enhanced Vocabulary</th>
<th>WITH Enhanced Vocabulary</th>
</tr>
</thead>
<tbody>
<tr>
<td>Matches expectations</td>
<td>85</td>
<td>130</td>
</tr>
<tr>
<td>Contains in part what is expected</td>
<td>92</td>
<td>85</td>
</tr>
<tr>
<td>Does not match expectations</td>
<td>78</td>
<td>58</td>
</tr>
<tr>
<td>Does not offer information</td>
<td>35</td>
<td>16</td>
</tr>
<tr>
<td>No response</td>
<td></td>
<td>1</td>
</tr>
</tbody>
</table>
Improving the encoding of surveys based on the International Standard Classification of Occupations (ISCO) and the North American Industry Classification System (NAICS) is another relevant objective for INEGI, and we are actively leveraging data science methodologies to achieve it.

This recurring task demands a substantial allocation of human resources, presenting a challenge to the pursuit of enhanced quality and efficiency in the generation of statistical information.

Addressing this bottleneck is critical for ensuring the continued refinement of our survey encoding processes and ultimately optimizing the delivery of accurate and timely statistical insights.
Coding of Economic Statistics

Codification Process

- Deterministic algorithms
- Manual codification

Release Batch

Quality validation
- Success
- Failure
Coding of Economic Statistics
Coding of Economic Statistics

INEGI
- Survey
  - Descriptions
  - Metadata

Wikipedia

Pre-processing

FastText

Bidirectional GRU

Softmax Classifier

Confidence Scores
Coding of Economic Statistics

Previous

Deterministic 80%

Manual 20%

New scheme

Deterministic 80%

AI

Manual
• INEGI has long been committed to promoting **equality** and **inclusion** within the country, recognizing the importance of ensuring that every segment of society is represented and supported.

• We have embarked on a project aimed at analysing and identifying economic units that offer services specifically designed for women.

• By focusing on economic units that cater to the needs of women, INEGI aims to shed light on the essential role women play in the Mexican economy and society at large.
Gender Oriented Services Analysis

Datasets
- Semi-automatic segmentation
- Manual train dataset

Text Representation
- Tokenization
- TF-IDF
- Model

Classification
- Gender Classifier
- Service Classifier
- Final Class
The system has demonstrated a great reliability level in classifying economic units specialized in delivering services for women.

We have achieved an outstanding accuracy rate of 93% for categories such as reproduction and maternity, women’s health, family health, child health, childcare, and child development.
The Way Forward: Use of Large Language Models

• These projects exemplify the impact of applying language models to improve our processes and the quality of our products and services.

• Each project yields invaluable insights and crucial lessons that are now being leveraged to drive us to keep innovating.

• Now, we are proactively moving forward with projects applying Large Language Models (LLMs), and using the insights and methodologies gleaned from these endeavours into our workflows.
Thank You!

Questions?