Natural Language Processing for Economic Activity and Occupation
Goal

To research the extent to which NLP can be incorporated in the current coding production-process:

- Reduce the human-assisted workload
- Reduce the timing
- Maintain or improve the quality
- The hardest ones: Economic Activity, Occupation
Current Coding Process = Ground Truth

Entry batch (some records with standardized-text items) to be coded

Deterministic Coding in the form of:
"if this text is equal to… then …"

Subsample (proportional to class-size)

Human-coding made by an expert ("guru")

Assisted Coding = Human-made coding

Subsample (Acceptance sampling for every load humans make)

Contrasting Process:
In the case of deterministic coding, the quality must be 95%, if it is less, adjustments are made to the algorithms. For assisted coding, the sampling design ensures 90 percent.
NLP Stages

- Preprocessing
- Text Vectorization
- Automatic Coding
2018 National Household Income and Expenditure Survey (ENIGH). 158k coded records

Features:

- **Occupation:** ‘COCINERO DE ANTOJITO MEXICANO’
- **Task:** ‘PREPARAR VENDER DE ANTOJITO MEXICANO EN LOCAL’
- **Business activity:** ‘PREPARAR VENDER DE ANTOJITO MEXICANO EN LOCAL AL PUBLICO EN GENERAL’
- **Company name:** ‘HUARACHE MIMI’
- **13 additional items:** academic level, company size, ...

Variable to classify (either Economic Activity or Occupation): {7221}

- Hierarchical code: first two digits account for the sector
Stages

- Preprocessing
- Vectorization
- Automatic Coding
Preprocessing for ML Coding

» Ortographic correction

» Lemmatization

Based on our experience instead of a library (e.g. python)

Same process applied to both current coding and ML coding

» Stemming

» Q-grams (n-letter sequence) and N-grams (n-word sequence)
Stages

- Preprocessing
- Vectorization
- Automatic Coding
Two TF-IDF matrices (each one of 30,000 columns):
  › 2-word sequence
  › 6-letter sequence

Concatenate the two matrices into one of 60,000 columns + 13 auxiliary variables
Stages

- Preprocessing
- Vectorization
- Automatic Coding
Ensemble ML

» ML algorithms used:
  › SVM (best accuracy among all)
  › Logistic regression
  › Random Forest
  › Neural Networks
  › XGBoost
  › K-NN

» The final assigned code is made by a voting process:
  › Equal weights for each ML-method call
  › Unequal weights (try and error)
Two sets: Training (75%) and Test (25%). The following are based on the test set

Economic Activity (157 classes)

<table>
<thead>
<tr>
<th></th>
<th>Accuracy</th>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ensemble with equal weights</td>
<td>0.8905</td>
<td>0.6925</td>
<td>0.6149</td>
<td>0.6365</td>
</tr>
<tr>
<td>Ensemble with different weights</td>
<td>0.8921</td>
<td>0.6767</td>
<td>0.6420</td>
<td>0.6512</td>
</tr>
</tbody>
</table>

Occupation (461 classes)

<table>
<thead>
<tr>
<th></th>
<th>Accuracy</th>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ensemble with equal weights</td>
<td>0.8447</td>
<td>0.6441</td>
<td>0.5384</td>
<td>0.5639</td>
</tr>
<tr>
<td>Ensemble with different weights</td>
<td>0.8505</td>
<td>0.6437</td>
<td>0.5637</td>
<td>0.5831</td>
</tr>
</tbody>
</table>
Accuracy vs records coded automatically (percentage)
Accuracy vs number of records (log), by code
The first two digits for each code (class) represent the sector which the code belongs to:

- 7751, 7755: same sector

Definitions:

- Within-sector error: when the ML-algorithm codes a record it does it wrong, but the code it assigns does belong to the true sector: 4431 (real) vs 4434 (ML coding)
- Between-sector error: when the ML-algorithm codes a record it does it wrong, and also the code it assigns does not belong to the true sector: 4431 (real) vs 3178 (ML coding)
Percentage of misclassified records
Between vs Within Errors by Sector

Economic Activity
Economic Activity
Two-stage Automatic Coding

» First Stage:
  › Develop a SVM model on the training set in which we try to predict the Sector based on the textual features
  › Apply the above model to predict the Sector for the test set: predicted_sector

» Second Stage:
  › Develop a SVM model on the training set in which we try to predict the Class (4 digits) based on the following features: text + real_sector
  › Apply the above model to predict the Class for the test set, based on the following features: text + predicted_sector (first stage)

» We have not developed an ensemble algorithm for this two-stage approach
### Economic Activity

<table>
<thead>
<tr>
<th></th>
<th>Accuracy</th>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>One stage (6-grams, 2-words)</td>
<td>0.8793</td>
<td>0.6372</td>
<td>0.6760</td>
<td>0.6511</td>
</tr>
<tr>
<td>Two stages (6-grams, 2-words)</td>
<td>0.8774</td>
<td>0.6600</td>
<td>0.6452</td>
<td>0.6443</td>
</tr>
</tbody>
</table>

### Occupation

<table>
<thead>
<tr>
<th></th>
<th>Accuracy</th>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>One stage (6-grams, 2-words)</td>
<td>0.8188</td>
<td>0.5353</td>
<td>0.5918</td>
<td>0.5531</td>
</tr>
<tr>
<td>Two stages (6-grams, 2-words)</td>
<td>0.8312</td>
<td>0.5786</td>
<td>0.5730</td>
<td>0.5648</td>
</tr>
</tbody>
</table>
What is next

» Try state-of-the-art NLP methods for vectorizing
» Measure the performance on a new wave of records (ENIGH 2020)
» Develop a framework in which ML can be a part of the whole coding process and be subjected to quality measure
Conociendo México

01 800 111 46 34
www.inegi.org.mx
atencion.usuarios@inegi.org.mx

Facebook Instagram Twitter YouTube INEGI Informa