Enhancing Text Classification for Accurate Statistics: Leveraging LLM at the Ghana Statistics Service

Laurent Smeets and Josephine Baako-Amponsah
Ghana Statistical Service (GSS)
Why?

- Aim is to empower the GSS with a tool that extracts relevant information from open-ended questions.
- To make automate tedious and error prone work and to speed up the validation of field work by minimizing the need for manual coding.
- Get validated data public quicker.
- Build cost-efficient system that is easy to maintain.
When?

- ISIC during IBES
- ISCO and ISIC during GLSS
- HS code for cross border trade
- To reclassify “other, specify” questions into answer categories
- The first level of application of this project stage involves comparing the outputs of the trained model against codes assigned by field officers during data collection.
What do we want?

• A model that can automatically predict the ISIC code from open text
• Should be robust to typos
• This tool should use “English” as used by field enumerators and not United Nations English
• https://isic-prediction-27a3d67ecbf2.herokuapp.com/
What is in de box?

- Transformer model (2017)
- BERT (2018)
- DistillBERT (2019)
- Fine-tuned for ISIC prediction (2023)
A transformer model is a neural network that learns context and thus meaning by tracking relationships in sequential data like the words in this sentence.

longer-distanced context around a word (as compared to recurrent neural networks)
What is in de box?

(\textit{Distill})BERT model

- Bidirectional Encoder Representations from Transformers
- Transfer learning describes an approach where a model is first pre-trained on large unlabeled text corpora using self-supervised learning:
  - Pretrained on BookCorpus, a dataset consisting of 11,038 unpublished books and English Wikipedia
  - This makes it possible to create a model that is context aware:
    - "Server, can I have the check" vs. "Looks like I just crashed the server"
- Created for "next" word prediction
What is in de box?

**Fine-tuning**

- Annotated (cleaned) data from previous surveys added as a final Fine-tuned classification layer
- This way the model can learn ‘Enumerator’ English
- Model was trained on Nvidia GPU cards on Runpod.io (couple of USD) in Python
- Results are hosted on Hugging Face
How did the model perform?

- 431 different classes
- 86.5% Accuracy (80/20 split)
- RoBERTta: had an accuracy of 86.7%
- Weighted F1: 85.5%
- Differences between ISIC codes
- We observe that categories with relevant few observations have a lower accuracy than categories with a high number of observations.
What (didn’t work as well)

- Random forest: Model of 23GB for 83.8% accuracy
- Support Vector Machines: $O(N^3)$
- Creating code using gpt-3.5-turbo: Accuracy 59.1%.

"come up with 10 correct descriptions for establishments with ISIC code {code} is: {description} one example would be '{example}' each example can be maximally 100 characters long. try to make the example relevant for the types of businesses you might find in Ghana, without mentioning the word Ghana. dont mention the ISIC code in the output"
What? (is next)

• Getting the code to run on Android (using Tensorflow Lite)
• At the different stages of the project, the tool is shared with thematic experts for their feedback
• Write a paper with our results
• Put it into production
Useful links

- http://jalammar.github.io/illustrated-transformer/
- https://blogs.nvidia.com/blog/2022/03/25/what-is-a-transformer-model/
- https://mchromiak.github.io/articles/2017/Sep/12/Transformer-Attention-is-all-you-need/#.XiW1zBNKjOR
- https://www.youtube.com/watch?v=SZorAJ4IsA&ab_channel=GoogleCloudTech