

Knowledgebase refinement strategies to support classification of economic survey responses within official statistics production

Jamie Milsom¹, Luke Roantree¹

¹Central Data Science and AI, Office for National Statistics, UK

The ClassifAI App is used to classify responses to two key UK economic surveys, the Business Register and Employment Survey (BRES) and Annual Survey of Hours and Earnings (ASHE), against the UK's standard industrial and occupational taxonomies. Its open-source classification backend (the ClassifAI Package) is also in trial or use for several more surveys and taxonomies within the Office for National Statistics, and other National Statistics Organisations.

The classification approach used by these tools relies on the use of LLM encoder models to semantically link new survey responses to historical, coded examples (a 'knowledgebase') – and as such, accuracy of the classification is tied closely to curation of these historical examples. While a sufficient number and variety of historical examples are required for each distinct code, there is a strong risk of systematic misclassification patterns if incorrectly coded historical examples are present in the knowledgebase, which can result in a preference towards / away from some codes. In fact, even if historical examples are correctly assigned some code but are **semantically similar** to examples which fall under another code, this risk is present.

Currently manual review by expert clerical coders is used where possible to improve knowledgebase quality, a time consuming and expensive task, in addition to post-hoc analysis comparing distributions of assigned codes to past distributions. To mitigate these resource demands and enable proactive identification of 'problem' entries, the ClassifAI team have begun developing a 'Knowledgebase Refinement Tool' to curate a subset of historical examples which can be confidently used from a larger knowledgebase of unknown quality.

In this talk, we will discuss the iterative refinement approach we are developing towards flagging problem entries, using leave-one-out cross-validation, Retrieval Augmented Generation (RAG) with additional context about the classification taxonomy, and engagement with expert clerical coders to embed their heuristics and chains-of-thought into the data curation process.