API DE CODIFICACIÓN AUTOMÁTICA PARA LA PRODUCCIÓN DE ESTADÍSTICAS OFICIALES

Autores: Ignacio Agloni, Klaus Lehmann, Ricardo Pizarro

Esta presentación oral se propone dos objetivos principales. El primero, exponer las características y resultados de un modelo basado en aprendizaje profundo desarrollado en R. La implementación en producción de este modelo permite un importante ahorro para el Estado, en términos de recursos y oportunidad, al automatizar el proceso de codificación de glosas abiertas en encuestas. El segundo objetivo es mostrar una estrategia eficiente para desplegar este servicio de codificación a quienes requieran consumirlo, sin necesidad de realizar instalaciones en sus computadoras locales, mediante la disposición de una API. De forma transversal, en esta presentación se busca reflexionar acerca del rol que juega el Estado en la innovación pública y en la transformación digital, y sobre la importancia del desarrollo y uso de herramientas de código abierto para la consecución de estos desafíos.

Generalmente las encuestas incluyen preguntas que deben ser respondidas como textos abiertos, por la dificultad o imposibilidad de obtener la información deseada mediante preguntas con categorías predefinidas. Luego, a través de un proceso de codificación, se asigna a cada uno de estos textos una categoría del clasificador de referencia de acuerdo con su contenido[1].

En la principal institución productora de estadísticas en Chile, el Instituto Nacional de Estadísticas, este proceso ha sido implementado hasta la actualidad – salvo contadas excepciones - a través de un equipo de personas codificadoras que han sido entrenadas para la tarea de codificación manual a partir de un protocolo y criterios establecidos. Sin embargo, este proceso acarrea un costo en dinero y oportunidad: en la Encuesta Nacional de Empleo, hasta el 2018, este proceso utilizaba aproximadamente 3.600 horas persona mensualmente y en la VIII Encuesta de Presupuestos Familiares tuvo una dedicación aproximada de 2.640 horas persona mensuales, durante 18 meses. Ahora, ¿es mejor la calidad de la codificación manual? No necesariamente, pues se han observado niveles similares de precisión entre la codificación manual y la automatizada.

Desde el punto de vista de la eficiencia, la puesta en marcha de estos clasificadores puede generar un ahorro sustantivo de recursos para el INE, pues permitiría reducir de manera importante la cantidad de horas de trabajo destinada a esta tarea. Las ganancias de eficiencia se reflejarán en una disminución de costos laborales y en un menor tiempo de procesamiento de datos durante el periodo de relevamiento de las encuestas. Desde el punto de vista de la eficacia, el hecho de que la asignación de códigos descanse en las decisiones de distintas personas, hace bastante complejo el aseguramiento de una completa uniformidad en la aplicación de criterios. Si bien cada clasificador cuenta con un marco conceptual definido, la aplicación de criterios en la práctica no siempre resulta una tarea sencilla. La implementación de sistemas automáticos permite estandarizar los criterios aplicados.

El modelo que se expondrá en esta presentación se entrenó con glosas de la Encuesta de Nacional de Empleo (ENE) y de la prueba piloto de la IX Encuesta de Presupuestos Familiares (EPF, 2020). Los textos

[1] Algunos de los clasificadores internacionales más utilizados en las encuestas de hogares son el de ocupaciones (CIUO), rama de actividad económica (CIIU), y gastos de consumo final (CCIF).

corresponden a descripciones breves de ocupaciones de las personas entrevistadas y de la rama de actividad económica a la que pertenecen las empresas donde ellas trabajan. Estas glosas fueron registradas por encuestadores y encuestadoras mediante dispositivos móviles de captura (tablets), y luego fueron etiquetadas manualmente, en el marco de un proyecto estratégico de la institución que buscaba generar bases de datos de entrenamiento con foco en la calidad de las clasificaciones realizadas.

Tanto en el caso de ocupaciones como en el de rama de actividad económica se trata de problemas de clasificación multiclase. Existen 41 categorías de ocupación, a nivel de subgrupo principal, y las actividades económicas se desagregan en 81 categorías, a nivel de división.

Se probaron dos arquitecturas de redes neuronales: *feed-forward* y *gated recurrent unit* (GRU). Adicionalmente, se testeó la utilización de dos estrategias de vectorización de textos: *word embeddings* y TF-IDF. Al combinar ambas estrategias (arquitectura y representación vectorial) se crearon 4 modelos, dentro de los que se seleccionó el de arquitectura GRU con *word-embeddings* para el modelo en producción. El *accuracy* aproximado para el clasificador de actividad económica es de 93% y de 90% para el de ocupación. Es preciso mencionar que el motivo por el cual se escogió dicha arquitectura no solo tiene que ver con las métricas finales de evaluación, sino que con la posibilidad que ofrece la estrategia de *word-embeddings* para hacer el modelo más flexible a cambios en la forma de describir a las ocupaciones y actividades económicas.

Para poner en producción el modelo se detectaron dos obstáculos principales. El primero de ellos guarda relación con las dependencias y herramientas requeridas para poder utilizar el modelo. Una estrategia basada en poner simplemente los archivos de los modelos a disposición de los usuarios y usuarias tiene una serie de limitaciones, ya que supone que las personas serán capaces de instalar correctamente una serie de componentes y librerías necesarias de forma local (Python, reticulate, tensorflow, keras, entre otras). El segundo obstáculo se vincula con la reproducibilidad, pues incluso en el caso de que los y las usuarias lleven a cabo correctamente las instalaciones mencionadas, es posible que no se obtengan resultados idénticos en dos computadoras diferentes, lo cual no es algo deseable en el contexto de producción de estadísticas oficiales.

A raíz de los problemas mencionados anteriormente, se pensó en un servicio que permita a las personas utilizar los modelos, sin necesidad de hacer instalaciones ni manejar herramientas diferentes a las utilizadas usualmente en el Instituto Nacional de Estadísticas. Para ello, se utilizó el paquete plumber, el cual permite crear APIs de manera relativamente sencilla. Así, los archivos y modelos fueron almacenados en un servidor Linux (con la distribución Centos) y mediante la API pueden ser utilizados por los usuarios y usuarias, quienes envían al servidor una glosa y reciben como respuesta una predicción, que es la categoría o clase asignada a dicha glosa; esto ya sea con glosas únicas o de forma masiva sobre listas de glosas. Mayores detalles sobre la implementación serán revisados durante la presentación.

Actualmente, la política del Estado de Chile no discrimina positiva o negativamente el uso de *software* libre y de código abierto. Aún hoy en día existen pocas iniciativas de *software* abierto desde el Estado, a pesar del importante avance al que se vio forzado el Estado en materia de transformación digital durante la pandemia por Covid-19. Desarrollos como el que se expone en el marco de esta presentación oral, se suman a una serie de proyectos de código abierto realizados desde el Instituto Nacional de Estadísticas en diversas temáticas, que buscan poner al Instituto al servicio de la comunidad y junto a la comunidad usuaria de estadísticas oficiales.