Aprendizaje automático para el estudio de eventos políticos: Aplicación de modelos semisupervisados en archivos de América Latina

Bastián González-Bustamante

Palabras clave: aprendizaje automático, reconocimiento óptico, modelos semisupervisados, eventos políticos, América Latina.

Abstract

Este trabajo presenta la creación de indicadores de estabilidad política en 12 países latinoamericanos desde mediados de la década de 1970 hasta la fecha. Nos centramos en la creación del conjunto de datos utilizando algoritmos de reconocimiento óptico sobre archivos e informes de prensa junto con modelos de aprendizaje automático. Los modelos entrenados consisten en un conjunto de clasificadores supervisados que nos permiten ensamblar modelos semisupervisados en un período de casi 50 años. Posteriormente, presentamos una validación cruzada comparando nuestros indicadores con otros similares y también presentamos dos aplicaciones prácticas que proporcionan una visión general de la estabilidad de los gobiernos en la región durante el período y temas clave de la agenda política. La discusión resume el proceso de entrenamiento de algoritmos y reflexiona sobre la contribución empírica-metodológica de este trabajo y su potencial impacto.

Creación del conjunto de datos

Se crea un novedoso conjunto de datos con indicadores de cuestionamientos públicos al gobierno y dimisión de ministros a partir de una revisión de prensa y archivos. Para esto se identificaron las menciones específicas con un algoritmo de Optical Character Recognition (OCR) programado para tal efecto con base en el motor Tesseract, cuya primera versión se desarrolló en Bristol a mediados de la década de 1990 y, desde 2005, es de código abierto (Smith, 2007). Tesseract ofrece soporte Unicode para un centenar de idiomas y varios lenguajes de programación (Ooms, 2021).

El proceso de reconocimiento óptico se realizó sobre más de 28 mil páginas de archivos. La aplicación del algoritmo nos permitió construir un corpus y aplicar criterios de búsqueda específicos para elaborar nuestros indicadores. Este proceso nos permitió filtrar las notas de prensa individuales, asociarlas a un país y a un gobierno y, a continuación, filtrar por el periodo.

Modelos de aprendizaje automático

Entrenamos diferentes modelos de aprendizaje automático supervisado y semisupervisado para identificar específicamente cuestionamientos al presidente y miembros del gabinete y distinguir estos eventos de las menciones identificadas con el proceso de OCR. La Figura 1 ilustra el proceso de entrenamiento de los algoritmos que incluye varios pasos, desde la división de la submuestra aleatoria de 1.000 reportes para el etiquetado manual hasta la predicción. Nos centramos en el entrenamiento de modelos naive Bayes, Support Vector Machine (SVM), Random Forest con 100 y 500 árboles y XGBoost. La precisión de nuestros modelos supervisados oscila entre un 72 y 76%. Considerando que para mejorar nuestros algoritmos necesitaríamos una cantidad significativa de datos adicionales etiquetados manualmente y que los archivos abarcan un período de casi medio siglo, por tanto, la precisión de los algoritmos tiende a decaer fuera de la muestra y en las comparaciones entre años, entrenamos modelos semisupervisados utilizando los algoritmos anteriores y un conjunto de *seed words* relevantes teóricamente y aquellas identificadas en nuestros modelos supervisados. Esto nos permitió ensamblar modelos con una precisión entre el 80 y 90% dependiendo del año.