US Congress Clasificación de Proyectos por Temas Mediante Máxima Entropía

Dany Efrain Rubiano Jiménez
Universidad de Santiago de Chile

Abstract. US Congress Bills es un conjunto de datos de muestra que contiene proyectos de ley del Congreso de los Estados Unidos, etiquetados y compilados por el profesor John D. Wilkerson de la Universidad de Washington, y E. Scott Adler de la Universidad de Colorado. En esta oportunidad se hace uso del método de máxima entropía, buscando obtener conocimiento acerca de los temas de los proyectos para su clasificación.

Keywords: máxima entropía, proyectos de ley, clasificación de textos, precisión.

1 Introducción

El análisis de textos dadas las características propias del lenguaje representa un problema complejo que no es posible ser abordado a través de métodos comunes. Dado esto, dentro del ámbito de la minería de datos, se contempla el uso de la máxima entropía.
Para el desarrollo de la presente experiencia, se analiza un conjunto de datos con información procedente del congreso de los Estados Unidos, con alrededor de 4449 proyectos de ley etiquetados según el contenido y destino propio del proyecto. El análisis se hace finalmente mediante el método de máxima entropía para lo cual se busca:

- Realizar pre-procesamiento de texto para la aplicación del método.
- Realizar proceso de calibración de parámetros del método de máxima entropía, para luego evaluar el rendimiento del algoritmo mediante índices de precisión y exhaustividad, definiendo un subconjunto de categorías "relevantes" para dicho propósito.

2 Métodos y datos

2.1 Métodos utilizados

El método utilizado corresponde al de máxima entropía, que está basado en el caculo de una función de probabilidad capaz de maximizar la entropía de la probabilidad a posteriori.
Antes de proceder en la aplicación del método, se debe realizar primero un preprocesamiento del conjunto de datos, para luego dividir el conjunto en dos muestras para entrenamiento y prueba. La muestra para el entrenamiento equivale al 70% y la prueba el 30%.

Una vez hecho esto, se seleccionan ciertos temas considerados relevantes, a fin de aplicar el método de máxima entropía. Posteriormente, se realiza un modelo por cada muestra a partir de la validación cruzada y la calibración de los parámetros.
A partir de los resultados, se evalúa el rendimiento del método mediante índices de precisión y exhaustividad.
Todo el proceso es reiterado un cierto número de veces, variando el preprocesamiento y balanceando los datos según las etiquetas a través de SMOTE

2.2 Datos utilizados

El conjunto de datos de los proyectos de ley del congreso de los Estados Unidos. Contiene 4449 instancias en total y presenta los siguientes atributos:

- ID: un identificador único para el proyecto de ley.
- cong: la sesión del congreso en la que apareció por primera vez el proyecto.
- billnum: el número del proyecto tal como aparece en el expediente del Congreso.
- h_or_sen: Un campo que especifica si el proyecto de ley fue presentado en la Cámara de Representantes (HR) o el Senado (S).
- major: un código de tema etiquetado manualmente que corresponde al tema del proyecto.

Los temas abordados por el atributo *major* se describen en la tabla a continuación.

Tabla 2.1: *Descripción de los temas de los proyectos.*

Etiqueta	Descripción	Etiqueta	Descripción
1	Rigths and liberties	13	Social
2	Economie	14	Urbanization
3	Health	15	Insurance and regulations
4	Agriculture and livestock	16	Defense
5	Inmigration	17	Telecomunications
6	Education	18	Trade
7	Environment	19	International
8	Energy	20	Goverment
10	Transportation	21	Parks and lands
12	Law and crime	99	Private bills

En lo que respecta a la cantidad de instancias por cada tema la tabla 2.2 lo refleja a modo resumen.

Tabla 2.2: *Cantidad de proyectos de ley por tema.*

Etiqueta	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**10**	**12**
Cantidad	163	84	617	133	262	222	201	138	171	291
Porcentaje %	3,66	1,89	13,87	2,99	5,88	4,99	4,51	3,10	3,84	6,54
Etiqueta	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**99**
Cantidad	94	80	279	219	90	402	121	380	472	30
Porcentaje %	2,11	1,80	6,27	4,92	2,02	9,03	2,72	8,54	10,61	0,67

El preprocesamiento de los datos abordado consiste en convertir las palabras a minúscula, remover espacios en blanco, números y puntuación, y eliminar los stopwords, palabras irrelevantes.
Para reflejar el preprocesamiento, se presentan a continuación un análisis de frecuencia de palabras del conjunto de datos con y sin preprocesamiento.

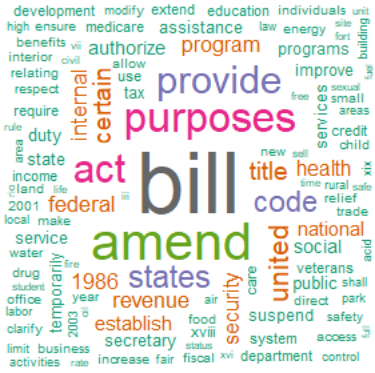

Figura 2.1*: Nube de palabras del conjunto de datos sin preprocesamiento.*

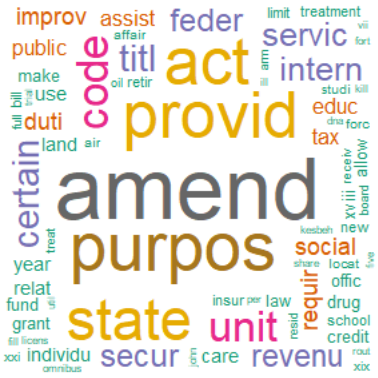

Figura 2.1*: Nube de palabras del conjunto de datos con preprocesamiento.*

3 Resultados

Cabe destacar que los temas elegidos como relevantes son Health, Inmigration, Law and crime, Insurance and regulations, Trade, Goverment, y Parks and lands,.

La primera prueba del método es en base a el conjunto de datos sin preprocesamiento, para lo cual la calibración de parámetros a través de la validación cruzada es la que se detalla a continuación.

Tabla 3.1. *Calibración de parámetros del modelo con los datos sin preprocesamiento.*

L1_regularizer	**L2_regularizer**	**Use_sgd**	**Set_heldout**	**Accuracy**	**Pct_besst_fit**
0	0,2	0	0	0,7133	1

Es importante tener en cuenta la cantidad de los datos considerados relevantes para el conjunto de prueba, correspondiendo a 822 instancias, mientras que las no relevantes corresponden a 513.

Tabla 3.2. *Resultados del modelo con el conjunto de datos sin preprocesamiento.*

	Recuperados	**No Recuperados**
Recuperados	734	88
No Relevantes	64	449

Precisión	0,6204
Recall	0,8929
F1	0,7321

A partir de los resultados anteriores, se procede a aplicar el modelo previo preprocesamiento del conjunto de datos. En este caso, la cantidad de datos considerados relevantes para el conjunto de prueba corresponden a 817 instancias y los no relevantes son 518.

Tabla 3.3. *Calibración de parámetros del modelo con los datos con preprocesamiento.*

L1_regularizer	**L2_regularizer**	**Use_sgd**	**Set_heldout**	**Accuracy**	**Pct_besst_fit**
0	0,4	0	0	0,7258	1

Tabla 3.4. *Resultados del modelo con el conjunto de datos con preprocesamiento.*

	Recuperados	**No Recuperados**
Recuperados	499	318
No Relevantes	179	339

Precisión	0,5954
Recall	0,6107
F1	0,6030

Dado que los datos por tema tienen cierto desbalance, se aplica SMOTE buscando obtener así mejores resultados del modelo.
La nueva configuración de temas se presenta en la tabla 3.5, teniendo ahora 802 datos para el conjunto de prueba.

Tabla 3.5. *Conjunto de datos una vez aplicado SMOTE*

Etiqueta	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**10**	**12**
Cantidad	76	43	331	79	146	142	111	71	99	153
Porcentaje %	2,85	1,61	12,40	2,96	5,47	5,32	4,16	2,66	3,71	5,73
Etiqueta	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**99**
Cantidad	50	42	153	103	41	231	71	218	240	270
Porcentaje %	1,87	1,57	5,73	3,86	1,54	8,65	2,66	8,16	8,99	10,11

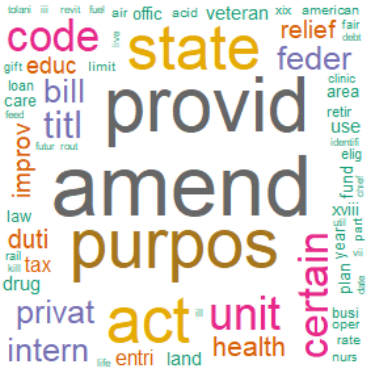

Figura 3.1. *Nube de palabras del conjunto de datos una vez aplicado SMOTE..*

Tabla 3.6. *Calibración de parámetros del modelo con los datos una vez aplicado SMOTE..*

L1_regularizer	L2_regularizer	Use_sgd	Set_heldout	Accuracy	Pct_besst_fit
0	0,2	0	0	0,7498	1

Tabla 3.7. *Resultados del modelo con el conjunto de datos una vez aplicado SMOTE.*

	Recuperados	No Recuperados
Recuperados	339	92
No Relevantes	118	253

Precisión	0,5726
Recall	0,7865
F1	0,6627

4 Discusión

Generalmente los proyectos de ley tienen palabras similares entre los diversos temas contenidos, ejemplo de ello se puede evidenciar en los proyectos de ley referidos a un tema internacional, que a su vez pueden abordar subtemas como comercio, salud, educación, entre otros, por lo cual presentan palabras similares entre sí. Así mismo, dentro de la mayoría de los temas se presentan casos de enmiendas a diversas leyes y regulaciones, aumentando la similitud de palabras entre los temas. Esto se puede evidenciar en los gráficos de nubes de palabras donde además se encuentran que las palabras de más alta frecuencia son propósito, proporcionar, acta, estado, seguridad, salud, entre otros, muchas transversales a varios temas.

Todo lo mencionado anteriormente son condicionantes para los resultados obtenidos para cada uno de los modelos, dependiendo en gran manera de la selección realizada para los documentos relevantes.

Resulta llamativo que los resultados del conjunto de datos sin preprocesamiento sean mejores en precisión y exhaustividad que los resultados sobre el conjunto de datos con preprocesamiento y al conjunto de datos con SMOTE. Esto puede ser debido al mismo preprocesamiento, específicamente al quitar las raíces de las palabras, se refleja pérdida de información. Además esto se puede deber a que los temas escogidos

como relevantes son los más frecuentes, y por lo tanto, se refleja el desbalance de los datos.
En una comparativa de los resultados con preprocesamiento y después de SMOTE, estos reflejan cierta similitud, dado que SMOTE es aplicado a los datos previamente preprocesados, a modo de no propagar ruido a las nuevas instancias generadas. Además, esto se puede deber a que los temas sobre-muestreados por la técnica, no afectan en gran manera a los temas considerados relevantes.

5 Conclusiones

Una vez desarrollada la experiencia, se puede reflejar la gran capacidad del método de máxima entropía para la extracción de conocimiento de textos, y cómo en su aplicación existe alta sensibilidad al preprocesamiento. Dicho preprocesamiento depende altamente del contexto de los textos, a manera de distinguir aquello que no aporta información y que por lo tanto corresponde a ruido.
Los resultados dependen claramente de lo que se considera como documentos relevantes. En esta ocasión se optó por elegir aquellos con las frecuencias más altas por tema, sin embargo, se podrían poder adoptado otras técnicas como la similitud entre temas, a través de un análisis más exhaustivo que mejore el contexto de clasificación.

6 Referencias

1. E. Adler, J. Wilkerson. Congressional Bills Project. University of Washington, 2004. URL http://www. congressionalbills.org/
2. Timothy P. Jurka and Yoshimasa Tsuruoka (2013). maxent: Low-memory Multinomial Logistic Regression with Support for Text Classification. R package version 1.3.3.1. https://CRAN.R-project.org/package=maxent
3. Chacón, M (2017) Taller de minería de datos avanzada, Capítulo 3: Máxima entropía.