# US Congress, clasificación de proyectos de ley por temas utilizando Máxima Entropía

Maximiliano Arévalo Sáez and Benjamín Muñoz Tapia

Universidad de Santiago de Chile

**Abstract.** La base de datos US Congress corresponde a un conjunto de datos que contiene diversos proyectos de Ley de Estados Unidos, los cuales fueron compilados y etiquetados por el profesor John D. Wilkerson de la Universidad de Washington junto con E. Scott Adler de la Universidad de Colorado. Por lo que en el presente documento se tiene como objetivo utilizar el método de Máxima Entropía, con el objetivo de analizar estos proyectos de ley para recuperar y clasificar documentos con el método indicado.

**Keywords:** Máxima Entropía · procesamiento de Lenguaje Natural · clasificación · proyectos de ley

## 1 Introducción

Los problemas de clasificación suelen ser utilizados para grupos que contengan variables numéricas, pero también pueden aplicados a problemas de clasificación de texto debido a la complejidad de este análisis al utilizar métodos comunes y tradicionales. Debido a las características propias del lenguaje natural, por lo que es necesario recurrir a métodos existentes dentro de la minería de datos para enfrentar estos problemas.

El procesamiento del lenguaje natural se utiliza hoy en día para diversas aplicaciones como lo es el análisis de sentimientos, estrategias de marketing, entre otros. Es por esto que se tiene como objetivo la aplicación del método de Máxima Entropía como técnica de clasificación para texto natural utilizando la base de datos US Congress, que cuenta con 4449 proyectos de ley de Estados Unidos etiquetados en diversas categorías. Por lo que los objetivos del presente trabajo corresponden a:

- Analizar la base de datos US Congress y realizar un pre-procesamiento al texto.
- Encontrar la mejor configuración para el método de Máxima Entropía.
- Evaluar los resultados del algoritmo utilizando las métricas *precision, recall y F1*

# 2 Métodos y datos

## 2.1 Métodos utilizados

El método a aplicar es el Método de Máxima Entropía, el cual está basado en el cálculo de una función de probabilidad capaz de maximizar la entropía de la probabilidad a posteriori. Para ser aplicado al lenguaje natural, se debe realizar primero un pre-procesamiento al texto, como la eliminación de *stopwords*, estandarización de palabras, entre otros, con el fin de obtener un texto que pueda ser analizado.

Luego de esto, se definen los temas considerados relevantes en base a la cantidad de proyectos de ley orientados hacia esos temas, bajo el criterio de que su importancia recae en la necesidad de proponer constantemente reformas para estos temas en particular. Para así aplicar el método de Máxima Entropía y realizar un modelo de las muestra utilizando validación cruzada junto con la calibración de los parámetros.

Finalmente es evalúa el rendimiento del método utilizado en base a las métricas correspondientes a *precision, recall y F1*,y luego se realiza un balancero de los datos en base a sus etiquetas utilizando la función *smote*.

## 2.2 Datos utilizados

En este caso, la base de datos US Congress posee 4449 instancias de proyectos de ley del congreso de los Estados Unidos, los cuales están divididos en 22 categorías. Entre los atributos del dataset se presentan los siguientes:

- ID: Identificador único del proyecto de ley.
- cong: Sesión del Congreso en la que apareció el proyecto de ley por primera vez.
- billnum: Número del proyecto de ley registrado en el Congreso.
- h_or_sen: Especifica si el proyecto de ley fue presentado en la Cámara de Representantes (HR) o en el Senado (S).
- major: Etiqueta asignada manualmente que corresponde al tema del proyecto de ley.

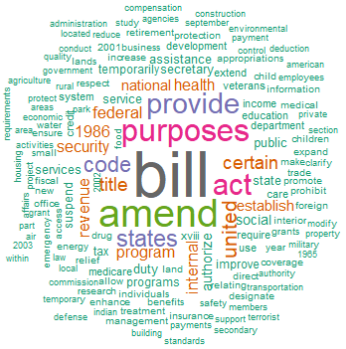
Para conocer los distintos temas abordados, se presenta a continuación una tabla que muestra los distintos valores de *major*:

| Etiqueta | Descripción | Etiqueta | Descripción |
|---|---|---|---|
| 1 | Rights and liberties | 13 | social |
| 2 | Economy | 14 | Urbanization |
| 3 | Health | 15 | Insurance and regulations |
| 4 | Agriculture and livestock | 16 | Defense |
| 5 | Immigration | 17 | Telecommunications |
| 6 | Education | 18 | Trade |
| 7 | Environment | 19 | International |
| 8 | Energy | 20 | Government |
| 10 | Transportation | 21 | Parks and land |
| 12 | Law and crime | 99 | Private bills |

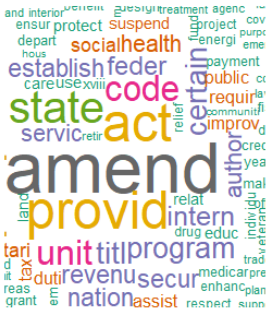
Table 1: Temas de proyectos de ley en el dataset US Congress

En cuanto al pre-procesamiento se tienen en cuenta la eliminación de *stop-words*, espacios en blanco, estandarización de palabras, números, puntuación y la palabra *bill*, ya que puede influir en la clasificación al contar con esta palabra como descripción de cada proyecto de ley. Esto se refleja en los siguientes *wordclouds*:

subcaption




(a) Wordcloud normal




(b) Wordcloud pre-procesado

Finalmente, para seleccionar los temas relevantes, se analiza la frecuencia de cada uno de estos como se muestra en la siguiente tabla:

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

Table 2: Cantidad y porcentaje de aparición de temas, US Congress

Por lo que los temas relevantes para este caso, corresponden a *Health*, *Immigration*, *Law and crime*, *Trade*, *Government* y *Parks and land*.

## 3 Resultados

Al realizar la aplicación del método de Máxima Entropía sobre el dataset original, se obtienen los siguientes resultados al realizar la calibración de parámetros a través de la validación cruzada:

| **L1_regularizer** | **L2_regularizer** | **Use_sgd** | **Set_heldout** | **Accuracy** | **Pct_bestfit** |
|---|---|---|---|---|---|
| 0.0 | 0.2 | 0 | 0 | 0.7409 | 0.9987 |

Table 3: Configuración dataset original

Por lo que los resultados del modelo con el conjunto de datos mencionado anteriormente se puede visualizar en la siguiente tabla, en la cual se puede visualizar que existe un mayor número de documentos recuperados relevantes:

| | Relevantes | No Relevantes |
|---|---|---|
| Recuperados | 1740 | 1446 |
| No Recuperados | 164 | 209 |

Table 4: Resultados del método

| Precision | Recall | F1 |
|---|---|---|
| 0.546139 | 0.91386 | 0.68369 |

Table 5: Valores de métricas

Al realizar el pre-procesamiento y seleccionando los temas relevantes, se realiza la calibración de parámetros a través de la validación cruzada indicada a continuación:

| **L1_regularizer** | **L2_regularizer** | **Use_sgd** | **Set_heldout** | **Accuracy** | **Pct_bestfit** |
|---|---|---|---|---|---|
| 0.0 | 0.2 | 0 | 0 | 0.7446 | 1 |

Table 6: Configuración dataset pre-procesado

En este caso, al obtener los resultados del modelo utilizando el conjunto de prueba, se pueden visualizar resultados que evidencian un problema de clasificación para los documentos recuperados que son relevantes, debido a que se observa que un gran cantidad de estos no pueden ser recuperados:

| | Relevantes | No Relevantes |
|---|---|---|
| Recuperados | 128 | 97 |
| No Recuperados | 281 | 34 |

Table 7: Resultados del método

| Precision | Recall | F1 |
|---|---|---|
| 0.568888 | 0.3129 | 0.4037 |

Table 8: Valores de métricas

## 4 Discusión

A través de los resultados se observa que hay una gran cantidad de documentos recuperados, dando a ver así la buena aplicación del modelo. En este caso se tuvo en cuenta una probabilidad del 55% para considerar un documento como recuperado, y entre estos, una gran cantidad son relevantes, lo que se debe a que los temas de proyecto seleccionados como relevantes son los que están más presentes en el Dataset. Por otro lado, se debe tener en cuenta por la naturaleza del Dataset que los proyectos de ley tienen un lenguaje determinado y en general usan palabras bastante similares, por lo que distintos temas pueden tener palabras parecidas, lo cual se refleja en las nubes de palabras obtenidas, aparenciendo con mayor frecuencia las palabras: *amend, provid, act, state, code, entre otras*. Por esto, se necesitaría un modelo más preciso para tener mejores resultados de clasificación.

En cuanto a las métricas obtenidas, el tener una precisión sobre el 50% es regularmente bueno, pero entre estos documentos se tiene un 91,38% de recall

para el conjunto de datos sin procesar, lo que indica una buena detección de documentos recuperados en este conjunto. Al analizar ambas métricas en conjunto, el tener una precisión relativamente alta y también un alto recall deja a ver que el modelo clasifica bien, pero sí hay elementos de otras clases que pueden influir en los resultados. Sin embargo, al utilizar el conjunto de entrenamiento, los resultados disminuyen drásticamente, ya que se utiliza solo un 20% del dataset y solo hay 6 temas considerados como relevantes de los 22 totales, y al tener una menor cantidad de documentos totales habrán menos ocurrencias de recuperación y relevancia.

## 5 Conclusiones

Se puede ver un cumplimiento de los objetivos específicos del trabajo, dando como resultado un buen cumplimiento del objetivo general del trabajo, dado a la aplicación exitosa del método de Máxima Entropía a través de un análisis de texto y procesamiento del lenguaje natural.

Los resultados por su parte, dependen de la persona que esté interesada en el problema de clasificación, ya que se tiene que definir manualmente para este caso qué temas son relevantes o no en las cuentas del congreso. Para las pruebas realizadas se escogieron los temas de mayor frecuencia en el dataset, aunque podrían haber sido otros como los que van más hacia la política y economía, otros como gobierno y legislación, etc.

## References

1. Wilkerson, J., Scott, E. (2004). USCongress: a sample dataset containing labeled bills from the United States Congress. https://rdrr.io/cran/RTextTools/man/USCongress.html.
2. Maxent function in R. https://www.rdocumentation.org/packages/dismo/versions/1.3-3/topics/maxent.