Universidad del Valle de Guatemala

Facultad de Ingeniería

Data Science

Proyecto 2. Resultados Parciales y Visualizaciones Estáticas

Mariana David 201055

Angel Higueros 20460

Fredy Velasquez 201011

Javier Valle 20159

Mario De León

Guatemala, 2023

Investigación de algoritmos

1. Naive Bayes

   El algoritmo clasificador Naïve-Bayes (NBC) es una técnica ampliamente utilizada en el campo de la clasificación de datos, conocida por su simplicidad y eficacia en una variedad de aplicaciones. Aunque asume una hipótesis de independencia entre las características, lo que a veces puede ser una suposición simplificada, el NBC a menudo proporciona resultados de clasificación precisos, especialmente en tiempo real y con conjuntos de datos de entrenamiento relativamente pequeños.

   Ventajas del Modelo Naive Bayes:

   - Facilidad de Implementación: Una de las principales ventajas del NBC es su facilidad de implementación y aplicación. Es un algoritmo que puede adaptarse a diversas situaciones de clasificación, incluyendo la categorización de texto y el etiquetado de documentos.

   - Requisitos de Datos de Entrenamiento Bajos: El NBC sobresale en situaciones en las que los datos de entrenamiento disponibles son limitados. A menudo, es capaz de proporcionar resultados sólidos incluso cuando la recopilación de datos es costosa o limitada.

   - Simplificación de la Preparación de Datos: Comparado con algoritmos más complejos, el NBC reduce la carga de trabajo asociada con la preparación de datos. Esto implica menos esfuerzo y recursos dedicados a la limpieza y procesamiento de datos.

   Desventajas del Modelo Naive Bayes:

   - Suposición de Independencia: Una de las desventajas notables del NBC es su suposición de independencia entre las características utilizadas para la clasificación. En situaciones de la vida real, esta independencia a menudo no se cumple completamente, lo que puede llevar a sesgos en los resultados.

   - Problema de Categorías no Observadas: Si una variable categórica en el conjunto de prueba contiene una categoría que no se observó en el conjunto de entrenamiento, el NBC asignará una probabilidad de 0 a esa categoría. Para abordar esta limitación, es necesario agregar datos adicionales a los conjuntos de entrenamiento.

2. Multinomial Naive Bayes

   El algoritmo clasificador Multinomial Naive Bayes (MNB) es una extensión del Naive Bayes, diseñado específicamente para enfrentar situaciones donde las características representan recuentos o frecuencias. Es muy útil en tareas de clasificación de documentos y análisis de texto, donde las características pueden ser la frecuencia de palabras o frases en el documento.

Ventajas del Modelo Naive Bayes según Sriram (2022):

- Facilidad de Implementación: Una de las características distintivas del NBC es su sencillez en cuanto a la implementación. El algoritmo es más sencillo de utilizar porque, según Sriram (2022), solo es necesario calcular la probabilidad.

- Versatilidad con los tipos de datos: NBC tiene una gama más amplia de aplicaciones porque se puede aplicar tanto a datos continuos como discretos, a diferencia de otros algoritmos que pueden estar limitados en el tipo de datos con los que trabajan.

- Sriram (2022) afirma que el NBC es apropiado para aplicaciones en tiempo real, lo que implica que puede procesar y categorizar datos a medida que se generan.

- Escalabilidad: la capacidad de NBC para manejar grandes conjuntos de datos es una de sus principales ventajas. Es adecuado para situaciones en las que el volumen de datos puede resultar abrumador para otros algoritmos debido a su naturaleza probabilística y simplicidad de implementación.

Desventajas del Modelo Naive Bayes según Sriram (2022):

- Precisión limitada: según Sriram (2022), la precisión predictiva de NBC puede ser menor que la de otros algoritmos probabilísticos. Puede que no siempre sea la mejor opción debido a esta restricción, especialmente si la precisión es crucial.

- No Adecuado para Regresión: A pesar de su utilidad en la clasificación, el NBC no es adecuado para tareas de regresión. Esto significa que no se puede utilizar para predecir valores numéricos, lo que restringe su uso en algunos campos.

- Supuesto de independencia: como se mencionó anteriormente, la NBC supone que los rasgos utilizados para clasificar algo no están relacionados entre sí. Esta suposición, frecuentemente, no se cumple en situaciones reales y puede afectar la precisión del modelo.

3. Gaussian Naive Bayes

El algoritmo de Gaussian Naive Bayes es una técnica de clasificación que se usa bastante en Aprendizaje Automático, el cual se basa en el enfoque probabilístico y la distribución gaussiana. El Gaussian Naive Bayes hace la suposición de que cada parámetro tiene una cierta capacidad independiente para poder predecir la variable de salida. La combinación de las predicciones para todos los parámetros es la predicción final, que devuelve una probabilidad de que la variable dependiente sea

clasificada en cada grupo. La clasificación final se asigna al grupo con la probabilidad más alta.(Martins, 2022)

Ventajas del Gaussian Naive Bayes

- Se puede usar en tiempo real para realizar predicciones.
- Es bastante preciso para realizar los problemas de clasificación en donde las características siguen una distribución normal.
- Es bastante fácil de entender y de implementar.(S, 2022)

Desventajas del Gaussian Naive Bayes

- La suposición de independencia de las características que se están clasificando puede ser inexacta.
- El algoritmo puede ser sensible a valores atípicos.
- Presenta un mal manejo de características numéricas continuas dentro de un dataset. (S, 2022)

Selección de algoritmos a probar

Se seleccionó el uso de naive bayes por las siguientes razones:

- Naive Bayes es conocido por ser uno de los algoritmos más simples en cuanto a clasificación, especialmente cuando se trata de texto. A pesar de su simplicidad, ha demostrado ser altamente efectivo en muchas aplicaciones reales, como el filtrado de spam en correos electrónicos y la categorización de documentos.

- Una de las ventajas principales de Naive Bayes es que se puede implementar con relativa facilidad. No requiere de un preprocesamiento extenso de los datos, y el proceso de entrenamiento es directo. Esta característica es especialmente valiosa en proyectos donde el tiempo es una limitante o cuando se busca una solución rápida y eficiente sin entrar en algoritmos más complejos.

- Contar con experiencia previa en el uso de una herramienta o técnica es una ventaja significativa en cualquier proyecto. Dentro del equipo ya se cuenta con miembros que han trabajado con Naive Bayes anteriormente. Esta experiencia permite evitar curvas de aprendizaje empinadas y aprovechar conocimientos previos para afinar y optimizar el modelo

Modelos

1. Preparación:

- Se importan las bibliotecas necesarias para el análisis, incluyendo tm para el procesamiento de texto, SnowballC y wordcloud para visualizaciones, RColorBrewer para colores, e1071 para implementar Naive Bayes y caret y gmodels para análisis y comparación de resultados.
- Se lee un conjunto de datos desde 'train.csv' y se seleccionan solo las columnas 3 y 5, renombradas como 'Msg' y 'Tag'.
- El campo 'Tag' es convertido en un factor.

2. Procesamiento del Corpus:

- Se crea un corpus a partir de la columna 'Msg' y se visualizan los primeros 5 mensajes.
- Se realiza una limpieza del corpus, transformando todo a minúsculas, eliminando números, palabras comunes (stop words) en inglés, puntuaciones y espacios en blanco adicionales.
- Se crea una matriz de términos del documento (DocumentTermMatrix) a partir del corpus limpio.

3. Preparación de Conjuntos de Entrenamiento y Prueba:

- El conjunto de datos se divide en un 70% para entrenamiento y un 30% para prueba.
- Se obtienen las proporciones de las etiquetas tanto para el conjunto de entrenamiento como para el conjunto de prueba.

4. Selección de Términos Frecuentes:

- Se identifican palabras que aparecen con frecuencia (al menos 5 veces) en el conjunto de entrenamiento y se seleccionan solo estas palabras frecuentes tanto para el conjunto de entrenamiento como para el conjunto de prueba.

5. Creación y Evaluación del Modelo Naive Bayes:

- Se crea un clasificador Naive Bayes utilizando el conjunto de entrenamiento.
- Se hace una predicción utilizando el conjunto de prueba y se evalúa la precisión de las predicciones utilizando una tabla cruzada (CrossTable).

6. Creación y Evaluación de un Modelo Mejorado:

- Se crea un segundo modelo Naive Bayes aplicando un suavizado de Laplace para tratar los términos que no aparecen en el conjunto de entrenamiento.

- Se realiza una nueva predicción con el modelo mejorado y se evalúa su precisión con otra tabla cruzada.

Tablas Cruzadas (CrossTable):

Estas tablas proporcionan un resumen de las predicciones del modelo versus las etiquetas reales. Se presentan los siguientes datos:

- 'Adequate', 'Effective', 'Ineffective' son las etiquetas que los mensajes pueden tener.
- Cada celda muestra el número de predicciones realizadas por el modelo para una etiqueta específica en comparación con la etiqueta real.
- Las proporciones dentro de cada fila y columna también se proporcionan para ofrecer una idea más clara del desempeño del modelo.
- La primera tabla corresponde al primer modelo Naive Bayes y la segunda tabla corresponde al modelo mejorado con suavizado de Laplace.
- Según las tablas, se puede ver cómo se desempeñaron los modelos al predecir las tres categorías. Por ejemplo, en la primera tabla, de los mensajes que el modelo predijo como 'Adequate', 4542 en realidad eran 'Adequate', pero 1278 eran 'Effective' y 1216 eran 'Ineffective'. Estas cifras y proporciones permiten evaluar la precisión, sensibilidad y especificidad de los modelos.

```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## |           N / Row Total |
## |           N / Col Total |
## |-------------------------|
##
##
## Total Observations in Table:  11030
##
##
##              | actual
##    predicted |    Adequate |   Effective | Ineffective |   Row Total |
## -------------|-------------|-------------|-------------|-------------|
##     Adequate |        4542 |        1278 |        1216 |        7036 |
##              |       0.646 |       0.182 |       0.173 |       0.638 |
##              |       0.726 |       0.455 |       0.619 |             |
## -------------|-------------|-------------|-------------|-------------|
##    Effective |         753 |        1331 |         118 |        2202 |
##              |       0.342 |       0.604 |       0.054 |       0.200 |
##              |       0.120 |       0.474 |       0.060 |             |
## -------------|-------------|-------------|-------------|-------------|
##  Ineffective |         962 |         198 |         632 |        1792 |
##              |       0.537 |       0.110 |       0.353 |       0.162 |
##              |       0.154 |       0.071 |       0.321 |             |
## -------------|-------------|-------------|-------------|-------------|
## Column Total |        6257 |        2807 |        1966 |       11030 |
##              |       0.567 |       0.254 |       0.178 |             |
## -------------|-------------|-------------|-------------|-------------|
##
##
```

Figura 1. Cross Table del primer modelo naive bayes

```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## |           N / Row Total |
## |           N / Col Total |
## |-------------------------|
##
##
## Total Observations in Table:  11030
##
##
##              | actual
##    predicted |    Adequate |   Effective | Ineffective |   Row Total |
## -------------|-------------|-------------|-------------|-------------|
##     Adequate |        4256 |        1201 |        1104 |        6561 |
##              |       0.649 |       0.183 |       0.168 |       0.595 |
##              |       0.680 |       0.428 |       0.562 |             |
## -------------|-------------|-------------|-------------|-------------|
##    Effective |         791 |        1362 |         117 |        2270 |
##              |       0.348 |       0.600 |       0.052 |       0.206 |
##              |       0.126 |       0.485 |       0.060 |             |
## -------------|-------------|-------------|-------------|-------------|
##  Ineffective |        1210 |         244 |         745 |        2199 |
##              |       0.550 |       0.111 |       0.339 |       0.199 |
##              |       0.193 |       0.087 |       0.379 |             |
## -------------|-------------|-------------|-------------|-------------|
## Column Total |        6257 |        2807 |        1966 |       11030 |
##              |       0.567 |       0.254 |       0.178 |             |
## -------------|-------------|-------------|-------------|-------------|
##
##
```

Figura 2. Cross Table del segundo modelo naive bayes mejorado con suavizado de Laplace

Discusión

Al analizar los dos modelos realizados, se observa que los resultados son muy parecidos. Cada uno muestra una precisión similar al realizar predicciones, con una pequeña variación en la clasificación de argumentos ineficaces. En este ámbito, el primer modelo alcanza una precisión del 35%, mientras que el segundo registra un 34%. Por otra parte, al evaluar argumentos categorizados como adecuados y efectivos, las precisiones ascienden al 65% y 60% para cada modelo, respectivamente.

Estos hallazgos refuerzan la idea de que el algoritmo Naive Bayes posee una notable capacidad en la tarea de clasificación de textos. Una característica distintiva de este algoritmo es su eficiencia operativa, logrando resultados óptimos sin exigir una inversión considerable en recursos.

Es importante destacar que los Clasificadores Bayesiano Ingenuos (NBC) presentan una excelente capacidad de escalabilidad. Esto implica que, al incrementar el número de características o dimensiones en el modelo, el algoritmo mantiene su desempeño de manera ágil y confiable. Aunque en ciertos contextos o proyectos, los NBC podrían no parecer la opción más idónea a primera vista, resultan ser herramientas de gran valor, al

menos como punto de referencia o comparación para otros métodos o enfoques de clasificación.

Visualizaciones

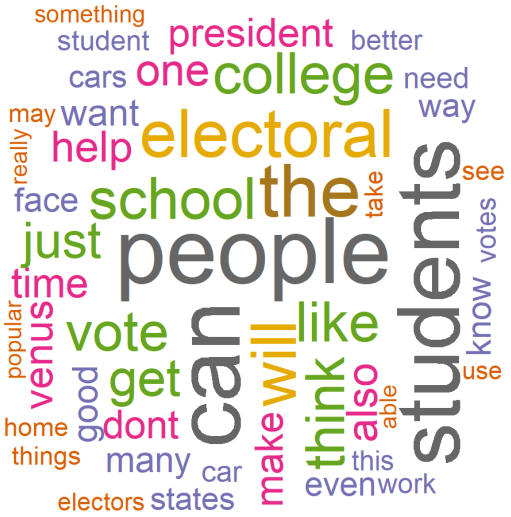

Figura 3. Frecuencia con la que aparecen ciertas palabras en los argumentos categorizados como "Adequate"

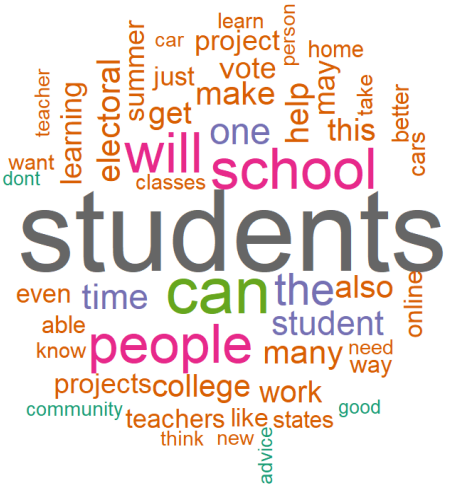

Figura 4. Frecuencia con la que aparecen ciertas palabras en los argumentos categorizados como "Effective".

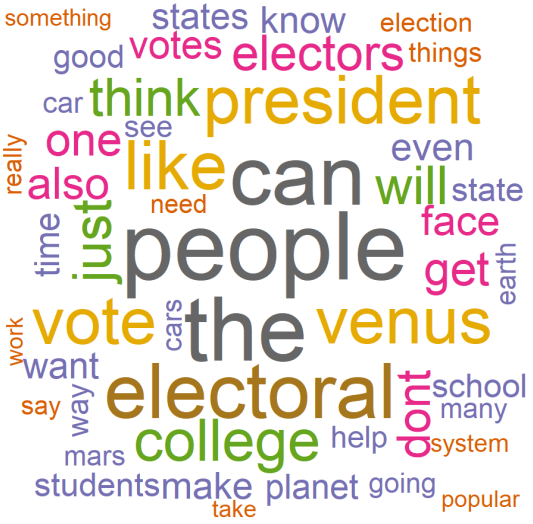

Figura 5. Frecuencia con la que aparecen ciertas palabras en los argumentos categorizados como "Ineffective".

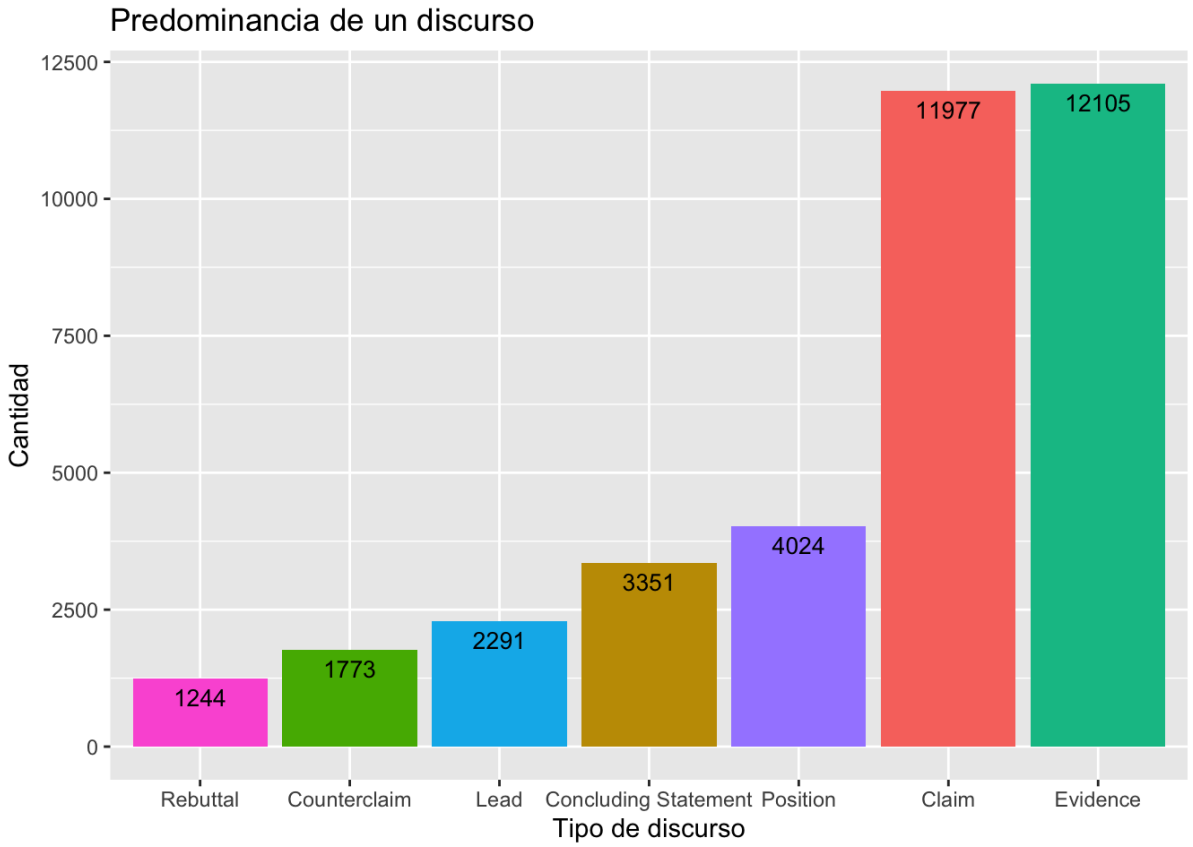

Figura 6. Predominancia que de los discursos

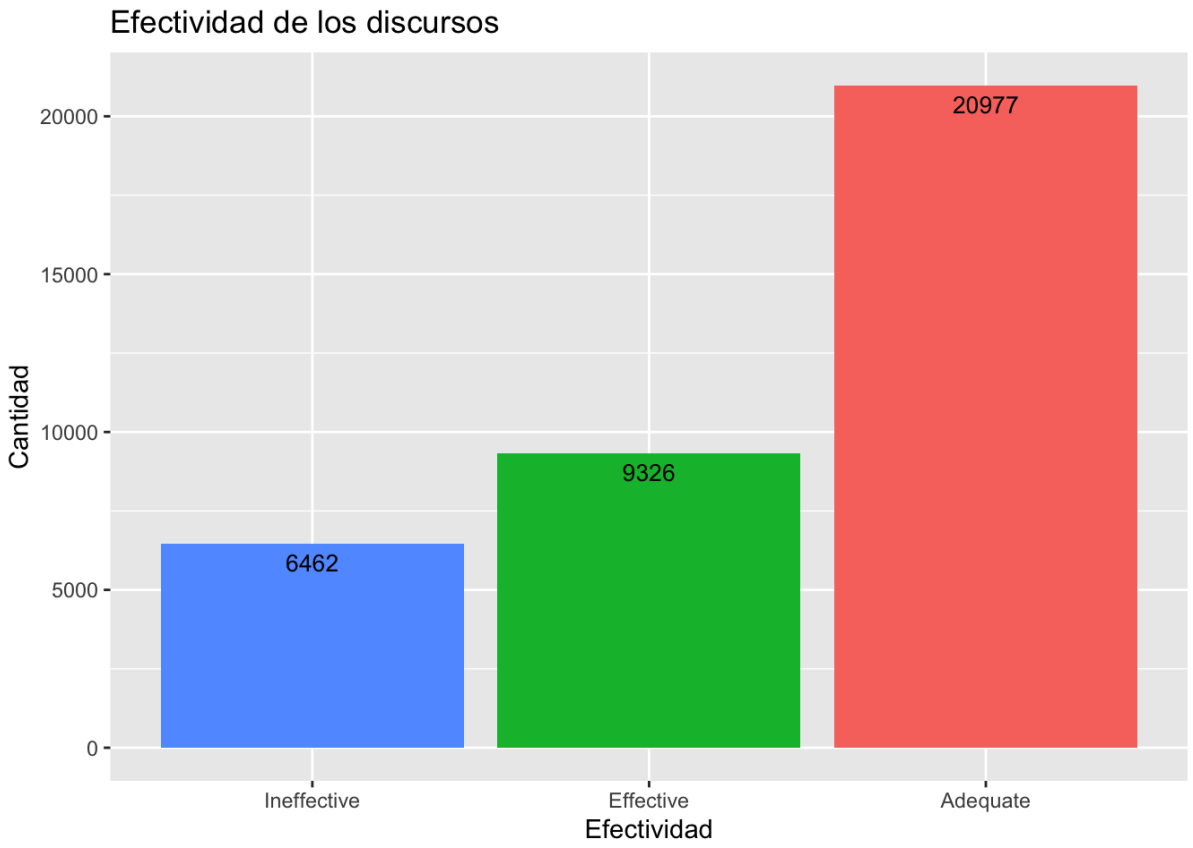

Figura 7. Efectividad de los discursos

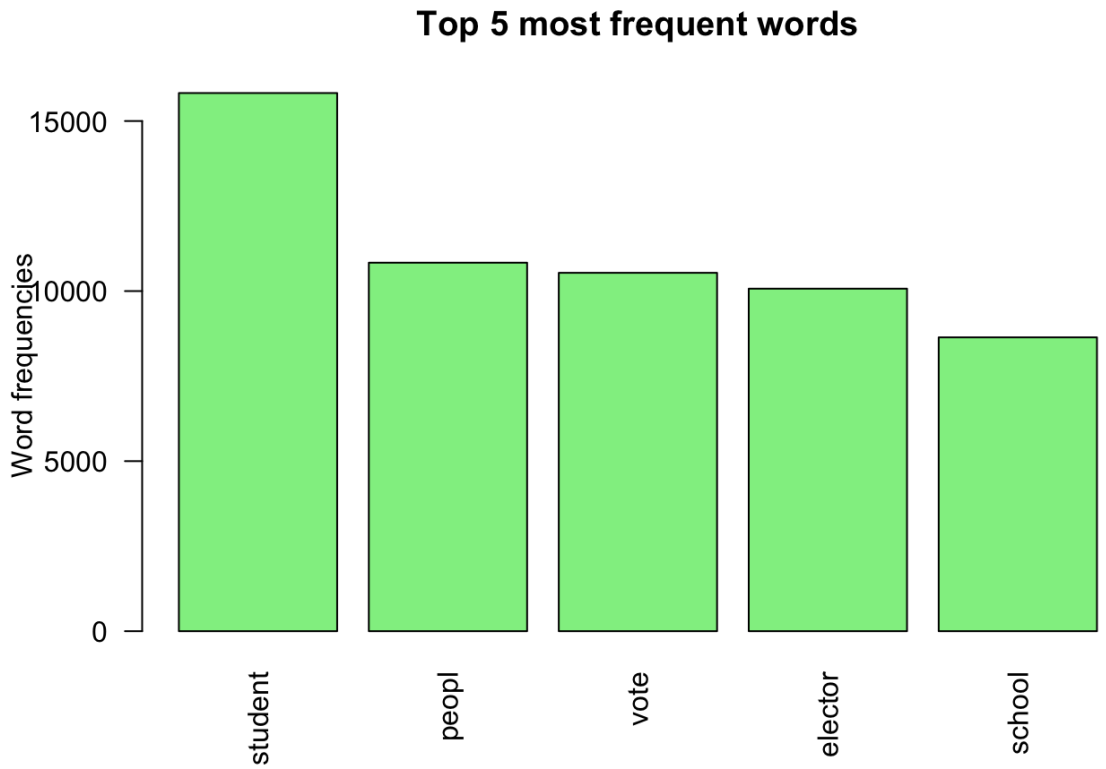

Figura 8. Top 5 de palabras que más aparecen en los discursos

Bibliografía

Medium. (25 de Abril de 2019). Algoritmos Naive Bayes: Fundamentos e Implementación. Obtenido de https://medium.com/datos-y-ciencia/algoritmos-naive-bayes-fudamentos-e-implementación-4bcb24b307f

Anguiano, E. (29 de Abril de 2009). Naive Bayes Multinomial para Clasificación de Texto Usando un Esquema de Pesado por Clases. México.

Sriram. (2022, 2 de octubre). Multinomial Naive Bayes Explained: Function, Advantages & Disadvantages, Applications in 2023. Blog Author. [https://www.upgrad.com/blog/multinomial-naive-bayes-explained]

S, L. (2022b). Gaussian Naive Bayes algorithm for credit risk modelling. Analytics Vidhya.https://www.analyticsvidhya.com/blog/2022/03/gaussian-naive-bayes-algorithm-for-credit-risk-modelling/

Martins, C. (2022, March 26). Gaussian Naive Bayes Explained and Hands-On with Scikit-Learn. Medium.

https://pub.towardsai.net/gaussian-naive-bayes-explained-and-hands-on-with-scikit-learn-4183b8cb0e4c