Universidad del Valle de Guatemala

Facultad de Ingeniería

Data Science

Proyecto #2 - Predicción de argumentos efectivos

Análisis exploratorio

Mariana David 201055

Javier Valle 20159

Angel Higueros 20460

Fredy Velásquez 201011

Situación problemática

La situación problemática que da lugar a esta investigación se centra en la escritura argumentativa de estudiantes en los grados de 6º a 12º en los Estados Unidos. A pesar de la importancia de la escritura argumentativa en el desarrollo de habilidades críticas y de participación cívica, solo el 13 por ciento de los profesores de octavo grado solicitan a sus estudiantes que escriban de manera persuasiva cada semana. Además, esta carencia afecta de manera desproporcionada a los estudiantes afroamericanos e hispanos, quienes tienen más probabilidades de escribir a un nivel considerado "por debajo del básico" en comparación con sus compañeros blancos. La disponibilidad limitada de herramientas de retroalimentación automatizada asequibles y efectivas agrava aún más esta problemática. La falta de evaluación precisa de elementos argumentativos, como organización, evidencia y desarrollo de ideas, obstaculiza el progreso de los estudiantes y la calidad de la retroalimentación proporcionada por los educadores. Esta situación destaca la necesidad de desarrollar modelos de Data Science que puedan abordar y mejorar la calidad de la escritura argumentativa de los estudiantes, reduciendo así las disparidades y proporcionando una retroalimentación más efectiva.

Problema científico

Dada la gran diversidad que existe entre los argumentos científicos, existe una cierta cantidad de argumentos que no son efectivos para un investigador, es por ello que se desea crear un sistema que procese el lenguaje natural para poder hacer una clasificación bastante exacta de los argumentos científicos que existen en internet.

Objetivos

General

- Investigar y establecer una relación cuantitativa entre la efectividad y el tipo de discurso en la escritura estudiantil, y validar que los discursos de la categoría "Effective" son predominantemente los más frecuentes mediante un análisis exhaustivo de datos.

Específicos

- Buscar una relación entre la efectividad y el tipo de discurso.
- Validar que los discursos con más existencia son los de la categoría “Effective”.

Descripción del conjunto de datos

El conjunto de datos contiene ensayos de tipo argumentativo escritos por estudiantes de los grados 6-12 de Estados Unidos. Estos ensayos fueron revisados y clasificados por expertos calificadores que identificaron los elementos comunes del discurso que se encuentran en la escritura argumentativa, los cuales son los siguientes:

- Lead: Un comienzo cautivador que puede iniciarse con una cifra impactante, una cita relevante, una breve narración o cualquier otro elemento diseñado no solo para capturar la atención del lector sino también para guiarlo hacia la idea principal o tesis del texto.
- Position: Una perspectiva clara o interpretación sobre la cuestión o tema en discusión.
- Claim: Una proposición fuerte que apoya y refuerza la postura adoptada.
- Counterclaim: Una declaración que desafía o presenta una visión alternativa a la postura establecida.
- Rebuttal: Una respuesta elaborada y pensada que desmantela o refuta el contraargumento presentado.
- Evidence: Ideas detalladas, ejemplos concretos o testimonios que validan y sostienen las afirmaciones, contrarrefutaciones o contraargumentos.
- Concluding Statement: Una sentencia final que no solo resume los puntos clave discutidos, sino que también reitera y fortalece la validez de los argumentos expuestos.

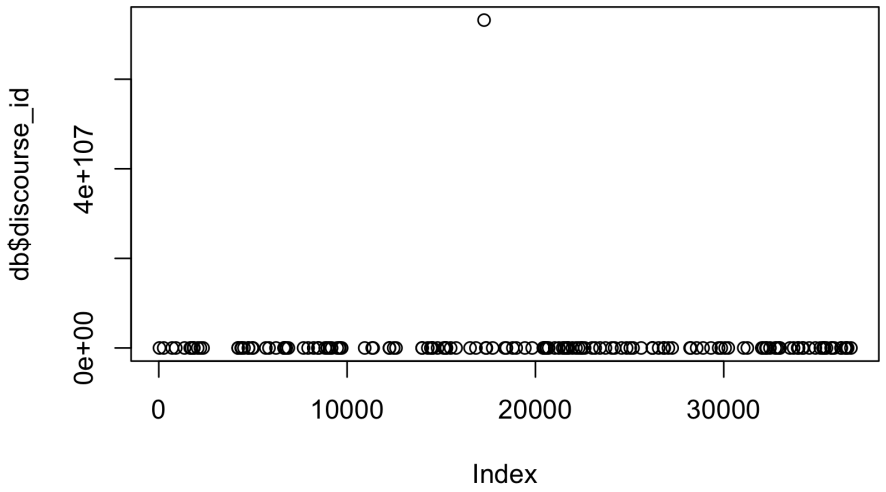
Análisis Exploratorio

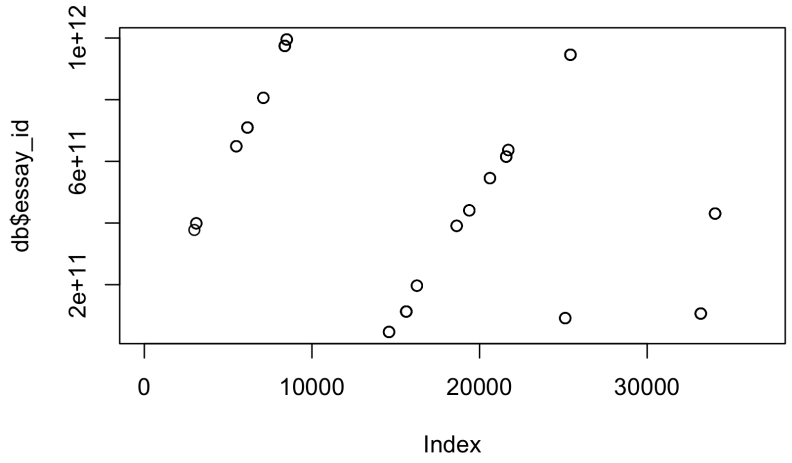
<u>Estudio de las variables cuantitativas</u>

Para ello, se analizan los elementos discursivos presentes en ensayos argumentativos escritos por estudiantes de los grados 6-12 en los Estados Unidos. Estos elementos discursivos incluyen aspectos como la introducción (Lead), la posición (Position), la afirmación (Claim), la contraargumentación (Counterclaim), la refutación (Rebuttal), la evidencia (Evidence) y la declaración conclusiva (Concluding Statement). El conjunto de datos de entrenamiento proporciona información esencial para desarrollar un modelo de aprendizaje automático que pueda realizar estas predicciones con precisión en un conjunto de prueba no visto. El estudio se centra en la tarea de procesar los textos y etiquetar cada elemento discursivo de manera

coherente con su calidad, lo que contribuirá a la mejora de las herramientas de retroalimentación automatizada en la escritura argumentativa estudiantil.

Representación gráfica de los datos






```
discourse_id         essay_id         discourse_text     discourse_type
Length:36765       Length:36765       Length:36765       Length:36765
Class :character   Class :character   Class :character   Class :character
Mode  :character   Mode  :character   Mode  :character   Mode  :character
discourse_effectiveness
Length:36765
Class :character
Mode  :character
```

- El conjunto de datos cuenta con 36765 filas y 5 columnas

1. Tipo de discurso predominante

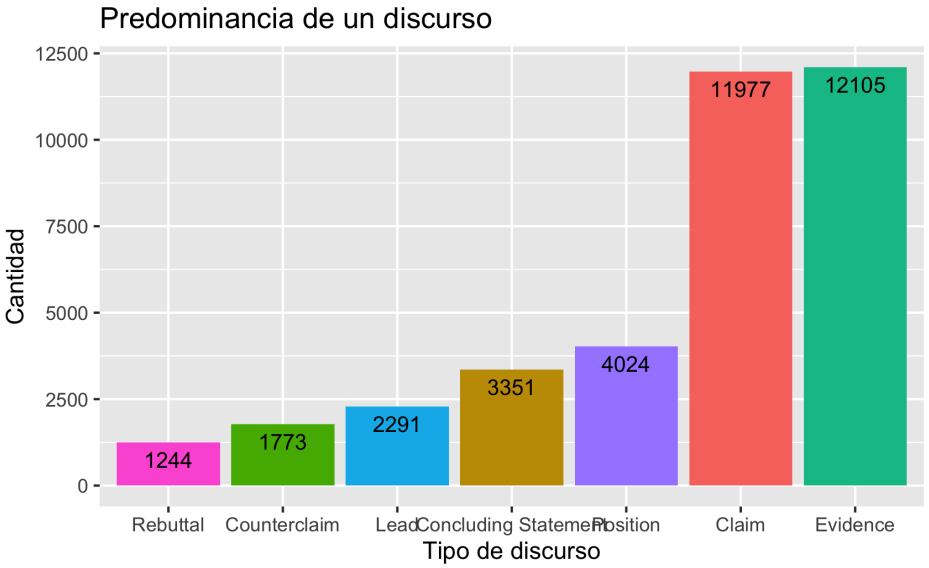


Figura 1. El tipo de discurso predominante es “Evidence”, este tipo de discurso expresa ideas o ejemplos que respaldan demandas, contrademandas o réplicas. Le sigue el tipo “Claim”, el cual es una demanda o reclamo que respalda alguna postura.

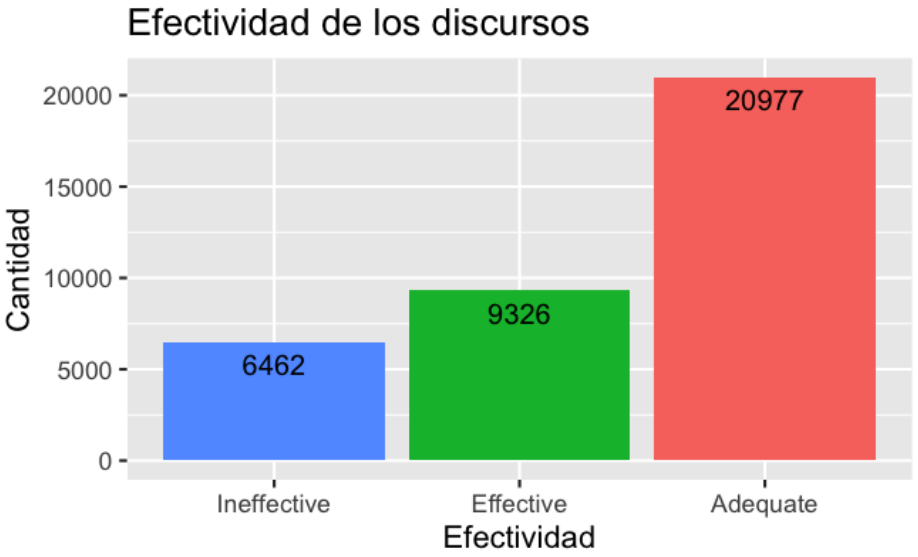


Figura 2. Es posible observar que no existe un balance correcto entre los tres tipos de efectividad.

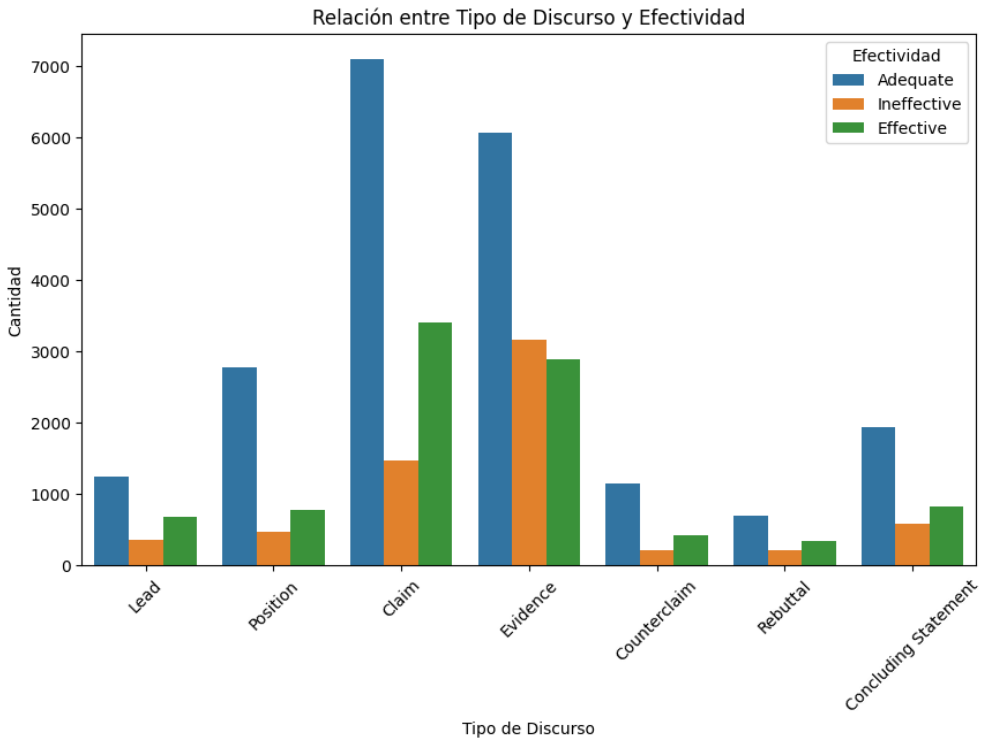


Figura 3. Se puede observar que en cada categoría de los discursos existen más discursos adecuados que ineficientes y efectivos. Por otro lado, se puede observar que los tipos de discursos con más clasificaciones son los de “Claim” y los de “Evidence”.

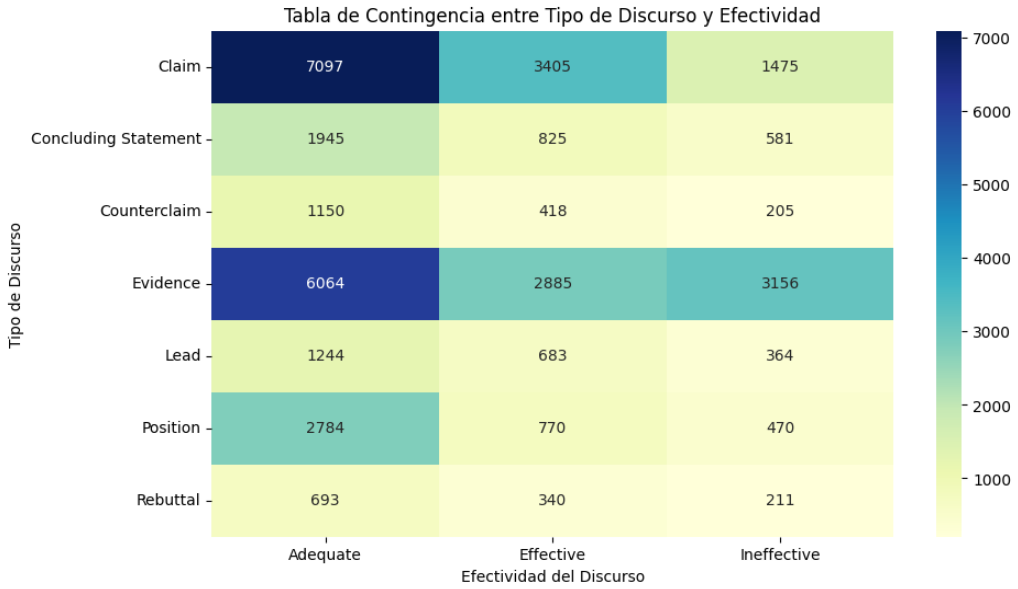


Figura 4. En la anterior tabla de contingencia se puede observar con más detalle la relación que existe entre las categorías de discursos y su efectividad.

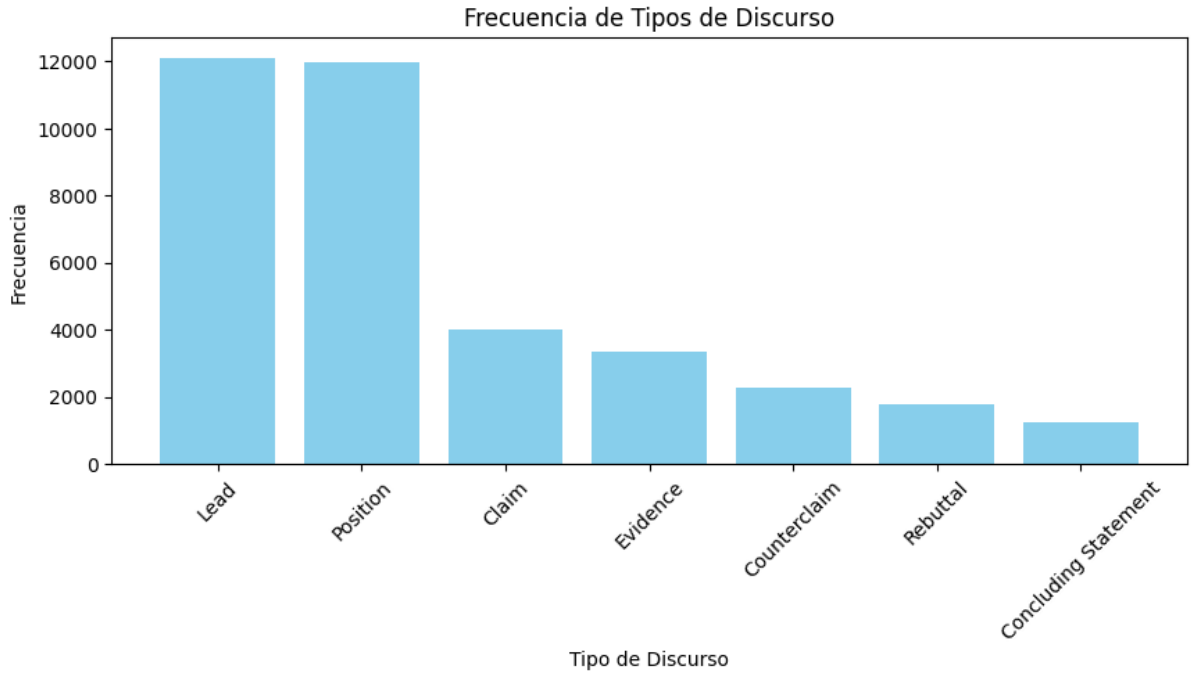


Figura 5. Se puede observar que las categorías que contienen más discursos dentro del dataset son “Lead” y “Position”.

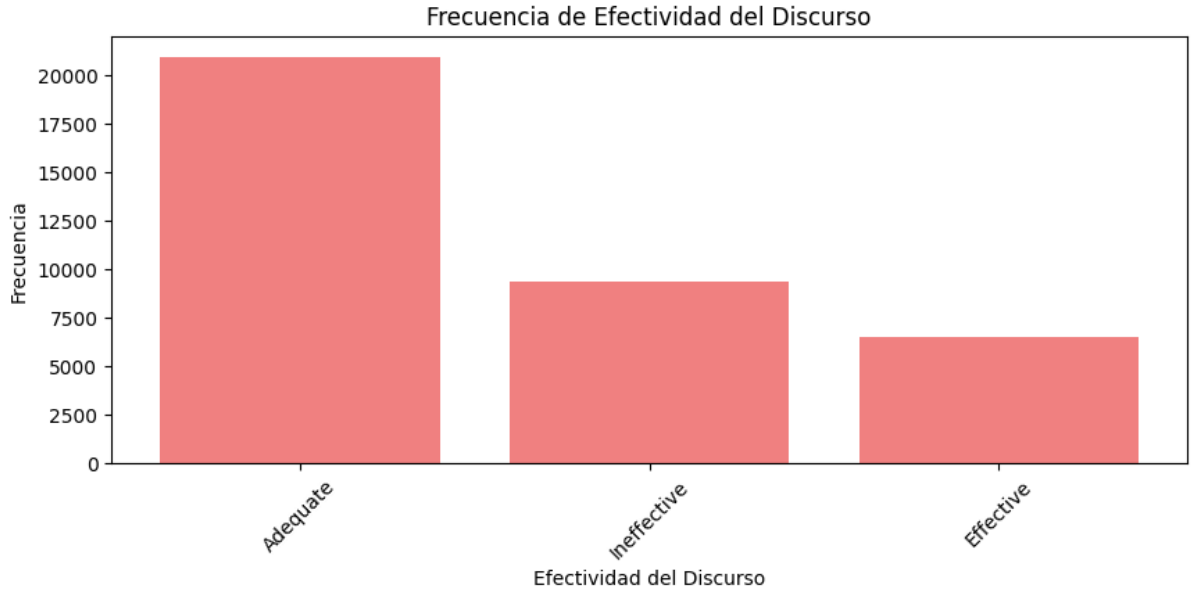


Figura 6. Es notorio que, dentro de nuestro dataset, la efectividad del discurso con más datos de estos mismos es de “Adequate”, luego le sigue “Inefective” y, por último, “Effective”.

Análisis de los discursos del conjunto de datos

- ☐ *Discurso sin limpieza*

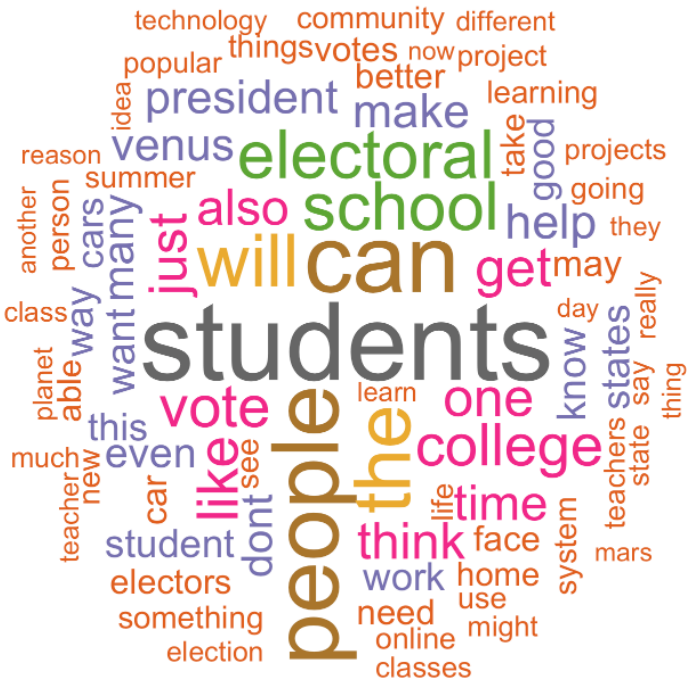


Figura 7. Las palabras con mayor frecuencia en el conjunto de datos son “students” y “people”, también es posible encontrar algunas como “vote” y “school”

- ☐ *Discurso con limpieza*

Métodos aplicados para la limpieza:

1. Se crea un Corpus desde la columna discourse_text
2. A continuación, se realizan una serie de transformaciones en el texto para estandarizar y limpiar el contenido:
    a. Convertir todo a minúsculas.
    b. Eliminar números.
    c. Eliminar palabras comunes del inglés (stopwords).
    d. Eliminar algunas palabras específicas.
    e. Eliminar puntuaciones.
    f. Eliminar espacios extra.

g. Derivar las palabras, es decir, reducir las palabras a su forma raíz (por ejemplo, "running" se convierte en "run").emplazan ciertos caracteres ("/", "@" y "|") por espacios.

3. Crear matriz que muestra la frecuencia con la que aparece cada término en cada documento del corpus.
4. Mostrar las 5 palabras más frecuentes y luego crea un gráfico de barras para esas palabras.
5. Se crea una nube de palabras que muestra las palabras más frecuentes en el corpus.

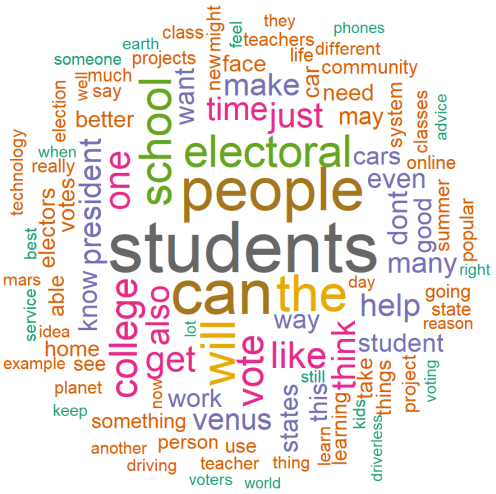


Figura 8. Tras aplicar la limpieza antes mencionada, la palabra “students” sigue siendo la más predominante en los discursos compilados.

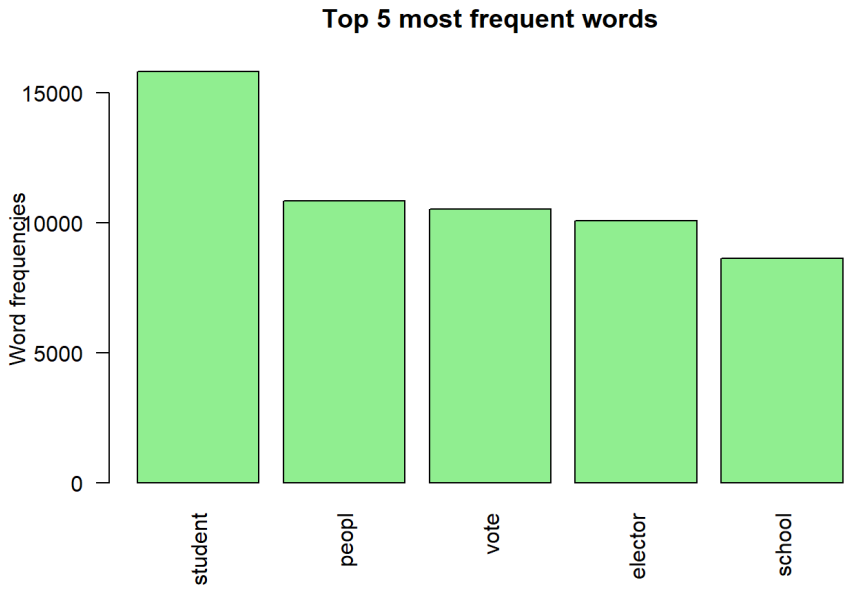


Figura 9. Es posible observar que “student” es la palabras más utilizada con 15823, “people” con 10835, “vote” con 10537, “elector” con 10071 y “school” con 8639.

Análisis de sentimiento

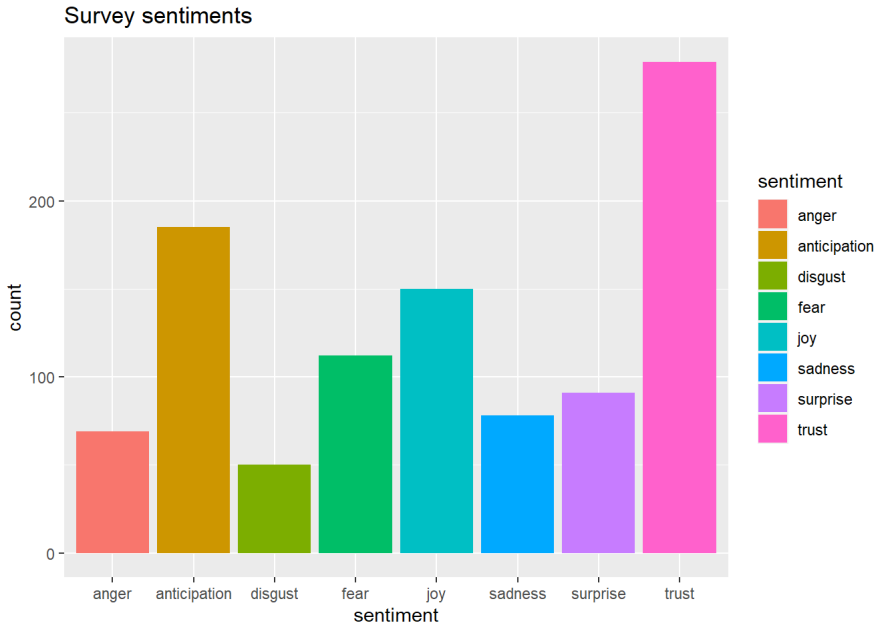


Figura 10. Conteo de palabras asociado a los sentimientos presentados.

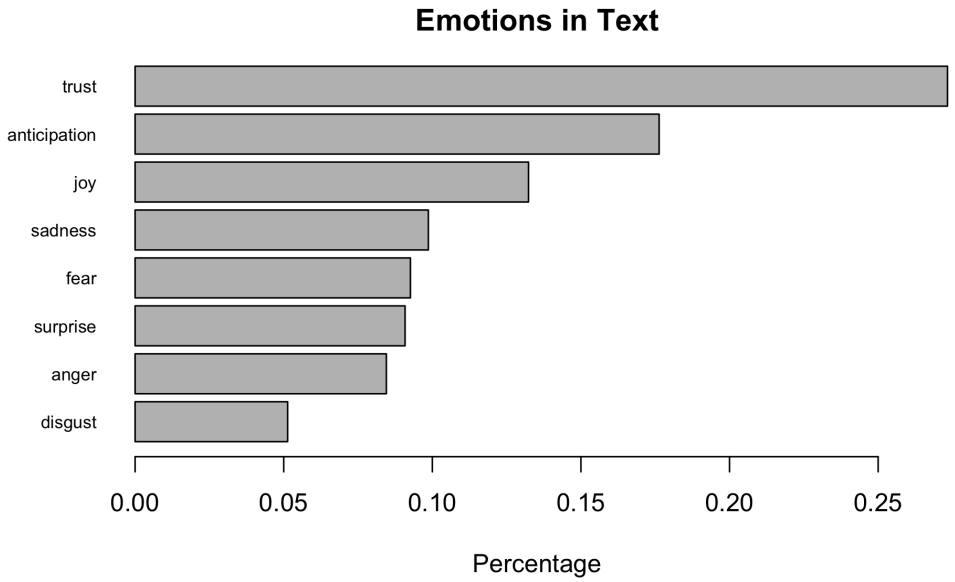


Figura 11. Porcentaje de emociones en el texto.

A partir de un meticuloso análisis de las gráficas presentadas, se puede inferir que una proporción significativa de los ensayos redactados por los estudiantes, específicamente más del 25%, logra comunicar un sentido de autenticidad y veracidad. Adicionalmente, es digno de resaltar que más del 15% de dichos ensayos exuda un marcado sentimiento de anticipación y esperanza, y que al menos un 10% de ellos transmite con éxito sensaciones de alegría y bienestar. Sin embargo, es imperativo señalar que las emociones menos representadas en estos ensayos son, en su orden, la sorpresa, el enojo y el disgusto.

```
discourse_id         essay_id         discourse_text     discourse_type
Length:6462        Length:6462        Length:6462        Length:6462
Class :character   Class :character   Class :character   Class :character
Mode  :character   Mode  :character   Mode  :character   Mode  :character



discourse_effectiveness     Enojo          Anticipacion        Disgusto
Length:6462             Min.   :0.0000   Min.   : 0.0000   Min.   :0.0000
Class :character        1st Qu.:0.0000   1st Qu.: 0.0000   1st Qu.:0.0000
Mode  :character        Median :0.0000   Median : 1.0000   Median :0.0000
                        Mean   :0.4559   Mean   : 0.8781   Mean   :0.2849
                        3rd Qu.:1.0000   3rd Qu.: 1.0000   3rd Qu.:0.0000
                        Max.   :9.0000   Max.   :10.0000   Max.   :7.0000
    Miedo           Felicidad          Tristeza          Sorpresa
Min.   : 0.0000   Min.   : 0.0000   Min.   : 0.0000   Min.   :0.0000
1st Qu.: 0.0000   1st Qu.: 0.0000   1st Qu.: 0.0000   1st Qu.:0.0000
Median : 0.0000   Median : 0.0000   Median : 0.0000   Median :0.0000
Mean   : 0.4449   Mean   : 0.6998   Mean   : 0.4989   Mean   :0.5118
3rd Qu.: 1.0000   3rd Qu.: 1.0000   3rd Qu.: 1.0000   3rd Qu.:1.0000
Max.   :11.0000   Max.   :10.0000   Max.   :10.0000   Max.   :9.0000
    Verdad          Negativo          Positivo
Min.   : 0.000   Min.   : 0.0000   Min.   : 0.000
1st Qu.: 0.000   1st Qu.: 0.0000   1st Qu.: 0.000
Median : 1.000   Median : 0.0000   Median : 1.000
Mean   : 1.338   Mean   : 0.8955   Mean   : 1.989
3rd Qu.: 2.000   3rd Qu.: 1.0000   3rd Qu.: 3.000
Max.   :14.000   Max.   :18.0000   Max.   :27.000
discourse_id         essay_id         discourse_text     discourse_type
Length:9326        Length:9326        Length:9326        Length:9326
Class :character   Class :character   Class :character   Class :character
Mode  :character   Mode  :character   Mode  :character   Mode  :character
```

Figura 12. Análisis estadístico de la segmentación inefectiva.

```
discourse_effectiveness     Enojo         Anticipacion      Disgusto
Length:9326            Min.   :0.0000   Min.   : 0.00   Min.   :0.0000
Class :character       1st Qu.:0.0000   1st Qu.: 0.00   1st Qu.:0.0000
Mode  :character       Median :0.0000   Median : 1.00   Median :0.0000
                       Mean   :0.5109   Mean   : 1.22   Mean   :0.2922
                       3rd Qu.:1.0000   3rd Qu.: 2.00   3rd Qu.:0.0000
                       Max.   :8.0000   Max.   :13.00   Max.   :7.0000
    Miedo           Felicidad          Tristeza          Sorpresa
Min.   : 0.0000   Min.   : 0.0000   Min.   : 0.0000   Min.   :0.0000
1st Qu.: 0.0000   1st Qu.: 0.0000   1st Qu.: 0.0000   1st Qu.:0.0000
Median : 0.0000   Median : 0.0000   Median : 0.0000   Median :0.0000
Mean   : 0.6288   Mean   : 0.8719   Mean   : 0.6514   Mean   :0.5298
3rd Qu.: 1.0000   3rd Qu.: 1.0000   3rd Qu.: 1.0000   3rd Qu.:1.0000
Max.   :14.0000   Max.   :10.0000   Max.   :13.0000   Max.   :6.0000
    Verdad          Negativo         Positivo
Min.   : 0.000   Min.   : 0.000   Min.   : 0.000
1st Qu.: 0.000   1st Qu.: 0.000   1st Qu.: 1.000
Median : 1.000   Median : 1.000   Median : 2.000
Mean   : 1.864   Mean   : 1.241   Mean   : 3.184
3rd Qu.: 3.000   3rd Qu.: 2.000   3rd Qu.: 5.000
Max.   :15.000   Max.   :18.000   Max.   :24.000
discourse_id         essay_id         discourse_text     discourse_type
Length:20977       Length:20977       Length:20977       Length:20977
Class :character   Class :character   Class :character   Class :character
Mode  :character   Mode  :character   Mode  :character   Mode  :character
```

Figura 13. Análisis estadístico de la segmentación efectiva.

discourse_effectiveness	Enojo	Anticipacion	Disgusto
Length:20977	Min. :0.0000	Min. : 0.000	Min. :0.000
Class :character	1st Qu.:0.0000	1st Qu.: 0.000	1st Qu.:0.000
Mode :character	Median :0.0000	Median : 0.000	Median :0.000
	Mean :0.3468	Mean : 0.677	Mean :0.216
	3rd Qu.:1.0000	3rd Qu.: 1.000	3rd Qu.:0.000
	Max. :7.0000	Max. :12.000	Max. :6.000

Miedo	Felicidad	Tristeza	Sorpresa
Min. :0.0000	Min. :0.0000	Min. : 0.0000	Min. :0.0000
1st Qu.:0.0000	1st Qu.:0.0000	1st Qu.: 0.0000	1st Qu.:0.0000
Median :0.0000	Median :0.0000	Median : 0.0000	Median :0.0000
Mean :0.3662	Mean :0.5154	Mean : 0.3905	Mean :0.3744
3rd Qu.:1.0000	3rd Qu.:1.0000	3rd Qu.: 1.0000	3rd Qu.:1.0000
Max. :7.0000	Max. :9.0000	Max. :11.0000	Max. :8.0000

Verdad	Negativo	Positivo
Min. : 0.000	Min. : 0.0000	Min. : 0.000
1st Qu.: 0.000	1st Qu.: 0.0000	1st Qu.: 0.000
Median : 1.000	Median : 0.0000	Median : 1.000
Mean : 1.069	Mean : 0.7057	Mean : 1.624
3rd Qu.: 2.000	3rd Qu.: 1.0000	3rd Qu.: 2.000
Max. :13.000	Max. :16.0000	Max. :25.000

Figura 14. Análisis estadístico de la segmentación adecuada.

Categoria <chr>	Enojo <dbl>	Anticipacion <dbl>	Disgusto <dbl>	Miedo <dbl>	Felicidad <dbl>	Tristeza <dbl>	Sorpresa <dbl>	Verdad <dbl>	Negativo <dbl>
Ineficiente	0.4558960	0.8780563	0.2848963	0.4449087	0.6997833	0.4989167	0.5117611	1.337512	0.8955432
Adecuado	0.3468084	0.6769795	0.2160461	0.3661629	0.5153740	0.3904753	0.3743624	1.069219	0.7056776
Eficiente	0.5109372	1.2201373	0.2921939	0.6287798	0.8718636	0.6514047	0.5298091	1.863607	1.2408321

3 rows | 1–10 of 11 columns

Figura 15. Media de sentimientos según la segmentación.

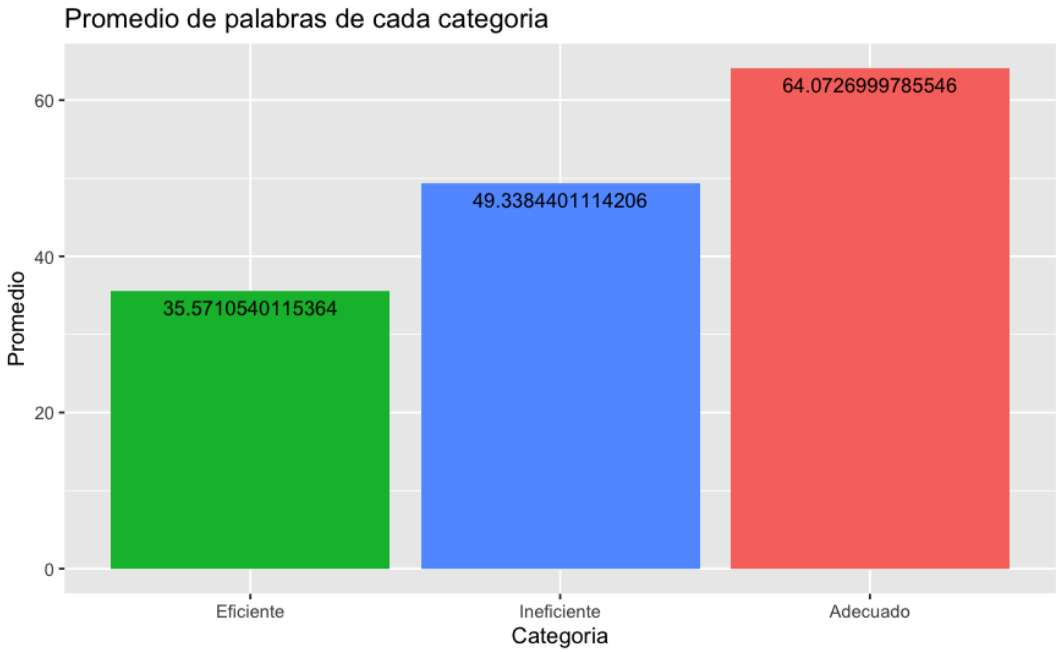


Figura 16. Palabras para cada categoría de eficacia discursiva.

Tras examinar la gráfica en cuestión, se destaca que la abundancia de palabras en un discurso o texto no garantiza su claridad o eficiencia. Este hallazgo subraya la premisa de que la extensión no siempre se traduce en excelencia.

Hallazgos y conclusiones

- Se evidencia que la longitud de un discurso o texto (medida en términos de cantidad de palabras) no necesariamente implica una mayor eficacia o comprensibilidad. En otras palabras, un discurso más largo no garantiza que sea de mejor calidad o más eficaz.
- Los argumentos que se articulan con una escritura refinada, una selección meticulosa de verbos y, esencialmente, un uso impecable del lenguaje, tienden a manifestar una riqueza emocional más profunda en su narrativa.
- La profundidad emocional no sólo embellece el contenido, sino que también establece un puente más sólido de conexión con los lectores, permitiéndoles sintonizar con las intenciones y sentimientos del autor.

- La escritura efectiva no solo eleva la calidad del contenido, sino que también facilita una comunicación efectiva y un entendimiento claro, enriqueciendo la experiencia de lectura y fortaleciendo el impacto del mensaje transmitido.
- Un argumento eficiente se distingue por su habilidad para expresar sentimientos profundos. Aunque la categoría "eficiente" prevalece en la expresión de emociones, es esencial entender que un tono puramente positivo no garantiza su eficacia.
- El análisis demuestra que la verdadera conexión emocional con el lector es lo que verdaderamente potencia la persuasión y autenticidad del mensaje.