Análisis cuantitativo del texto II

Unsupervised learning y GeminiAI

Matías Deneken

May 22, 2025

¿Qué veremos hoy?

1. ¿Cómo hacer cosas con las palabras?
2. Breve repaso
3. Conceptos claves
    a. Unsupervised learning
    b. Modelamiento de tópicos
    c. LLM
    d. ¿Qué es una Api?
4. Uso y práctica en el software
    a. LDA
    b. Api GeminiAI
5. Preguntas empíricas

Resultados esperados

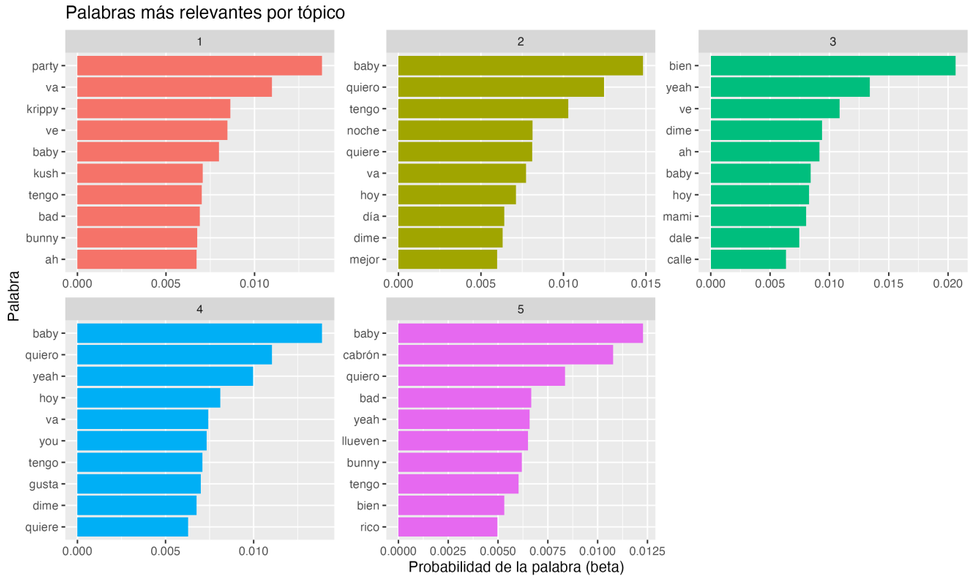
Análisis de texto con IA 🤭

```
Source on Save | Run | Source
86   "Eres un analista político chileno especializado en discurso ideológico.\n",
87   "La derecha chilena se define como el conjunto de fuerzas políticas identificadas con ",
88   "valores conservadores o liberales en lo económico, que históricamente han defendido ",
89   "el orden institucional, el mercado libre y una visión tradicional de la sociedad. ",
90   "Esto incluye partidos como la UDI, RN, Evópoli y el Partido Republicano.\n\n",
91
92   "Analiza este texto en base a las siguientes dimensiones. Si no aplica, responde con 'no aplica'
93   "Solo da un CONCEPTO concreto por línea, sin explicaciones.\n\n",
94   "1. Tópico temático principal:\n",
95   "2. Frame ideológico dominante:\n",
96   "3. Sentimiento global del texto:\n",
97   "4. Adversario(s) principal(es):\n",
98   "5. Posición respecto a Chile Vamos:\n",
99   "6. Alianzas simbólicas o redes mencionadas:\n",
100  "7. Personas nombradas:\n\n",
101  "Texto a analizar:\n{texto}"
102  )
172:1  PROMPT PARA DISCURSO DE ODIO  Copilot: No completions available.  R Script

Console  Terminal  Background Jobs
R 4.4.3 · ~/Dropbox/CIIR/curso_interculturales/

> cat(resultado)
1.  Seguridad ciudadana.
2.  Ley y orden.
3.  Crítica.
4.  La izquierda.
5.  No aplica.
6.  Carabineros.
7.  José Antonio Kast.
> #
```

Modelamiento de tópicos (Bad Bunny songs)



¿Cómo hacer cosas con las palabras?

Pues bien; si a un cervantista se le ocurriera decir: el Quijote empieza con dos palabras monosilábicas terminadas en n: (en y un), y sigue con una de cinco letras (lugar), con dos de dos letras (de la), con una de cinco o de seis (Mancha), y luego se le ocurriera derivar conclusiones de eso, inmediatamente se pensaría que está loco. La Biblia ha sido estudiada de ese modo.

Jorge Luis Borges en "La Cábala". Conferencias denominadas Siete Noches

Conceptos claves

Conceptos claves (Aprendizaje supervisado)

En este enfoque, cada dato de entrenamiento cuenta con una etiqueta (label), es decir, la respuesta correcta que queremos enseñar al modelo.

- Correo electrónico → spam o no spam.
- Texto de opinión → sentimiento: positivo, negativo, neutral.

Cómo funciona:

- Alimentamos el modelo con ejemplos (entrada, etiqueta) –> El modelo aprende a mapear entradas a etiquetas. Luego puede predecir etiquetas en datos nuevos.

Ejemplos en ciencias sociales:

- Clasificar tweets por ideología política: cada tweet con etiqueta izquierda, derecha, etc.
- Predecir estado cultural: texto de entrevistas etiquetado como “optimista” o “pesimista”.

Conceptos claves (Aprendizaje supervisado)

Ejemplo: Clasificación de tweets por orientación ideológica

Objetivo

Clasificar tweets según si expresan una **posición política de derecha** (vs. centro o izquierda).

Texto del tweet	Etiqueta
“Los empresarios crean empleo, no el Estado. Menos intervención, por favor.”	derecha
“No más ideología de género en las escuelas. Dejen educar a los padres.”	derecha
“El gobierno debe garantizar salud pública y educación gratuita”	izquierda
“La constitución debe reconocer los derechos de los pueblos originarios”	izquierda

Conceptos claves (Aprendizaje supervisado)

¿Qué contienen los tweets de derecha?

- **Temas frecuentes**:
  - orden, ley, seguridad
  - libertad económica, emprendimiento
  - defensa de Carabineros o militares
  - rechazo al Estado grande o a políticas redistributivas
  - crítica a ideologías progresistas (feminismo, género, multiculturalismo)
- **Lenguaje característico**:
  - “ideología de género”, “orden”, “delincuencia”, “progreso real”, “libertad”

Conceptos claves (Aprendizaje supervisado)

¿Cómo usar esto para clasificar?

1. **Recolectas más tweets**.
2. **Etiquetas una muestra manualmente** (`derecha`, `izquierda`, `centro`, `ambiguo`).
3. Tokenizas el texto y extraes características (TF-IDF o embeddings).
4. Entrenas un clasificador supervisado (por ejemplo: regresión logística).
5. Evalúas su precisión en una muestra de test.

Conceptos claves (Aprendizaje No Supervisado)

El aprendizaje **no supervisado** es un enfoque del machine learning en el cual **no contamos con etiquetas previas**. En lugar de predecir una clase o valor, el algoritmo **explora patrones internos en los datos**, buscando:

- Agrupar elementos similares (**clustering**)
- Detectar **temas latentes** o estructuras (ej. con LDA)

¿Por qué usarlo en texto?

- Los textos muchas veces **no vienen etiquetados**.
- Etiquetar manualmente puede ser **costoso, subjetivo y lento**.
- El no supervisado permite:
  - **Explorar corpus grandes**
  - **Detectar patrones sin sesgo**
  - **Agrupar textos automáticamente** (temas, posturas, estilos)

Conceptos claves (Aprendizaje No Supervisado)

ID	Texto del tweet
T1	“Boric quiere acabar con las instituciones democráticas. Grave.”
T2	“La libertad individual está por sobre el Estado. Basta de intervencionismo.”
T3	“Apoyo total a nuestras Fuerzas Armadas. Orden, disciplina y respeto.”
T4	“La educación pública debe ser gratuita y de calidad. El Estado tiene un rol central.”
T5	“No a la ideología de género. Proteger la familia es prioridad nacional.”

Limpieza y tokenización

Después de limpiar (sin stopwords, sin hashtags ni puntuación), quedan tokens como:

- T1: boric, quiere, acabar, instituciones, democráticas
- T2: libertad, individual, estado, intervencionismo
- T3: apoyo, fuerzas, armadas, orden, disciplina, respeto

Conceptos claves

Tuit	boric	quiere	acabar	instituciones	democráticas	libertad	individual	esta
T1	1	1	1	1	1	0	0	
T2	0	0	0	0	0	1	1	
T3	0	0	0	0	0	0	0	
T4	0	0	0	0	0	0	0	
T5	0	0	0	0	0	0	0	

- T4: educación, pública, gratuita, calidad, estado, rol
- T5: ideología, género, proteger, familia, prioridad, nacional

Conceptos claves

- **Comparar tweets según palabras compartidas**
  Para cada par de tweets, puedes contar cuántas palabras tienen en común.
- **Calcular similitud manualmente**
  La **similitud del coseno** se puede calcular a mano usando su fórmula:

$$\text{sim}(A, B) = \frac{A \cdot B}{\|A\| \cdot \|B\|}$$

- (A B): es el **producto punto** entre los vectores (A) y (B). Mide cuántas palabras tienen en común, y puede incorporar pesos si se usan frecuencias o valores como TF-IDF.
- (|A|) y (|B|): son las **normas** (o magnitudes) de los vectores (A) y (B).

```
#calidad, educación, estado, gratuita, individual, intervencionismo, libertad, pública, rol

T2 <- c(0, 0, 1, 0, 1, 1, 1, 0, 0)
T4 <- c(1, 1, 1, 1, 0, 0, 0, 1, 1)
```

Conceptos claves

Cálculo paso a paso

1. Producto punto

$$A \cdot B = (0 \cdot 1) + (0 \cdot 1) + (1 \cdot 1) + (0 \cdot 1) + (1 \cdot 0) + (1 \cdot 0) + (1 \cdot 0) + (0 \cdot 1) + (0 \cdot 1)$$

2. Normas

$$\|T_2\| = \sqrt{1^2 + 1^2 + 1^2 + 1^2} = \sqrt{4} = 2$$

$$\|T_4\| = \sqrt{1^2 + 1^2 + 1^2 + 1^2 + 1^2 + 1^2} = \sqrt{6} \approx 2.449$$

3. Similitud del coseno

$$\text{sim}(T_2, T_4) = \frac{1}{2 \cdot 2.449} = \frac{1}{4.898} \approx 0.204$$

Conceptos claves: Modealmiento de tópicos (LDA).

El modelamiento de tópicos es una técnica de aprendizaje automático no supervisado que permite descubrir temas latentes (tópicos) en un conjunto de documentos de texto.

Un tópico es una distribución de probabilidad sobre palabras. Cada documento es una mezcla de estos tópicos. El algoritmo más usado es LDA (Latent Dirichlet Allocation).

1. Representación vectorial (matriz documento-palabra):

- Antes transformaste los tweets en vectores binarios con 0 y 1 (presencia/ausencia de palabras).

- Esa matriz es el insumo de LDA.

Modelamiento LDA en R: Tweets.

```r
# Crear dataset
tweets <- tibble::tibble(
  id = paste0("T", 1:10),
  texto = c(
    "Boric quiere acabar con las instituciones democráticas. Grave.",
    "La libertad individual está por sobre el Estado. Basta de intervencionismo.",
    "Apoyo total a nuestras Fuerzas Armadas. Orden, disciplina y respeto.",
    "La educación pública debe ser gratuita y de calidad. El Estado tiene un rol central.",
    "No a la ideología de género. Proteger la familia es prioridad nacional.",
    "El orden público debe garantizarse. Apoyo a Carabineros.",
    "El Estado no debe intervenir en la economía. Basta de burocracia.",
    "La patria necesita líderes firmes. Basta de debilidad institucional.",
    "Los padres tienen derecho a decidir sobre la educación de sus hijos.",
    "La delincuencia ha superado todos los límites. Más penas y más cárceles."
  )
)
```

Modelamiento LDA en R

Stopwords y Token

```
stop_es <- stopwords::stopwords("es")

# Tokenización y limpieza con stopwords en español
tokens <- tweets %>%
  unnest_tokens(word, texto) %>%
  filter(!word %in% stop_es) %>%
  count(id, word)

tokens |> head(4)
```

```
# A tibble: 4 × 3
  id    word             n
  <chr> <chr>        <int>
1 T1    acabar           1
2 T1    boric            1
3 T1    democráticas     1
4 T1    grave            1
```

Convertir en matrix

```
# Convertir a DocumentTermMatrix
dtm <- tokens %>%
  cast_dtm(document = id, term = word, value = n); dtm
```

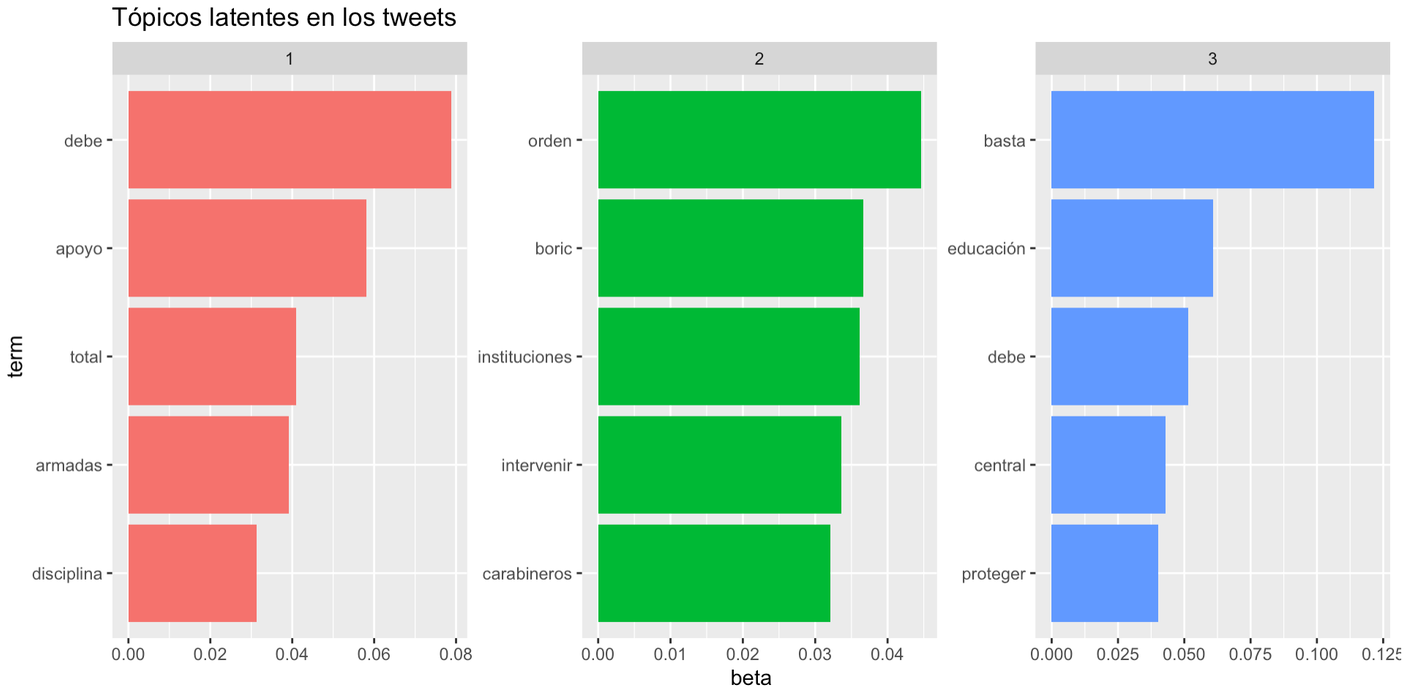
```
<<DocumentTermMatrix (documents: 10, terms: 52)>>
Non-/sparse entries: 59/461
Sparsity           : 89%
Maximal term length: 16
Weighting          : term frequency (tf)
```

LDA en R

```r
# Ajustar modelo LDA (3 tópicos)
lda_model <- LDA(dtm, k = 3, control = list(seed = 1234)) # Seed es para reproducibilidad

# Distribución de palabras por tópico (beta)
topics_terms <- tidy(lda_model, matrix = "beta")

# Mostrar las 5 palabras más probables por tópico
topics_terms %>%
  group_by(topic) %>%
  slice_max(beta, n = 5) %>%
  arrange(topic, -beta) %>%
  mutate(term = reorder_within(term, beta, topic)) %>%
  ggplot(aes(beta, term, fill = factor(topic))) +
  geom_col(show.legend = FALSE) +
  facet_wrap(~topic, scales = "free") +
  scale_y_reordered() +
  labs(title = "Tópicos latentes en los tweets")
```



LDA en R

Distribución de probabilidad por tópico (gamma)

document	Tópico_1	Tópico_2	Tópico_3
T1	0.327	0.342	0.331
T10	0.336	0.339	0.325
T2	0.327	0.332	0.341
T3	0.347	0.326	0.327
T4	0.330	0.332	0.338

- Cada celda muestra la probabilidad de que un tweet esté asociado a un determinado tópico. Por ejemplo:
  - T1 tiene: 32.7% de pertenencia al Tópico 1, 34.2% al Tópico 2, 33.1% al Tópico 3.

DESPUÉS VEREMOS DATOS REALES

USO DE IA GENERATIVA PARA DATOS TEXTUALES

¿Qué es un *prompt*?

Un **prompt** es el **texto de entrada** que le das a un modelo de lenguaje (como ChatGPT) para obtener una respuesta.

- Puede ser una instrucción, una pregunta o un texto para completar.
- Es la **forma en que dialogas con la IA**.

🔷 *Ejemplo:*

> “Resume este artículo en 3 frases”
> “Clasifica este tuit como positivo o negativo”

En modelos generativos, **la calidad del resultado depende en gran medida del prompt.**

¿Qué es un *LLM*?

Un **LLM** (*Large Language Model*) es un **modelo de lenguaje entrenado con enormes volúmenes de texto** para predecir, generar o analizar lenguaje natural.

🔷 Algunos conocidos:

- GPT-4 (OpenAI)
- BERT (Google)
- LLaMA (Meta)
- Claude (Anthropic)

Estas arquitecturas permiten tareas como:

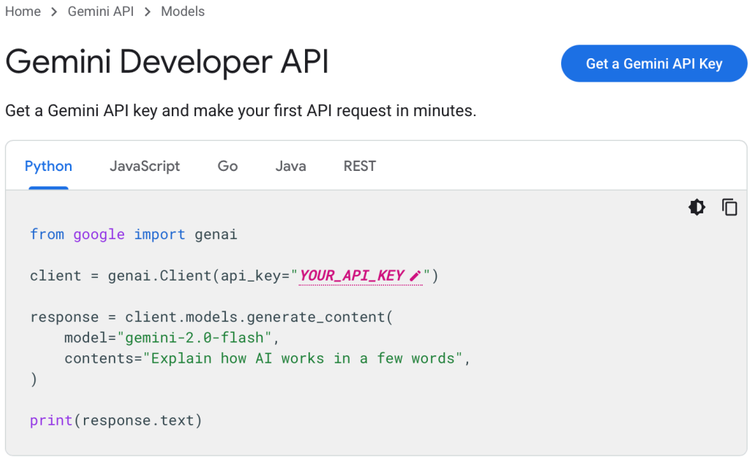
- Traducción automática
- Resumen de textos
- Análisis de sentimientos
- Extracción de tópicos o entidades

¿Qué es una *API*?

Una **API** (*Application Programming Interface*) es una **interfaz que permite que dos programas se comuniquen entre sí**.

🔷 En este contexto:

Una API te permite **enviar texto a un LLM desde tu código** (por ejemplo en R o Python), y recibir la respuesta automáticamente.



VAMOS AL CÓDIGO!