Informe Final MA5406

Procesamiento de datos textuales EthicApp con algoritmos de procesamiento de lenguaje natural

Camilo Carvajal Reyes

21 de julio 2023

Resumen

La aplicaión *EthicApp* es una herramienta que permite recolectar preferencias de estudiantes ante dilemas éticos y sus justificaciones. Lamentablemente, el gran volumen de datos (del orden de dos mil textos) dificulta el análisis de estos. Este trabajo aborda el uso de modelos de aprendizaje de máquina supervisados y no-supervisados para modelar la estructura textual de las respuestas y con esto apoyar el análisis que puedan hacer los equipos docentes. Pese al potencial de mejora, la metodología propuesta ofrece una visión general de las respuestas y conceptos utilizados, lo que permitirá tomar decisiones informadas y justificadas para evaluar las competencias en éticas de estudiantes de la FCFM.

1. Introducción

1.1. Descripción de la problemática

En el marco de la enseñanza de la formación ética en la FCFM, se ha utilizado la aplicación EthicApp para la obtención y posterior análisis de las decisiones morales de estudiantes ante un dilema ético. Estos casos de estudio consisten en una problemática, que se plantea en forma de pregunta. Como respuesta a esta pregunta, los estudiantes manifiestan una preferencia en la forma de un número entre 1 y 6, donde los extremos de esta escala representan posturas dispares en cuanto a la decisión a tomar.

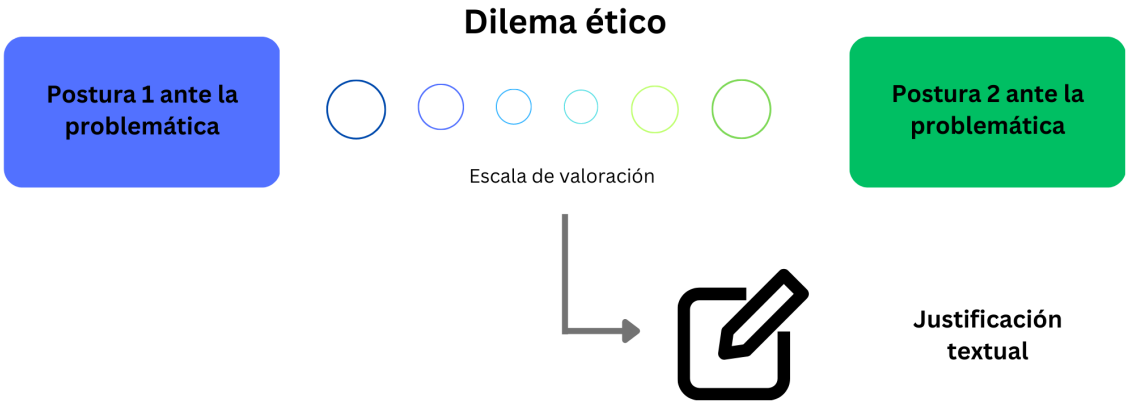


Figura 1: Estructura de respuestas de estudiantes.

Luego de una primera respuesta se realiza una instancia de deliberación grupal, donde se toma una nueva decisión en conjunto. Finalmente, los estudiantes otorgan nuevamente una calificación, que puede diferir de la señalada en las dos instancias anteriores.

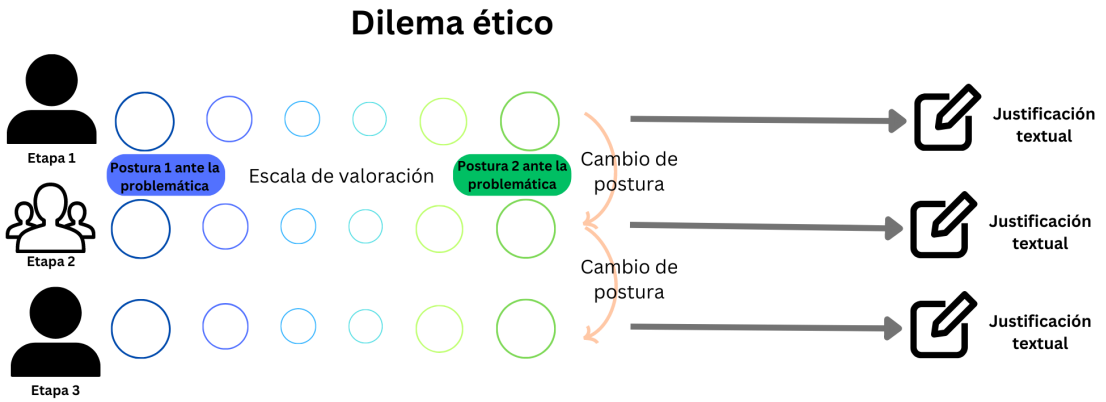


Figura 2: Iteraciones de las respuestas de estudiantes.

El Área de Ética de la FCFM ha llevado a cabo un completo análisis de las posturas de los estudiantes. No obstante, un aspecto difícil de procesar son las justificaciones que deben colocar luego de cada instancia de decisión. Pese a que análisis cualitativos de algunas respuestas han permitido plantear hipótesis preliminares respecto a los juicios morales y justificaciones de las decisiones, la gran magnitud de datos presentes dificultan la tarea de tomar conclusiones acerca de las competencias éticas exprimidas en la instancia, y como consecuencia la verificación de la adquisición de esta durante los cursos formativos de la escuela.

1.2. Proposición de solución

La ciencia de datos es una disciplina en constante crecimiento, en particular en nuestra facultad. En particular, los métodos basados en aprendizaje de máquinas han experimentado un aumento considerable en sus capacidades, ayudado por la mejora en capacidad de cómputo en las últimas décadas. Los algoritmos de procesamiento han sido una demostración de aquello, con modelos conversacionales como *chatGPT* tomando protagonismo entre los medios y el público.

Dentro de esta rama se encuentran los modelos de lenguaje, que intentan aproximar modelar uno o más lenguajes humanos al inducir una probabilidad a secuencias de palabras (frases, oraciones o documentos). Estos modelos se implementan usando redes neuronales profundas y siendo entrenados en grandes volúmenes de datos textuales (córpuses). Finalmente, un modelo sirve para resolver variadas tareas de procesamiento de texto, incluyendo clasificación, pregunta-respuesta e categorización/identificación de elementos relevantes de una secuencia. Entre las desventajas que presentan estos modelos están su costo de entrenamiento y capacidad limitada de interpretabilidad, ambas consecuencias de su gran tamaño.

Se propone la implementación de estos modelos pero también el uso de algoritmos más simples y más interpretables, para procesar los argumentos escritos por estudiantes en sus decisiones éticas. Más precisamente, se procederá a:

1. Explorar las justificaciones textuales de las respuestas usando técnicas de minería de datos.

2. Implementar modelos estadísticos para texto, que sean interpretables, para predecir la respuesta (número en la escala de 1 a 6) de los estudiantes utilizando el texto de justificación.

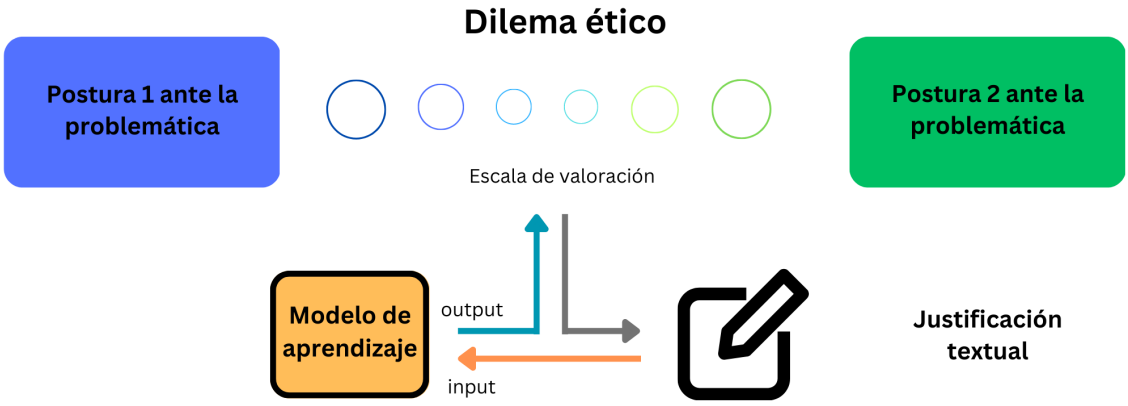


Figura 3: Ejemplo de modelo para predecir posturas de estudiantes.

3. Identificar, a través del texto, elementos semánticos que justifiquen los argumentos dados por los estudiantes. Esto usando tanto el análisis exploratorio de datos como los algoritmos.

Si es que los modelos muestran una buena capacidad de predicción, se pueden usar como herramienta que a la larga servirá para evaluar la progresión de competencia ética de los estudiantes con menor inversión humana. Esto es de particular relevancia para los objetivos finales del área de ética. Por otro lado, existe un interés en verificar hasta que punto los algoritmos pueden modelar relaciones semánticas complejas como lo son las justificaciones morales de estudiantes ante a una problemática. Este es un objetivo complementario y que logrará plantear nuevas perspectivas de investigación cualquiera sea el resultado, tanto del punto de vista computacional como del estudio de la ética.

Este proyecto consistirá de la utilización de diversas técnicas de modelamiento de lenguaje que permitan tomar una fotografía global de las justificaciones de estudiantes antes las preferencias escogidas. De manera sistemática se considerarán los datos del caso Adela, a modo de no sobrecargar el informe de información. No obstante, los códigos para cada caso pueden encontrarse en el repositorio de github del trabajo. Se provee un resumen del caso Adela a continuación:

En Chile, la deficiencia de vitamina D es un problema serio tanto en adultos mayores como en niños. Para abordar esta preocupante situación, un grupo de profesionales creó una startup que encontró una fruta ancestral de las comunidades diaguitas con alta concentración de vitamina D. Adela, una ingeniera del equipo, diseña el proceso de producción de un nuevo alimento a base de esta fruta. Sin embargo, se enfrenta a desafíos, ya que el árbol solo crece cerca de los ríos y necesita abundante luz solar, lo que dificulta llevarlo a zonas más australes afectadas por la sequía. Además, para conservar la vitamina D durante el transporte, el equipo decide liofilizar la fruta y agregar conservantes. Aunque aún no tienen la obligación legal de integrar a las comunidades diaguitas en el proyecto, Adela escucha sus preocupaciones sobre cómo estos cambios afectarían sus tradiciones. Aunque los cambios son necesarios para ayudar a quienes sufren deficiencia de vitamina D, las comunidades prefieren mantener sus prácticas tradicionales, ya que estas son parte fundamental de su identidad.

Adela debería priorizar:

(1) Producir el alimento contra el déficit

(6) Resguardar las tradiciones indígenas

2. Estado del arte

El procesamiento del lenguaje natural (NLP) ha ganado la atención de los medios y la sociedad en general en los últimos meses. Los últimos modelos conversacionales como chatGPT han generado mucha atención del público. A pesar de las capacidades que han demostrado, este tipo de modelos no son capaces de procesar ideas de la misma manera que lo hacen los humanos. Este es especialmente el caso de las decisiones que implican juicios morales.

En este contexto, responder éticamente a una pregunta se ha estudiado en el contexto de grandes modelos de lenguaje para mejorar su capacidad de ayudar a los seres humanos. Además, cuando entrenamos modelos para predecir respuestas similares a las humanas a dilemas éticos y otras tareas generales de NLP, a veces somos capaces de detectar patrones y estructuras que podrían explicar cómo las diferentes culturas enfrentan problemas morales.

Varios son los trabajos que buscan la predicción/identificación de elementos morales en texto natural. En esta sección abordaremos algunos de ellos. Primeramente, Garten et al. 2016 busca la detección automática de retóricas morales [1]. Para esto usan el diccionario de fundamentos morales [2] (de la Teoría de fundamentos morales TMF [3]), combinado con word embeddings. Otro trabajo similar es el de Xie et al. que evalua modelos en la clasificación de dilemas morales en los distintos fundamentos, gracias a lo cual concluyen que modelos de lenguaje tienen ventaja sobre modelos como las representaciones distribucionales [4].

Por otro lado, Kennedy et al. buscan la predicción de preocupaciones morales propias a un individuo usando evidencias de lenguaje moral escritas por este [5] (estados de Facebook de usuarios que hayan contestado el cuestionario de fundamentos morales). Se utilizaron distintas técnicas de procesamiento de lenguaje para predecir los puntajes obtenidos por los usuarios, para cada una de las dimensiones morales planteadas en la TMF [3]. Se destacan la variedad de métodos para vectorizar texto testeados, incluyendo *latent dirichlet allocation* (LDA) [6], *word embeddings*, conteo de ocurrencias del DFM [2] y BERT [7], que corresponde a un modelo de lenguaje profundo. Es este último que obtiene mejores resultados. Finalmente, tanto conteos de diccionario como LDA se usaron para interpretar que elementos linguísticos específicos explicaban cada fundamento por separado.

Los trabajos mencionados en este reporte muestran la alta variedad de formatos en los cuales se ha evaluado la presencia y grado de categorías morales, así como también la capacidad de distintos modelos de procesamiento de lenguaje natural para modelarlos. No obstante, ninguno de los artículos estudiados enfrenta un desafío tan específico como el nuestro. Los datos que poseemos tienen la ventaja de poder verse desde distintos ángulos, lo cual plantea dificultades pero también abre puertas a que las conclusiones que se puedan tomar sean reflejo de nuevos descubrimientos en el área. Para finalizar, muchos de los artículos nos confirman la pertinencia de los modelos a usar, en el caso de modelos de lenguaje, y nos sugieren algunos tipos de modelos más simples que tengan capacidad interpretativa.

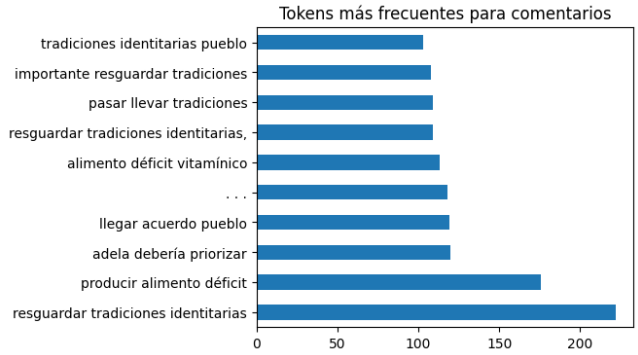
3. Análisis exploratorio de datos

Un análisis exploratorio adecuado se vuelve particularmente relevante a la hora de tomar conclusiones acerca de los datos textuales. Nos remitiremos al reporte de los largos de texto y n-gramas más frecuentes tanto de manera global como restringiéndonos a la etapa a evaluar o postura escogida por los estudiantes. A modo informativo se incluyen las frecuencias para cada uno de los casos en la tabla 1.

Tabla 1: Resultados de clasificación de posturas para caso Adela

Caso	**Cursos**	**Cantidad de estudiantes**	**Cantidad de grupos**
Julieta	**1**	*819*	247
Adela	**3**	**1866**	**515**
Laura	1	602	335
Alicia	2	1628	549

(a) 1-gramas más frecuentes



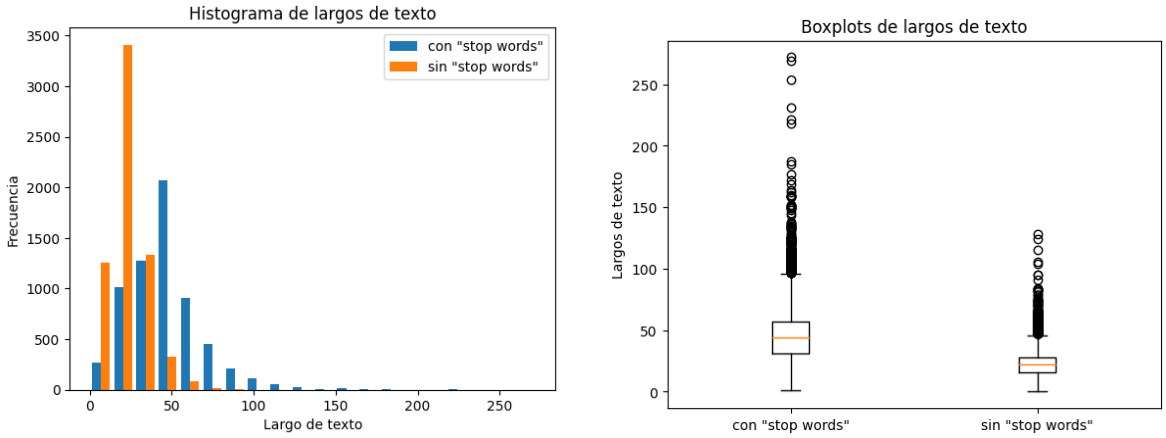
(b) 3-gramas más frecuentes

(c) Nube de palabras

Figura 4: Visualización de textos más frecuentes del dataset para el caso Adela

En la figura 4 se observan los n-gramas más frecuentes para $n \in \{1, 3\}$. Se ignoraron los 2-gramas pues en general eran concatenaciones de dos palabras que carecían de significado sin tener una tercera. En el caso de los 1-gramas, i.e., palabras más frecuentes, destacan “agua”, “recursos”

y "tradiciones", que son elementos importantes en el platemiento del dilema.



(a) Histogramas de largo de texto

(b) Boxplot de largos de texto

Figura 5: Visualización de largos de texto para caso Adela, con y sin *stop-words*.

En cuanto a los largos en general de los textos para el caso, estos son incluidos en la figura 5. Los largos de mensaje tienen un promedio de palabras inferior al 50 % para el caso Adela. También se incluye la cantidad de palabras que son relevantes dentro del texto. Para esto, se removieron las llamadas *stop-words*, que son palabras que no aportan información al mensaje y solo contribuyen a la grámaica de este (algunos ejemplos incluyen ciertos artículos y conjunciones).

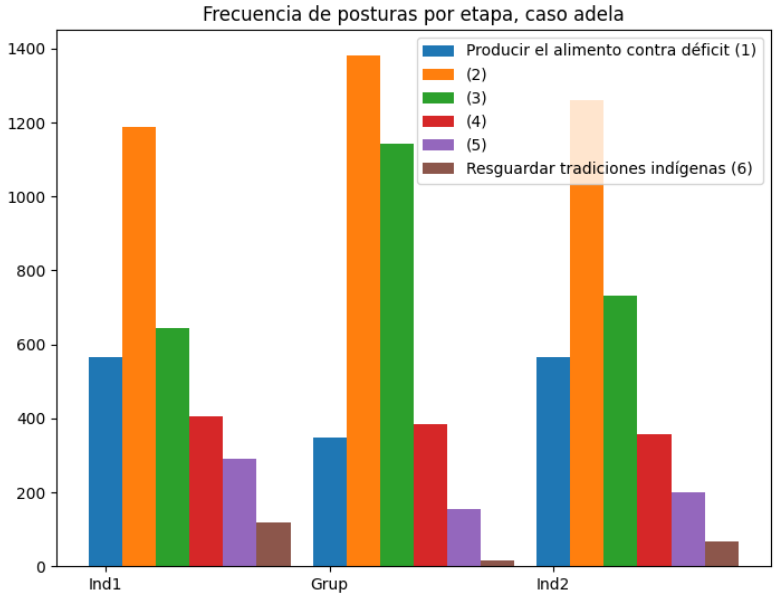


Figura 6: Frecuencias de posturas por etapa en caso Adela

Las frecuencias de cada elección son mostradas en la figura 6, donde se aprecia el hecho de que no hay un cambio brusco entre una etapa y otra, en especial entre las etapas individual 1 e individual 2. Recordemos que entre medio de esas existe una etapa grupal, para la cual se observa una mayor frecuencia de la postura (3), lo cual puede deberse al hecho de que estudiantes con posiciones similares están tratando de llegar a un consenso para colocar la valoración, lo cual por contraste disminuye la frecuencia de posiciones "extremas" como (1) y (6).

3.1. Condicionamiento a selección de postura y etapa de actividad

Una hipótesis a considerar en cuanto al rol del largo de palabras en la postura de estudiantes es que aquellas posiciones extremas tendrán una explicación más larga en extensión para explicarla. Por el contrario, se observa un efecto inverso, donde las posturas (1) y (6) tienen menor largo promedio que las preferencias más moderadas. Esto puede responder al hecho de que una persona absolutamente convencida de su postura vea menos necesidad de explicar esta. De cualquier modo la diferencia es sutil, por ende no se usará la variable implícita de largo de texto para las tareas de aprendizaje, pues se postula que aportará poca información.

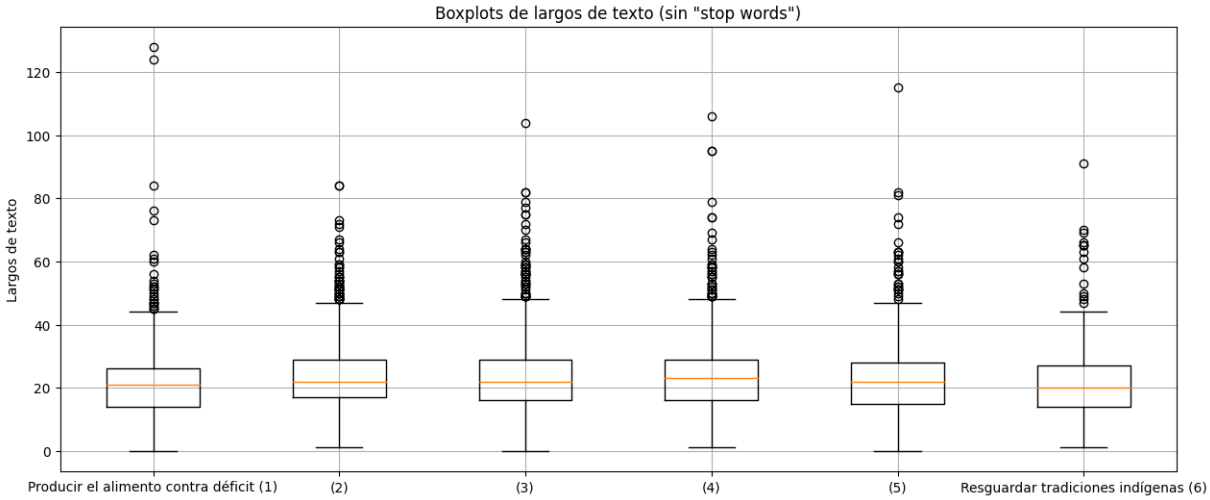


Figura 7: Boxplots de largos de texto para cada elección de postura.

Por otro lado, las palabras más frecuentes por cada postura se condicen con la justificación más comúnmente utilizada en cada caso. Se aprecia por ejemplo que para apoyar la producción del alimento se evocan conceptos como “salud” y “personas”, asu como también acciones como “producir” y “priorizar”, este último reflejando que estudiantes que asumen esa postura probablemente priorizan el bien común y/o salud de niños y ancianos en su decisión. En contraste, sustantivos como “cultura” y “pueblos” (en plural) aparecen al escoger resguardar las tradiciones. Así mismo, acciones como “resguardar”, “respetar” y “debería” son usadas con el mismo propósito. Este simple análisis nos permite también identificar palabras ampliamente repetidas sin importar la postura del estudiante, como es el caso de “tradiciones”.

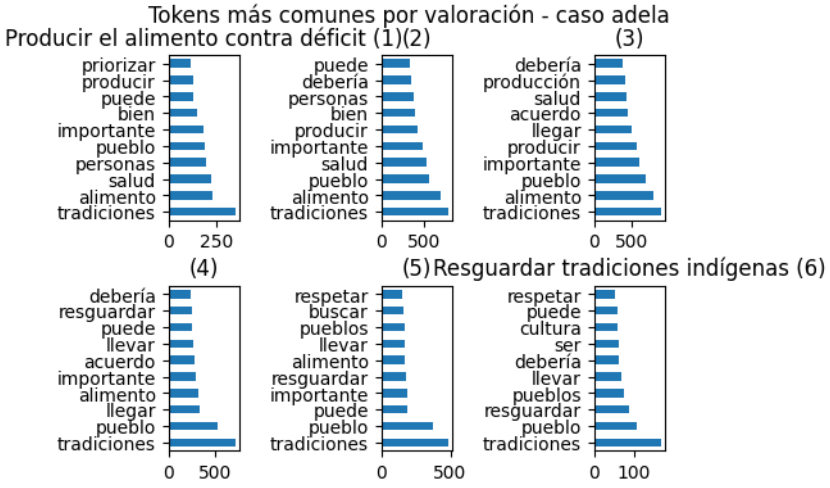


Figura 8: Frecuencias de posturas por etapa en caso Adela

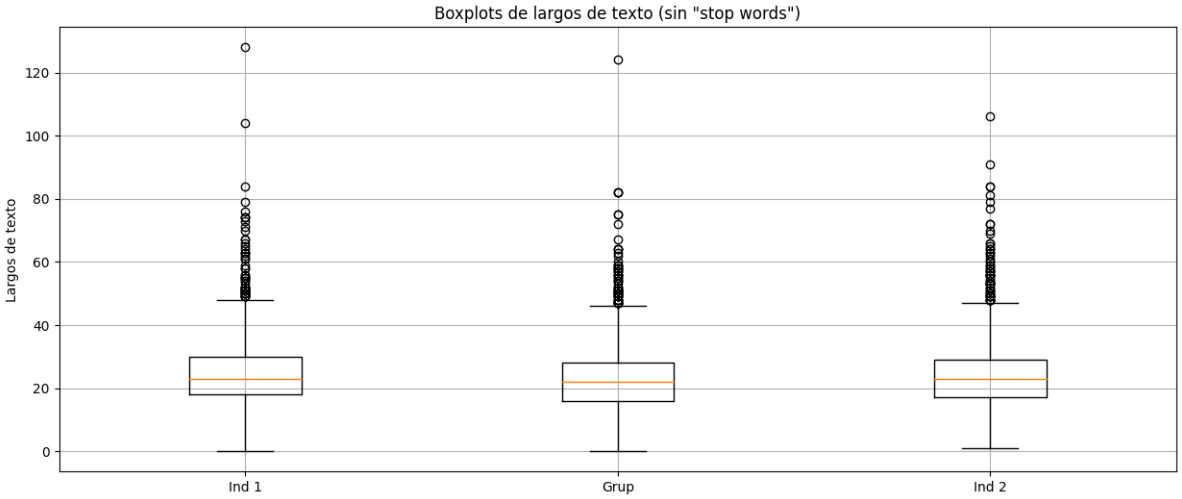


Figura 9: Boxplots de largos de texto para cada etapa de la actividad.

4. Metodología

4.1. Preprocesamiento

Se procedió a eliminar palabras que no aportasen contenido a la justificación en cuestión. Esta poda de palabras incluyó las llamadas *stop words*. Se consideró en un princió la *stemmisation* de palabras. Esto corresponde a dejar la raiz de modo que palabras con distintas conjugaciones o modos pero que denotan lo mismo correspondan a la misma variable. Esto no mejoró las métricas, por lo cual procedemos a dejar las palabras tal cual. Además se eliminó la puntuación, en particular los paréntesis, que agregaban ruido especialmente para la modelación con tópicos. Estos elementos, sumados a la limpieza de ciertos elementos no informativos en el dataset contribuyeron a la sustancial mejora de las métricas mostradas en la presentación parcial, como se discute más adelante.

4.2. Predicción supervisada de posturas con *Naive-Bayes*

Los métodos de Naive Bayes son un conjunto de algoritmos de aprendizaje supervisado basados en la aplicación del teorema de Bayes con la suposición "naive" de independencia condicional entre cada par de características dadas las clases variables. Consideramos vectores de n dimensiones de ocurrencias introducidas anteriormente, cada uno correspondiente a un documento sobre un vocabulario de longitud n. Dado (un documento) $x = (x_1, \ldots, x_n)$, nos gustaría etiquetarlo con una de las siguientes etiquetas: $C_1, \ldots, C_k, \ldots, C_K$. Gracias al teorema de Bayes, se tiene que:
$\forall k \in \{1, \ldots, K\}$

$$\begin{aligned}\mathbb{P}(C_k|x_1, \ldots, x_n) &= \frac{\mathbb{P}(C_k)\mathbb{P}(x_1, \ldots, x_n|C_k)}{\mathbb{P}(x_1, \ldots, x_n)} \\ &\propto \mathbb{P}(C_k)\mathbb{P}(x_1, \ldots, x_n|C_k)\end{aligned}$$

El término $\mathbb{P}(C_k)$ se puede estimar con la frecuencia de la etiqueta C_k en los datos. Para calcular la probabilidad de verosimilitud $\mathbb{P}(x_1, \ldots, x_n|C_k)$, primero asumimos una distribución de probabilidad (como la Gaussiana o Multinomial). En el caso de clasificación binaria ($k \in 0, 1$)

utilizando una distribución de Bernoulli, obtenemos, para cada característica j (palabra):

$$\mathbb{P}(x_j|C_k) = \mathbb{P}(j|C_k)x_j + (1 - \mathbb{P}(j|C_k))(1 - x_j)$$

$\mathbb{P}(j|C_k)$ también se puede estimar, tomando la proporción de documentos que contienen la palabra j entre los de clase C_k. Este modelo permite la detección de spam y el análisis de sentimientos, y tiene un rendimiento aceptable en el caso de textos cortos [8]. Además, es posible interpretar la probabilidad que cada palabra le asigna a cada clase. En efecto, ese vector de probabilidad discreto nos permite identificar que palabras son las que más "empujan" hacia alguna elección en específico.

4.3. Modelamiento no supervisado con *Latent Dirichlet Allocation*

La modelización de tópicos (*topic modelling* en ingles) se refiere a un tipo de modelado estadístico en el procesamiento del lenguaje natural que tiene como objetivo extraer la estructura semántica oculta de un texto. La suposición subyacente es que un documento está compuesto por una mezcla de temas abstractos, y los métodos existentes infieren esos temas basándose en las palabras de cada documento. La Asignación Latente de Dirichlet (LDA, por sus siglas en inglés) es el modelo de temas más ampliamente utilizado. Los documentos se representan como mezclas aleatorias de temas latentes. De manera similar, los temas se caracterizan por una distribución sobre todas las palabras.

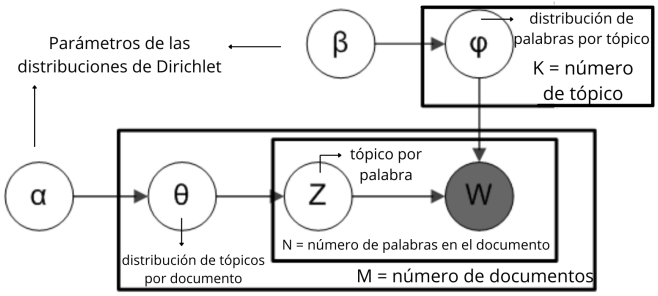


Figura 10: Diagrama de un modelo de tópicos LDA aplicado a texto.

Dado que el número de temas a considerar en LDA es un hiperparámetro que debe establecerse de antemano, utilizaremos la familia de métricas çoherencia", que se suele usar para evaluar la calidad e interpretabilidad de los temas generados por un modelo de temas. Esta familia de métricas utiliza las relaciones entre palabras dentro de cada tópico y está diseñada para capturar la similitud semántica entre palabras dentro de un tema, con el objetivo de identificar temas que contengan palabras que tienden a frecuentemente al mismo tiempo en documentos. Los temas con una mayor coherencia se consideran más interpretables y más propensos a representar un concepto concreto.

De las diversas métricas de coherencia comúnmente utilizadas en el modelado de temas, usaremos la coherencia C_v. Esta métrica calcula la coherencia basada en la co-ocurrencia de palabras dentro de un tema, y busca medir en qué medida las palabras en un tema tienden a aparecer juntas en el mismo contexto. Para más detalles sobre cómo se calcula esta métrica, nos referimos a *Syed & Spruit (2017)* [9].

Para poder visualizar los tópicos de manera adecuada se usarán los siguientes conceptos vistos en el curso:

- **Teoría de la información**: Se utiliza la divergencia de Jensen-Shanno, que es una métrica de disimilitud entre distribuciones de probabilidad. Dadas dos distribuciones de probabilidad,

p y q, la divergencia de Jensen-Shannon se define de la siguiente manera:

$$\frac{1}{2}(KL(p\|m) + KL(q\|m))\,,$$

donde $m = \frac{1}{2}(p+q)$ es la distribución promedio y KL denota la divergencia de Kullback-Leibler.

- Reducción de dimensionalidad: se utiliza el método *multidimensional scalling*, que puede traducirse como "escalamiento multidimensional". Esta es una reducción de dimensionalidad no lineal en la cual la proyección encontrada pretende que las distancias en el nuevo espacio (en este caso de dos dimensiones) difiera lo menos posible de las distancias en el espacio original.

$$\frac{1}{2}(KL(p\|m) + KL(q\|m))$$

Esto en base a la metodología de visualización de tópicos que se propone en *Sievert & Shirley (2014)* [10]. El algoritmo, llámese como LDAvis, presenta una vista global del modelo de temas que responde preguntas sobre la prevalencia de cada tema y cómo se relacionan entre sí. En esta visualziación, los temas se representan como círculos en un plano, y los centros de los círculos se determinan calculando la distancia entre los temas y utilizando *multidimensional scaling* para proyectar las distancias inter-temas en dos dimensiones. La distribución de términos de un conjunto de temas se define como el promedio ponderado de las distribuciones de términos de los temas individuales en el conjunto.

5. Resultados

5.1. Tarea de clasificación

La tarea de clasificación arroja las siguientes métricas para los cuatro casos estudiados, mostrados en la tabla 2.

Tabla 2: Resumen de resultados de clasificación por caso

	Métrica multiclase				**Métrica binaria**			
caso	**precisión**	**sensib.**	**f1**	**exactitud**	**precisión**	**sensib.**	**f1**	**exactitud**
Julieta	0.36	*0.37*	0.36	*0.37*	0.70	*0.70*	0.70	*0.70*
Adela	**0.43**	**0.43**	**0.43**	**0.43**	**0.79**	**0.79**	**0.79**	***0.79***
Laura	0.47	*0.48*	0.48	*0.48*	0.77	*0.77*	0.77	*0.77*
Alicia	0.44	*0.45*	0.44	*0.45*	0.83	*0.83*	0.83	*0.83*

Por otro lado, mostramos el detalle de la clasificación para el caso Adela. La tabla 3 muestra la tarea de clasificación fina multiclase, mientras que la versión binaria se muestra en la tabla 4.

Para entender la naturaleza de la complejidad de la clasificación fina, se muestra la matriz de confusión en la figura 11. Se aprecia como hay una ligera tendencia a etiquetar texto como de clase (2), que es aquella mayoritaria. Esto es un problema recurrente en cualquier tipo de modelo,

Tabla 3: Resultados de clasificación de posturas para caso Adela

	precision	**recall**	**f1-score**	**support**
(1) Producir el alimento contra el déficit	0.40	0.38	0.39	157
(2)	0.47	0.47	0.47	363
(3)	0.47	0.54	0.50	434
(4)	0.35	0.32	0.33	257
(5)	0.41	0.40	0.41	173
(6) Resguardar las tradiciones indígenas	0.37	0.19	0.25	57
accuracy			0.43	1441
macro avg	0.41	0.38	0.39	1441
weighted avg	0.43	0.43	0.43	1441

Tabla 4: Resultados de clasificación binaria de posturas para caso Adela

	precision	**recall**	**f1-score**	**support**
(1) Producir el alimento contra el déficit	0.85	0.83	0.84	954
(6) Resguardar las tradiciones indígenas	0.69	0.71	0.70	487
accuracy			0.79	1441
macro avg	0.77	0.77	0.77	1441
weighted avg	0.79	0.79	0.79	1441

pero lo es con mayor razón en un método Naive-Bayes, pues la predicción consiste en el voto de las palabras presentes, que naturalmente aparecerán más en aquellas clases recurrrentes.

Cabe señalar que este efecto se reduzco drásticamente respecto a la versión mostrada en la presentación de avance. En esta, la matriz de confusión estaba radicalmente cargada hacia la clase mayoritaria, confundiendo cualquier clase con la clase (2). En esta versión la confusión ocurre más fuertemente con las clases vecinas (1) y (3), que corresponden a grados de convencimiento de la misma postura. Esto grafica la dificultad de la clasificación fina respecto a la versión binaria. La mejora de esta situación y de las métricas de la clasificación multiclase en general se le atribuye al procesamiento adecuado realizado al texto, no realizado en una primera instancia (incluyendo la elimación de ruido, separación de casos en Adela y eliminación de puntuación). A modo de referencia, la sensibilidad de la clase (2) antes de aplicar la metodología completa fue de 86 %, contra menos del 6 % de las clases (1), (2) y (6). Se conjetura que realizar un sobresampleo o subsampleo podría mejorar lo anterior aún más, pero nos conformamos con la mejora señalada en el la versión actual.

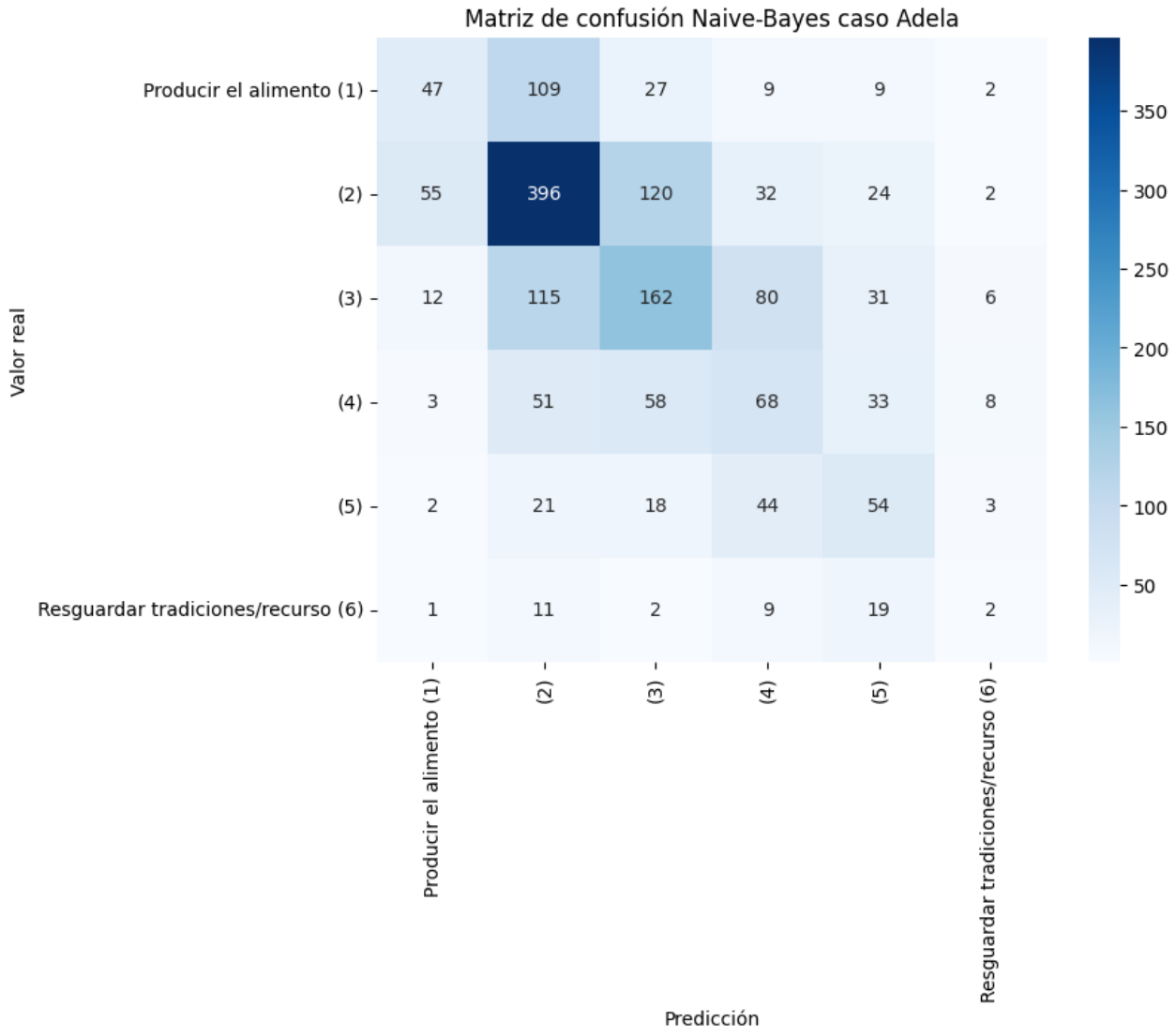


Figura 11: Matriz de confusión de clasificación caso Adela.

5.2. Interpretabilidad de modelos

5.3. Modelamiento de tópicos

Se obtuvieron los siguientes números de tópicos en cada case estudiado:

- Julieta: 2 tópicos
- Adela: 4 tópicos
- Laura: 4 tópicos
- Alicia: 4 tópicos

En la tabla 6 reportamos tanto la magnitud de cada tópico como los principales tokens que lo componen para el caso Adela.

Tabla 5: Palabras con más probabilidad por cada clase

	Producir el alimento		**Resguardar tradiciones**	
	Palabra	**Probabilidad**	**Palabra**	**Probabilidad**
1	***salvar***	0.9850757167193909	***siglos***	0.8807510221140485
2	***juego***	0.9589416426994265	***suplementos***	0.8630617786067957
3	***vidas***	0.9525060938146097	***sol***	0.8253861418636597
4	***tribu***	0.9507237155401059	***existen***	0.8227740010872212
5	***religiosas***	0.9481281470801085	***única***	0.7700217160507917
6	***poniendo***	0.9420201060316521	***usuario***	0.761934254557929
7	***riesgo***	0.9383929112652638	***consentimiento***	0.761934254557929
8	***tiempos***	0.9342815991664724	***integridad***	0.7303224790335253
9	***cambian***	0.9342815991664724	***sacar***	0.7242191780159098
10	***ayudaria***	0.9342815991664724	***obtener***	0.723129422679366

Por otro lado en la figura 12 mostramos una visualización de los tópicos obtenidos utilizando la biblioteca de python *pyLDAvis* [10]. Esta nos muestra un grado relativamente alto de sobreposición entre los tópicos 1 y 2, que son los más numerosos. Esto se condice con la similitud de los términos utilizados, poniendo en cuestión la optimalidad del número de tópicos encontrado con la métrica de coherencia. Se destaca el tópico número dos, con lo cual a la dereche de la figura se pueden ver la frecuencia relativa de las palabras más importantes de tal tema encontrado.

Tabla 6: Términos más importantes por tópico.

	Tópico (porcentaje de palabras)			
Prioridad	**Tópico 1 (42,4 %)**	**Tópico 2 (33,5 %)**	**Tópico 3 (14,9 %)**	**Tópico 4 (9,2 %)**
1	*tradiciones*	tradiciones	*tradiciones*	pueblo
2	*alimento*	alimento	*fruto*	si
3	*pueblo*	personas	*pueblo*	originario
4	*llegar*	salud	*fruta*	producción
5	*importante*	importante	*si*	acuerdo
6	*acuerdo*	pueblo	*pueblos*	producto
7	*llevar*	producir	*ser*	bien
8	*producción*	niños	*proyecto*	alimento
9	*originario*	priorizar	*originarios*	puede
10	*Producir*	puede	*manera*	tradición

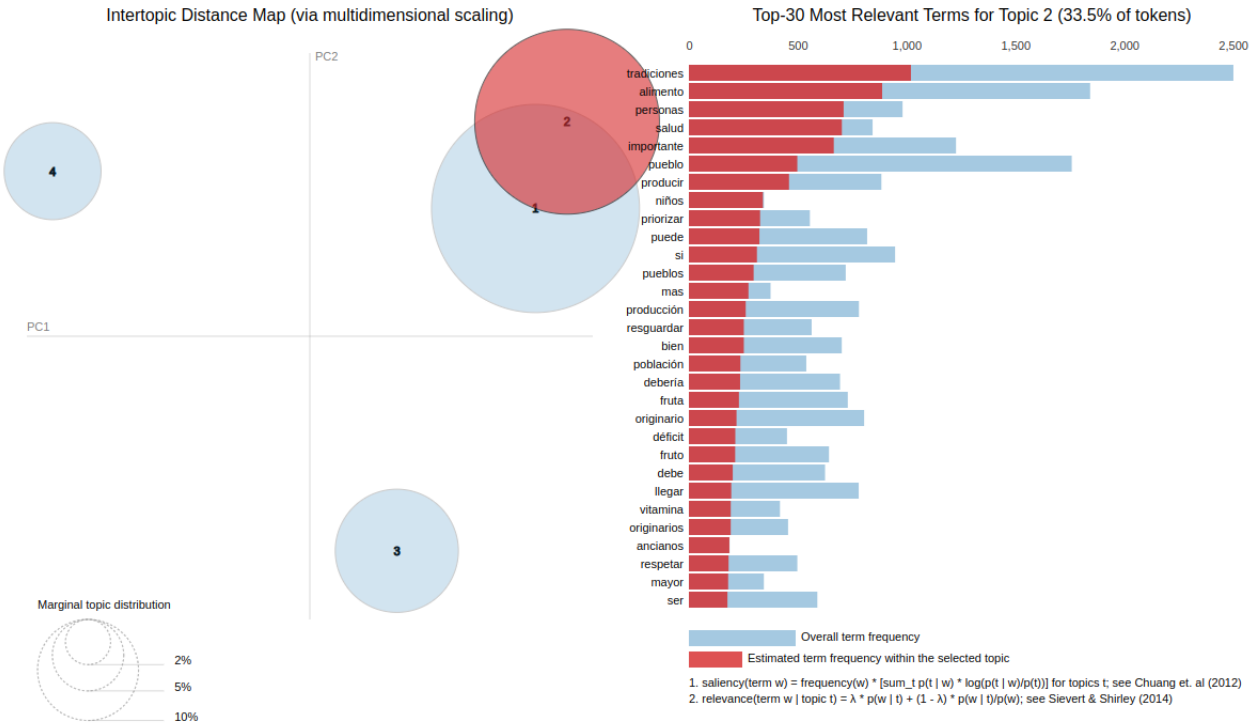


Figura 12: Visualización de tópicos ajustados con LDA.

6. Trabajo Futuro

Lamentablemente no se abordaron la gran variedad de técnicas a disposición para enfrentar este proyecto. Además, la naturaleza del *dataset* permitía plantear distintas tareas de predicción, cosa que se mencionó en las presentaciones del curso.

1. Ampliar el respectro de modelos utilizados tanto para clasificación como para modelamiento de tópicos, incluyendo modelos profundos. Comparar la capacidad de predicción tanto con los algoritmos básicos como con la capacidad humana. En particular, se mencionan algunos métodos que no se alcanzaron a abordar en este trabajo:

    - Topic Modelling con *BERT Topic*
    - Modelos basados en árboles usando vectores basados en la ocurrencia de los distintos tópicos.
    - Utilización de modelos de lenguaje para la vectorización de las distintas respuestas de texto.

2. Implementar los modelos utilizados en este trabajo pero para la predicción del cambio de respuesta de una etapa a otra, usando las justificaciones de la etapa intermedia y última etapa. Además, se pretende identificar elementos semánticos en cambios de valoraciones entre distintas etapas de la actividad, tanto con elementos diferentes como comunes entre ambas justificaciones.

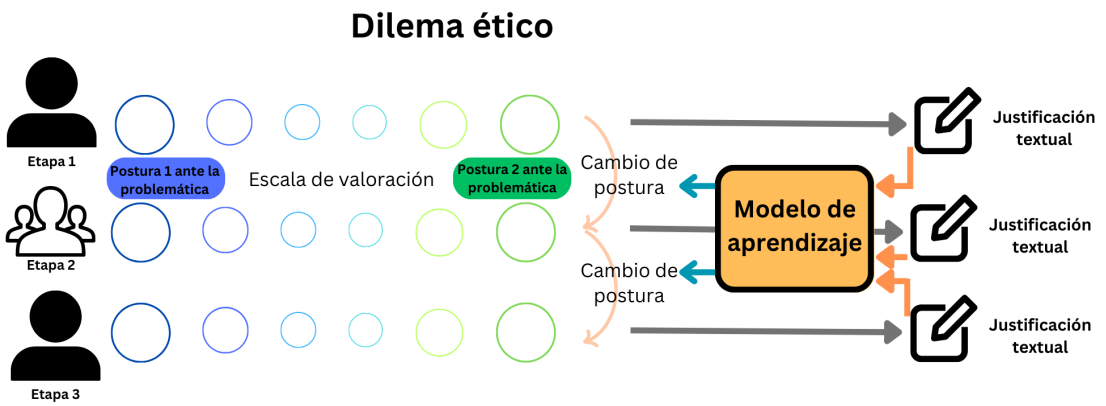


Figura 13: Ejemplo de modelo para predecir cambios de postura.

3. Utilizar modelos predictivos de texto para predecir el grado de competencia ética en las justificaciones, utilizando tanto técnicas simples como avanzadas de procesamiento de lenguaje natural. Incluir modelos de reconocimiento de entidades para la identificación automática de elementos textuales que denoten elementos positivos y negativos en cuanto a la calidad de la respuesta otorgada. Notemos sin embargo que estas tareas requieren la creación de un dataset con etiquetas especiales.

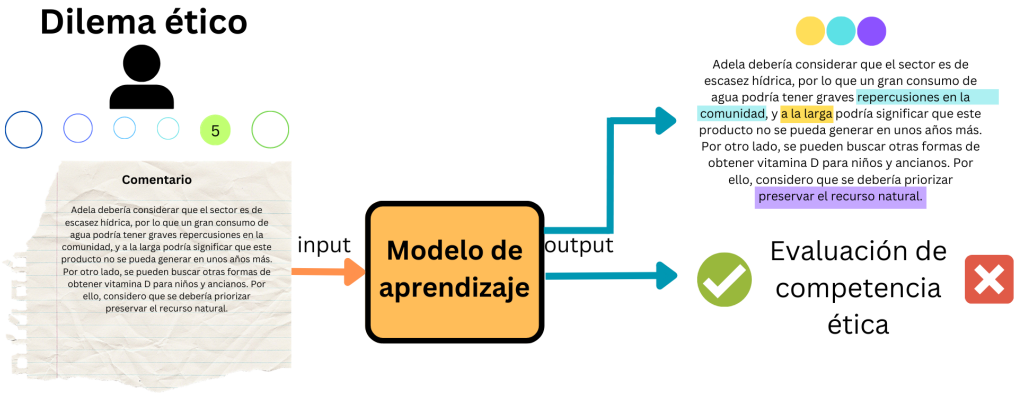


Figura 14: Ejemplo de modelo para asistencia a la evaluación de competencia ética.

7. Conclusiones

Los modelos usados en este trabajo fueron modelos simples basados en conteos de palabras, pero cuya formulación probabilística los dota de la capacidad de clasificar nuevos textos, tanto de manera supervisada como no supervisada. Estas técnicas son además interpretables. En el caso específico de LDA y Naive-Bayes, tenemos tanto una distribución de palabras por clase como una distribución de clase por palabra, lo cual es particularmente útil para el equipo de la unidad de ética tome conclusiones respecto al lenguaje empleado en las decisiones de estudiantes. Esto marca un contraste con los métodos de aprendizaje profundo. Estos han mostrado una gran capacidad predictiva, pero esta capacidad es inútil si en este caso particular, pues la predicción como tal no es de interés directo de la unidad de ética.

Por lo demás, el desarrollo de un método de aprendizaje profundo que si sea capaz de declarar aquellas variables (deseablemente, conceptos) que sean destacables a la hora de reflejar la competencia ética, es un trabajo futuro a considerar. Existe una buena perspectiva de que funcione dada la naturaleza de los textos que han mostrado tanto el análisis exploratorio como los modelos utilizados y sus métricas de desempeño. En particular modelos profundos podrían captar situaciones semánticas que modelos basados en conteo no logran reflejar, como lo son las negaciones.

En cualquier caso, el presente trabajo muestra una metodología con la capacidad de dar una fotografía global de la naturaleza de las respuestas y los conceptos empleados para justificarla. Se espera que esta herramienta ayude a graficar la coherencia de las respuestas cada año, con lo cual se pueden tomar decisiones informadas y justificadas respecto a la metodología para la enseñanza de la ética en la facultad de ciencias físicas y matemáticas.

Referencias

[1] Garten, J., Boghrati, R., Hoover, J., Johnson, K. M., Dehghani, M. (2016). Morality Between the Lines: Detecting Moral Sentiment In Text. Proceedings of IJCAI 2016 Workshop on Computational Modeling of Attitudes.

[2] Graham, J., Haidt, J., Nosek, B. A. (2009). Liberals and conservatives rely on different sets of moral foundations. Journal of Personality and Social Psychology, 96, 1029–1046. https://doi.org/10.1037/a0015141

[3] Haidt, J. (2007). The New Synthesis in Moral Psychology. Science, 316(5827), 998–1002. https://doi.org/10.1126/science.1137651

[4] Xie, J. Y., Hirst, G., Xu, Y. (2020). Contextualized moral inference (arXiv:2008.10762). arXiv. https://doi.org/10.48550/arXiv.2008.10762

[5] Kennedy, B., Atari, M., Mostafazadeh Davani, A., Hoover, J., Omrani, A., Graham, J., Dehghani, M. (2021). Moral concerns are differentially observable in language. Cognition, 212, 104696. https://doi.org/10.1016/j.cognition.2021.104696

[6] Blei, D. M., Ng, A. Y., Jordan, M. I. (2003). Latent dirichlet allocation. The Journal of Machine Learning Research, 3(null), 993–1022.

[7] Devlin, J., Chang, M.-W., Lee, K., Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), 4171–4186. https://doi.org/10.18653/v1/N19-1423

[8] Metsis, V., Androutsopoulos, I., Paliouras, G. (2006, July). Spam filtering with naive bayes-which naive bayes?. In CEAS(Vol. 17, pp. 28-69)

[9] Syed, S., Spruit, M. (2017, October). Full-text or abstract? examining topic coherence scores using latent dirichlet allocation. In 2017 IEEE International conference on data science and advanced analytics (DSAA) (pp. 165-174). IEEE. https://doi.org/10.1109/DSAA.2017.61

[10] Sievert, C., Shirley, K. (2014, June). LDAvis: A method for visualizing and interpreting topics. In Proceedings of the workshop on interactive language learning, visualization, and interfaces (pp. 63-70). https://github.com/bmabey/pyLDAvis