Medidas de importancia en textos mediante *DistilBert* para la detección de la adicción al juego

Exploración de mecanismos de auto atención de variables de textos.

- Segundo Avance -

Sacbe García García

Carlos Cuauhtemoc Gutiérrez Salazar

Centro de Investigacion en Matematicas, A.C.

16 de Noviembre, 2023

Contenido

Datos eRisk Ludopatía

```
<INDIVIDUAL>
  <ID>subject1</ID>
  <WRITING>
    <TITLE>Is it true that nowadays people view virgin girls in 30s as weirdo?</TITLE>
    <DATE>2022-10-24 16:10:32</DATE>
    <TEXT>[eliminado]</TEXT>
    <INFO>Reddit post</INFO>
  </WRITING>
  <WRITING>
    <TITLE />
    <DATE>2022-10-24 16:30:11</DATE>
    <TEXT>Used to be A Level candidate here. Economics doesn't really teach much out of demand and supply so it doesn't
matter if you take this paper or not. The best would be Pure Maths, Further Maths and Computing. Maybe Accounting too
if you are a Business stream. But really, you just need Further maths.</TEXT>
    <INFO>Reddit post</INFO>
  </WRITING>
  <WRITING>
    <TITLE />
    <DATE>2022-10-24 16:41:15</DATE>
    <TEXT>To be fair I feel like the good thing about Singapore is that it is less prevalent here. When I was in Tw/
Tokyo/Korea/China it was way worse, it might be because most girls are really pretty/well endowed and all dressed up
over there.</TEXT>
```

Figure: Datos originales

	user	date	title	text	comment	cat
0	6157	2021-08-25 13:03:25	NaN	He is just playing GTA 5 man	. He is just playing GTA 5 man	0
1	6157	2021-08-26 12:44:29	NaN	kariyu will stare than play	. kariyu will stare than play	0
2	6157	2021-08-26 12:45:56	NaN	me in my dreams be like:	. me in my dreams be like:	0
3	6157	2021-08-27 10:19:17	NaN	727	. 727	0
4	6157	2021-08-29 18:38:24	NaN	Doom bc that game is soooo good i liked the gu...	. Doom bc that game is soooo good i liked the ...	0
...	...	...	...	...	...	...
1102866	3988	2022-10-24 19:34:52	Pussy eaters here's a gift	NaN	Pussy eaters here's a gift.	0
1102867	3988	2022-10-24 19:43:31	Can this doll sit on your face?	NaN	Can this doll sit on your face?.	0
1102868	3988	2022-10-24 19:45:16	Born to have your dick in me	NaN	Born to have your dick in me .	0
1102869	3988	2022-10-24 20:10:54	I need pussy eaters	NaN	I need pussy eaters.	0
1102870	3988	2022-10-24 20:13:17	Here's a gift for pussy eaters	NaN	Here's a gift for pussy eaters.	0

1102871 rows × 6 columns

Figure: Datos Formateados

Tokenización

Tokenizer: WordPiece

1. **WordPiece** es el algoritmo de tokenización de *subpalabras* utilizado para BERT, DistilBERT y Electra. El algoritmo se describió en **Japanese and Korean Voice Search** (Schuster et al., 2012) y es muy similar a BPE. WordPiece primero inicializa el vocabulario para incluir todos los caracteres presentes en los datos de entrenamiento y aprende progresivamente una cantidad determinada de reglas de combinación. A diferencia de BPE, WordPieza no elige el par de símbolos más frecuente, sino el que maximiza la probabilidad de que los datos de entrenamiento se agreguen al vocabulario.

[CLS]	-100
S	0
##po	-100
##nge	-100
Bob	-100
chocolate	-100
milk	-100
(	-100
lost	-100
episodes	-100
)	-100
.	-100
[SEP]	-100
[PAD]	-100
[PAD]	-100
[PAD]	-100
[PAD]	-100
[PAD]	-100

Figure: Conversión de texto a tokens

Offsets mapping

1. Cada token generado por un tokenizer tiene un inicio y un final en el texto original. **El offset mapping es un par de números que indica la posición de inicio y fin de un token en el texto original.**
   Esto es esencial para mapear los resultados del modelo nuevamente al texto original después de la tokenización.

Análisis Exploratorio

Frecuencia de palabras

```
like: 83536
would: 58774
one: 54940
get: 50563
people: 46898
https: 40509
think: 39287
time: 37140
know: 35707
good: 34126
```

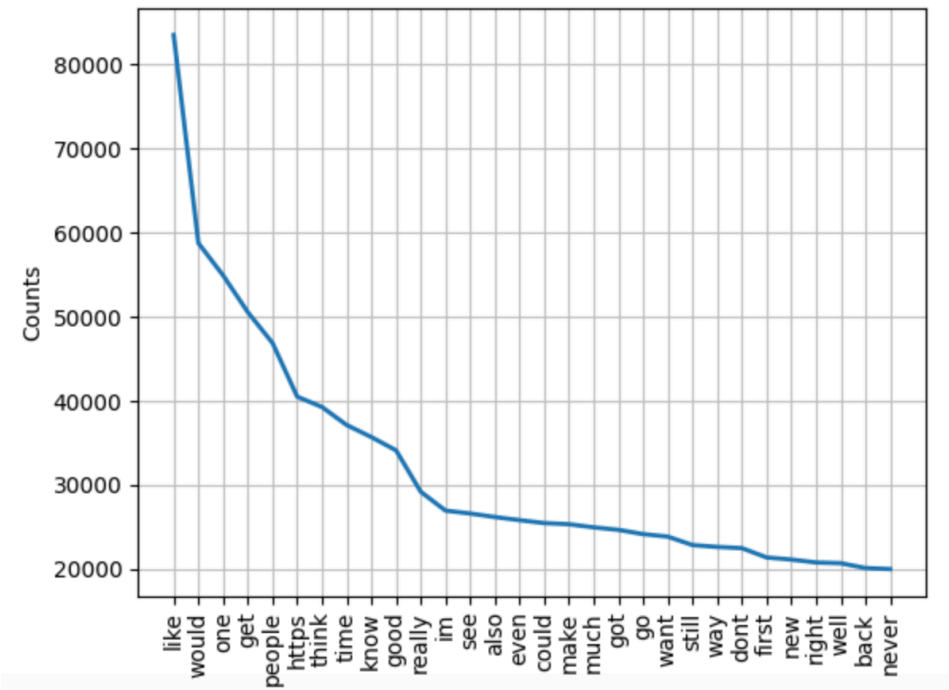

Figure: Frecuencia de tokens en dataset general

Nubes de palabras no diagnosticados

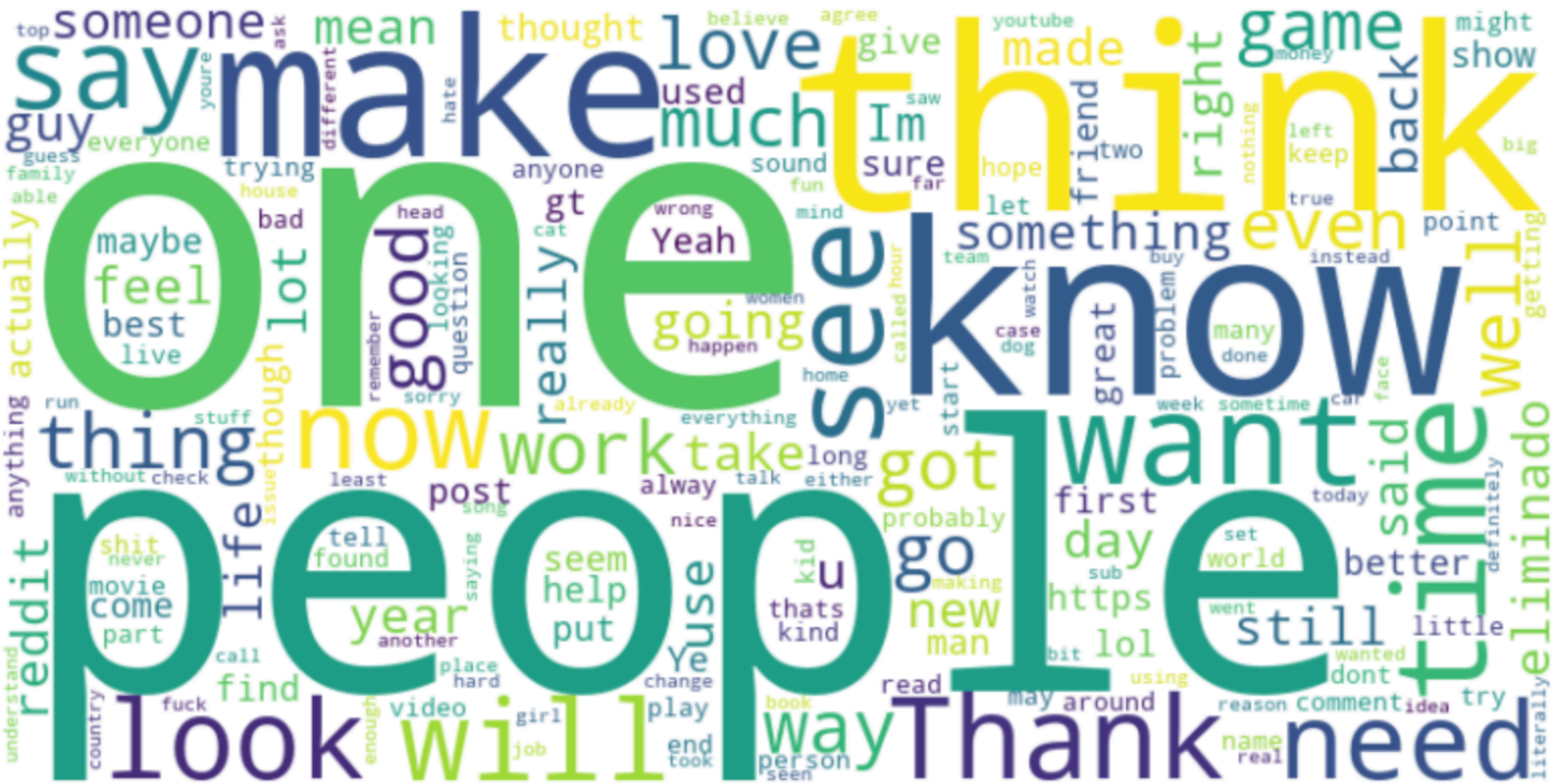

Figure: Nube de palabras de no ludopatías

Nubes de palabras diagnosticadas

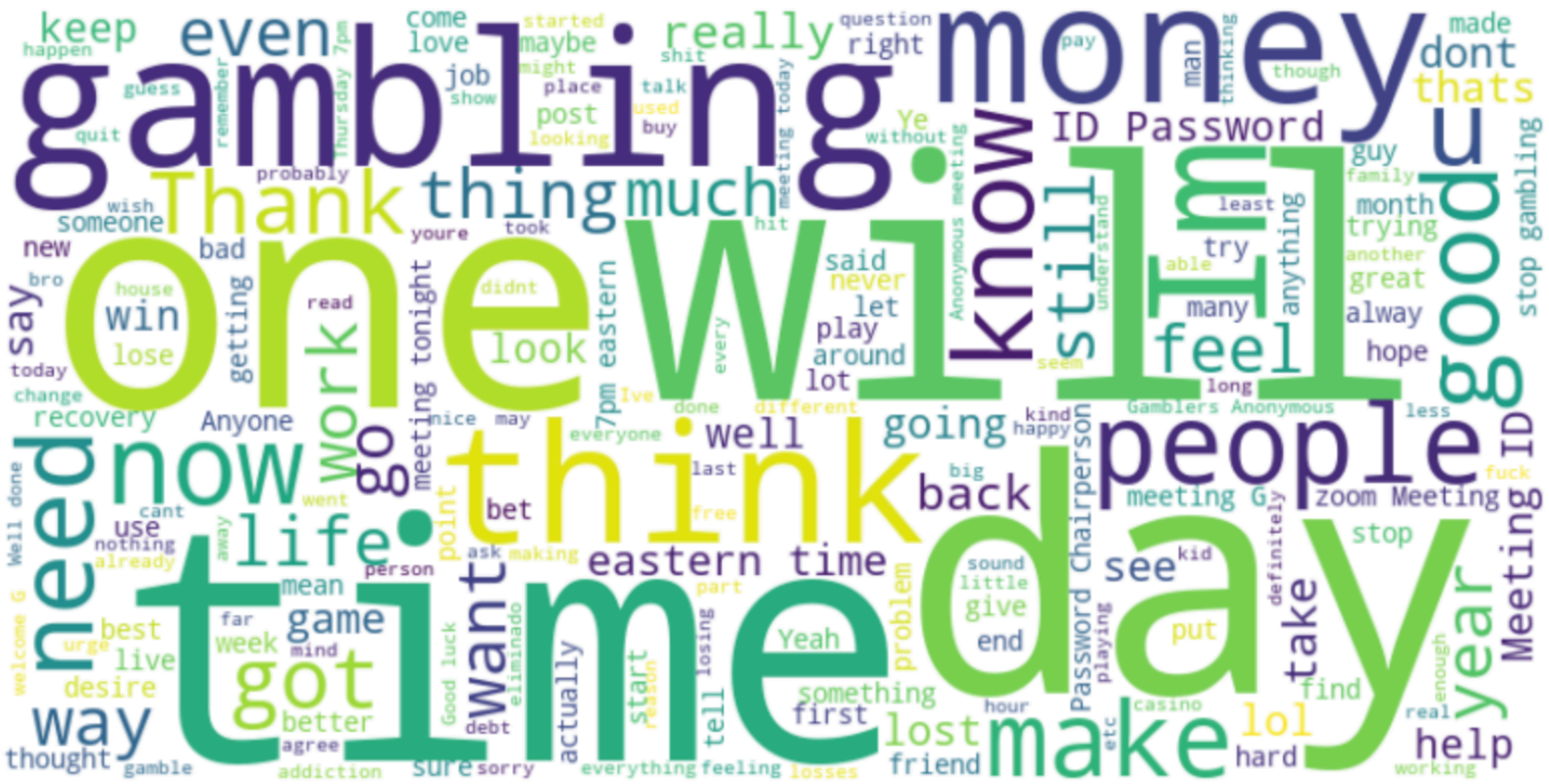

Figure: Nube de palabras de no ludopatías

Tópicos de individuos diagnosticados

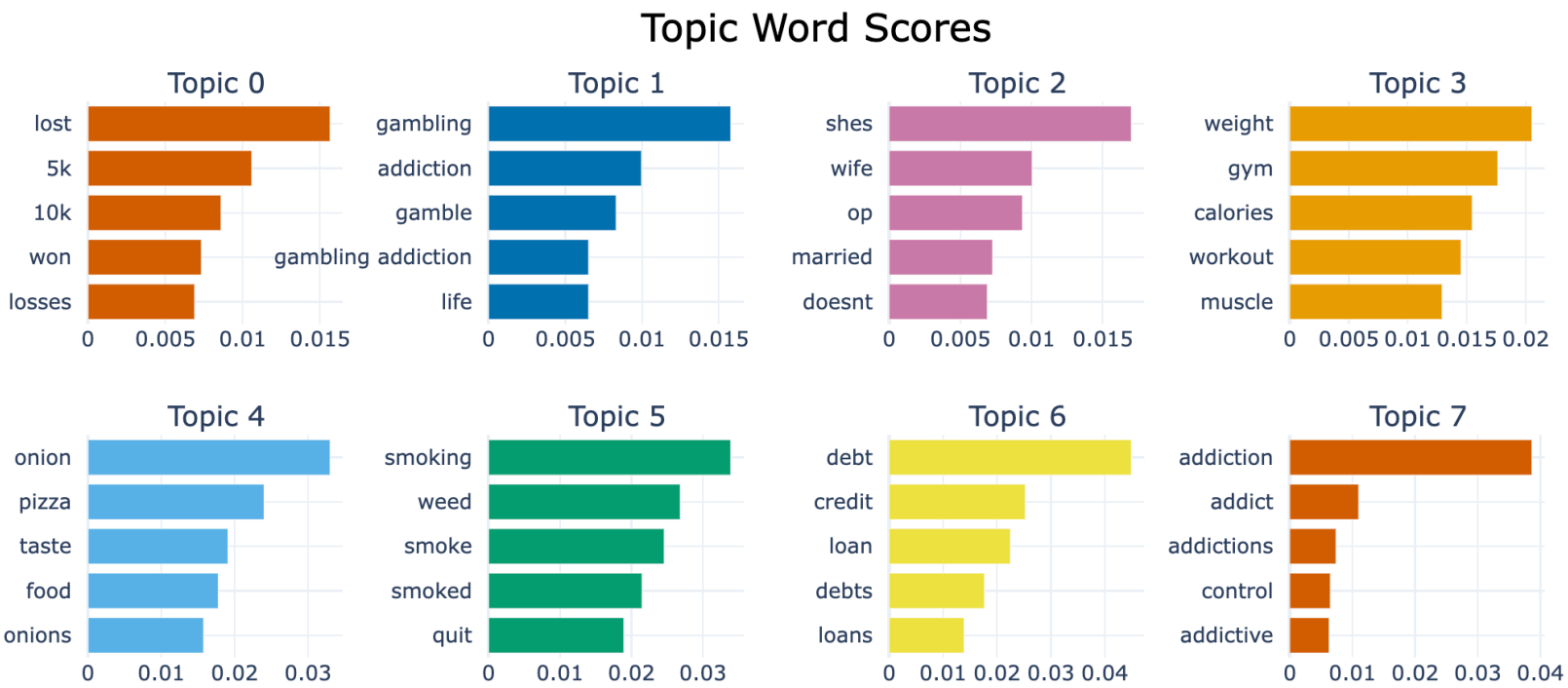

Figure: Tópicos más relevantes

Tópicos de individuos no diagnosticados

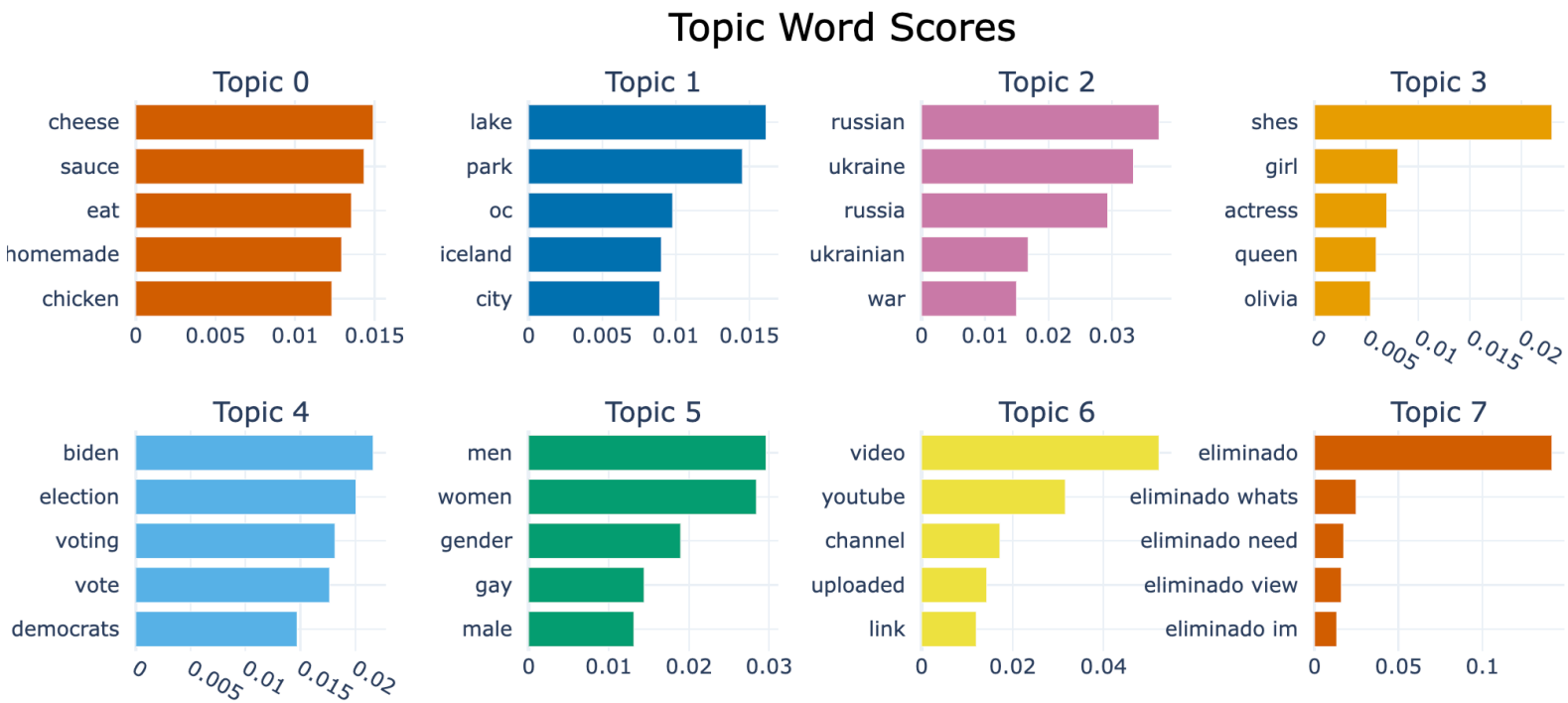

Figure: Tópicos más relevantes

Conteo de categorías

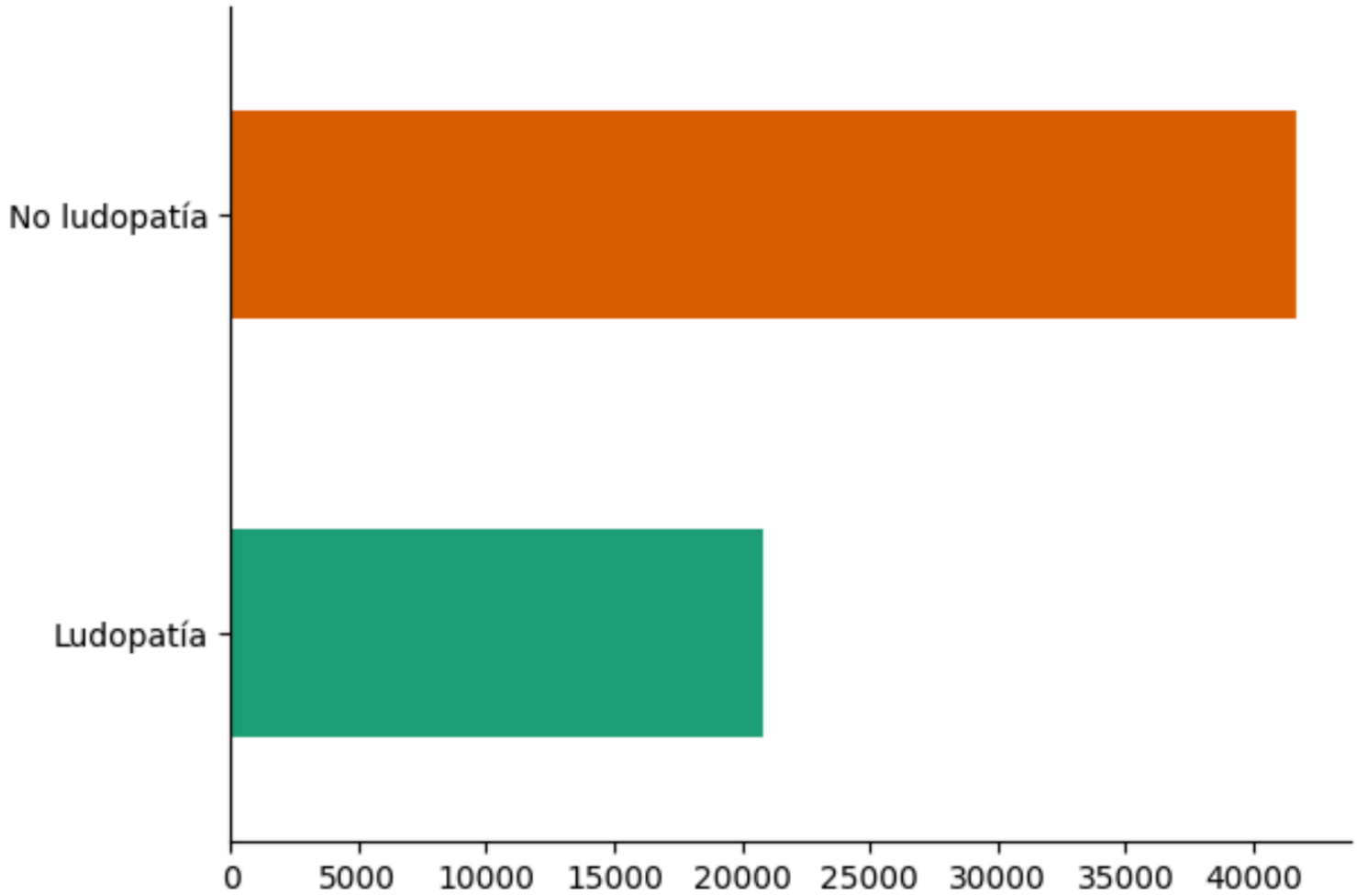

Figure: Nube de palabras de ludopatías

Ajuste Fino

Hiperparámetros

- MAX_LEN = 128
- TRAIN_BATCH_SIZE = 32
- VALID_BATCH_SIZE = 32
- EPOCHS = 10
- LEARNING_RATE = 1e-05
- MAX_GRAD_NORM = 10

Arquitectura

```
  (pre_classifier): Linear(in_features=768, out_features=768, bias=True)
  (classifier): Sequential(
    (0): Dropout(p=0.5, inplace=False)
    (1): Linear(in_features=768, out_features=256, bias=True)
    (2): Tanh()
    (3): Dropout(p=0.3, inplace=False)
    (4): Linear(in_features=256, out_features=32, bias=True)
    (5): Tanh()
    (6): Linear(in_features=32, out_features=16, bias=True)
    (7): Linear(in_features=16, out_features=2, bias=True)
    (8): LogSoftmax(dim=1)
  )
  (dropout): Sequential(
    (0): Dropout(p=0.2, inplace=False)
  )
)
```

Figure: Capas agregadas para mejorar la clasificación

Interpretación del modelo

Interpretación

- Para este proyecto buscaremos inspeccionar los cabezales de atención en búsqueda de posibles razones fundamentales para la clasificación, es decir buscaremos relaciones entre una entrada dada y la clasificación de un modelo.
- Buscaremos entender que lleva un modelo BERT, dada una muestra, llegue a una decisión.

Interpretación

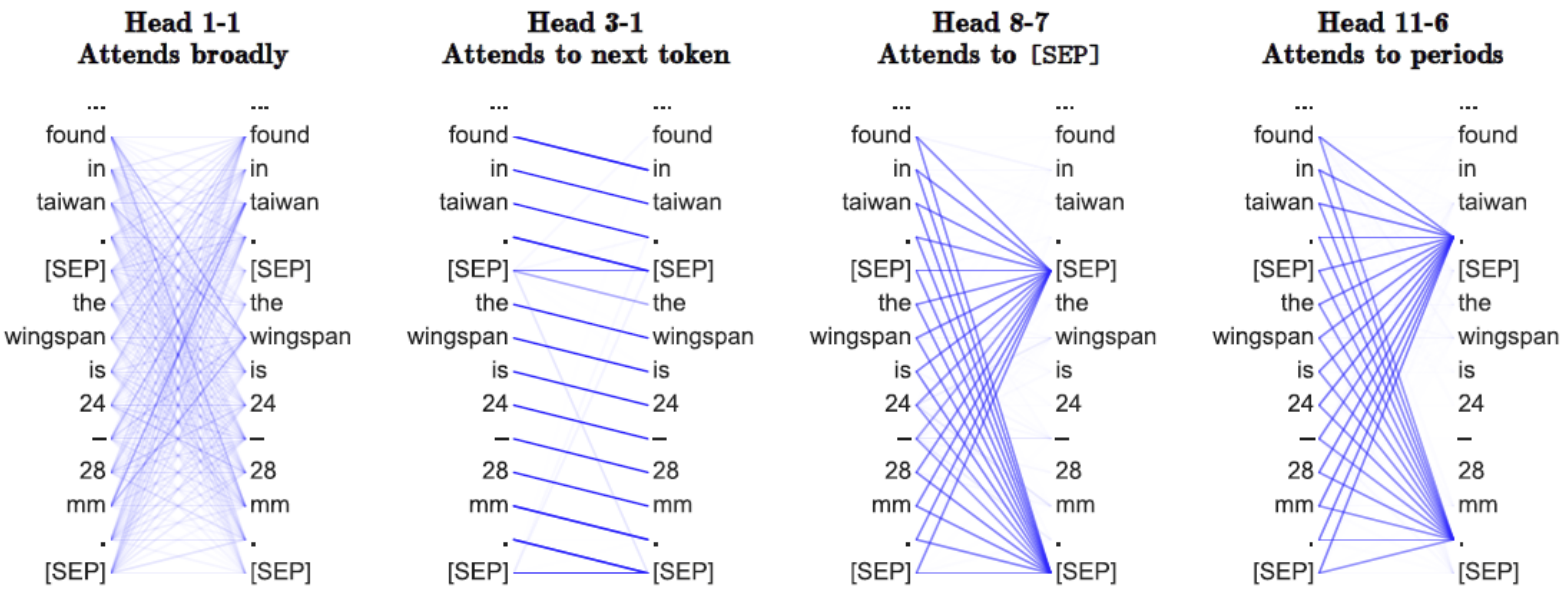

Figure 1: Examples of heads exhibiting the patterns discussed in Section 3. The darkness of a line indicates the strength of the attention weight (some attention weights are so low they are invisible).

Figure: Clark K. et. al.(2019) - *What Does BERT Look At? An Analysis of BERT' s Attention*

Gradientes Integrados

- Sundararajan et.al. (2017) introdujeron en el articulo 'Axiomatic Attribution for Deep Networks' un método de atribución que cumple con propiedades de importancia para la interpretación de modelos. Axiomas de Sensibilidad y de Invarianza de Implementación.
- Supongamos que tenemos una función $F : \mathbb{R}^n \to [0, 1]$ la cual representa una red profunda. Específicamente, donde sea $x \in \mathbb{R}^n$ sea una entrada y $x' \in \mathbb{R}^n$ sea la entrada nula. Para redes convolucionales, la entrada nula puede ser la de una imagen completamente negra, mientras que para textos podría ser del vector de embedding cero. Se considera un camino secuencial (en R^n) desde la entrada nula x' de la entrada x, y computamos los gradientes para todos los puntos del camino de funciones. Los gradientes integrados se obtienen al acumular estos gradientes, se definen como el camino de integrales de gradientes a lo largo del camino secuencial desde la base x' a la entrada x.

Gradientes Integrados

- El gradiente integrado a lo largo de la i-ésima dimensión para una entrada x y la entrada nula x' se define a continuación.

$$\text{IntegratedGrads}_i(x) ::= (x_i - x_i') \times \int_{\alpha=0}^{1} \frac{\partial F(x' + \alpha(x - x'))}{\partial x_i} d\alpha$$

- donde, $\frac{\partial F(x)}{\partial x_i}$ es el gradiente de $F(x)$ a lo largo de la dimensión i.
- Este método de atribución cumple con dos axiomas:

Gradientes Integrados

- **Sensibilidad**

  Un método de atribución se dice que satisface con este axioma si para cada valor y su valor nulo difieren cada una característica, pero tiene diferentes predicciones de manera que la característica debería de tener una atribución diferente de cero. En el caso contrario, si una red profunda no depende de alguna variable, esa atribución debe de ser de cero. Para ejemplificar, consideremos una variable en una red ReLU, $f(x) = 1 - ReLU(1 - x)$. Supongamos que su valor nulo es de $x' = 0$ y de entrada tenemos que $x = 2$. La función cambia de 0 a 1, pero dado que f tiene gradiente de 0 en a partir del valor de $x = 1$ el método de gradiente da una atribución de 0 a $x = 2$.

- **Invarianza de Implementación**

  Se dicen que dos redes son *funcionalmente equivalentes* si para todas sus salidas son iguales para todas las entradas, sin importar la implementación.

Resultados

Métricas

	precision	recall	f1-score	support
Ludopatía	0.76	0.55	0.64	5222
No ludopatía	0.65	0.82	0.73	5197
accuracy			0.69	10419
macro avg	0.70	0.69	0.68	10419
weighted avg	0.70	0.69	0.68	10419

Figure: Métricas obtenidas

Auto atención ludopatía

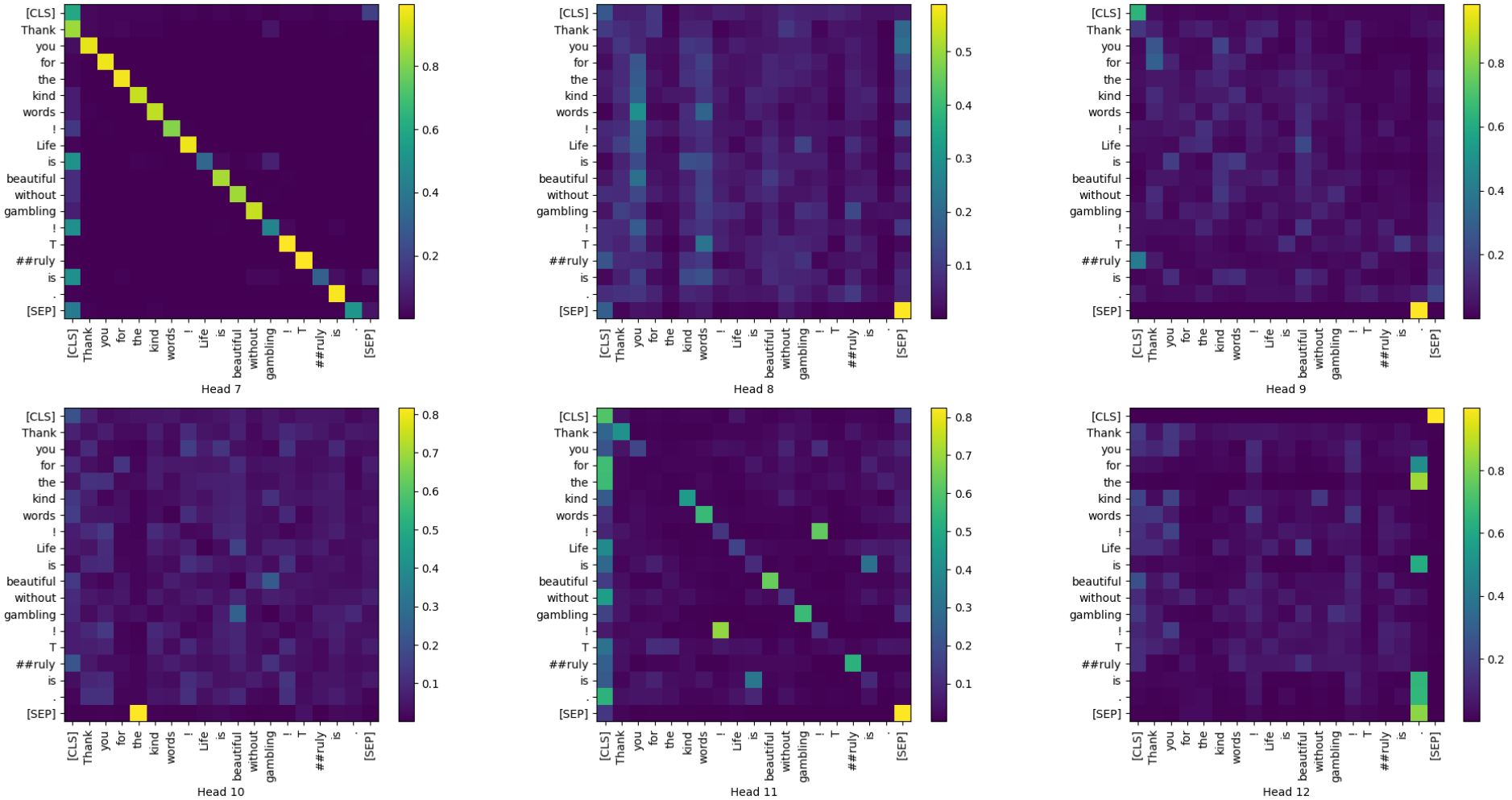

Figure: Heatmap comentarios de personas diagnosticadas

Auto atención no ludopatía

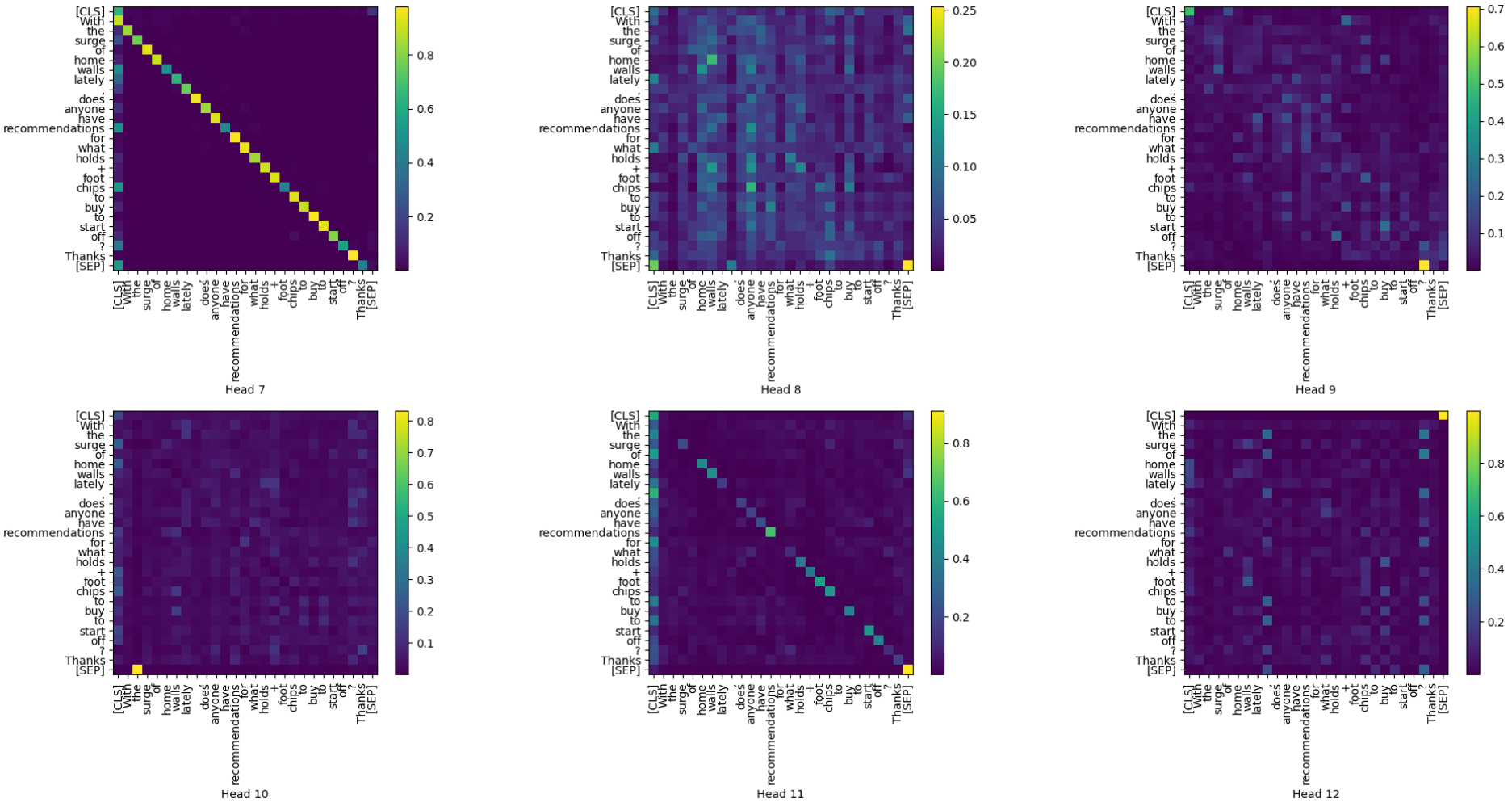

Figure: Heatmap comentarios de personas no diagnosticadas

BertViz

Ver notebook

Importancia de características ludopatía

Legend: Negative Neutral Positive

True Label	Predicted Label	Attribution Label	Attribution Score	Word Importance
Ludopatía	Ludopatía (-0.35)	Give up your money man. Find someone, make a friend if you have to, and give them the lock to that money to safeguard it from your self. This is the only thing that worked for me. Guess what, I still have the urge to gamble, but I also don't have the urge to jump off a cliff every other week. I gave up just a bit of my pride to get back everything.	-5.33	[CLS] Give up your money man . Find someone , make a friend if you have to , and give them the lock to that money to safe ##guard it from your self . This is the only thing that worked for me . Guess what , I still have the urge to g ##amble , but I also don ' t have the urge to jump off a cliff every other week . I gave up just a bit of my pride to get back everything . [SEP]

Legend: Negative Neutral Positive

True Label	Predicted Label	Attribution Label	Attribution Score	Word Importance
Ludopatía	Ludopatía (-0.41)	Lol white American mutt. Just asking because I saw the name and I know a lot of Chinese have gambling issues. Look man, your family is definitely dissapointed, I'm sure you are as well, but they mean well and are doing that to correct you. You know what my family did? They told me it's no big deal and that maybe I should try again, but this time be smarter. Trust me, Id rather have your family talking to me than mine lol, maybe I wouldn't be where I am today	-3.61	[CLS] Lo ##l white American m ##utt . Just asking because I saw the name and I know a lot of Chinese have gambling issues . Look man , your family is definitely di ##ssa ##point ##ed , I ' m sure you are as well , but they mean well and are doing that to correct you . You know what my family did ? They told me it ' s no big deal and that maybe I should try again , but this time be smart ##er . Trust me , I ##d rather have your family talking to me than mine lo ##l , maybe I wouldn ' t be where I am today [SEP]

Figure: Palabras relevantes basadas en información del gradiente para personas diagnosticadas

Importancia de características no ludopatía

Legend: Negative Neutral Positive

True Label	Predicted Label	Attribution Label	Attribution Score	Word Importance
No ludopatía	Ludopatía (-0.51)	Hey it was a mistake. Can happen to anyone. While I understand it is so tough to get over it and forgive yourself, please dont beat yourself over it. I know the feeling of loosing a dog. Please hold it together for your daughter. Maybe you can raise awareness so people dont make the same mistake again. Im sorry what has happened to you. Im sure the dog is in a good place and has forgiven you.	-4.96	[CLS] Hey it was a mistake . Can happen to anyone . While I understand it is so tough to get over it and forgive yourself , please don ##t beat yourself over it . I know the feeling of lo ##os ##ing a dog . Please hold it together for your daughter . Maybe you can raise awareness so people don ##t make the same mistake again . I ##m sorry what has happened to you . I ##m sure the dog is in a good place and has forgive ##n you . [SEP]

Legend: Negative Neutral Positive

True Label	Predicted Label	Attribution Label	Attribution Score	Word Importance
No ludopatía	No ludopatía (-0.60)	With the surge of home walls lately, does anyone have recommendations for what holds + foot chips to buy to start off? Thanks	-2.16	[CLS] With the surge of home walls lately , does anyone have recommendations for what holds + foot chips to buy to start off ? Thanks [SEP]

Figure: Palabras relevantes basadas en información del gradiente para personas no diagnosticadas

Conclusiones

Conclusiones

- Se mostró que la autoatención no puede ser tomada como un criterio de importancia. De manera general el token [CLS] no parece relacionarse con palabras especificas, sino que parece sesgarse hacia el token [SEP].
- Por otra parte, medidas de importancia como la de gradientes integrados, detectan palabras relacionadas con los tópicos y nubes de palabras de personas diagnosticadas, demostrando la relación limitada entre estas dos medidas.

¡Gracias!