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

Sacbe García García, Carlos Cuauhtémoc Gutiérez Salazar

Centro de Investigación en Matemáticas, A.C.
Unidad Monterrey
Monterrey, Nuevo León
Deep Learning.

12 de diciembre de 2023

Introducción

Los mecanismos de atención permiten que los modelos se centren en diferentes partes de la secuencia de entrada al producir la entrada de cada modulo consecuente.

A menudo se afirma que los pesos de atención, ya sea implícita o explícitamente, ofrecen información sobre el funcionamiento interno de los modelos. Si bien los pesos de atención pueden proporcionar una interpretabilidad valiosa, su interpretación puede ser subjetiva o depender del modelo y tareas específicas.

Al inspeccionar entradas con pesos de atención grandes, las entradas a las que el modelo asignó grandes pesos de atención, los analistas o investigadores pretenden comprender qué elementos de entrada fueron particularmente influyentes en la generación del resultado correspondiente.

Recientemente, ha habido un interés creciente en si estas representaciones intermedias ofrecidas por los módulos pueden usarse o no para explicar el razonamiento de la predicción de un modelo y, en consecuencia, alcanzar conocimientos sobre el proceso de toma de decisiones del modelo[6].

La atención sea o no una explicación depende de la definición de explicabilidad que uno esté buscando.

Lipton [8] clasifica la transparencia, o comprensión humana general de un modelo, y la explicabilidad post-hoc como dos nociones de interpretabilidad. El sentido relevante de transparencia, tal como lo define Lipton, se refiere a la forma en que una porción específica de un modelo corresponde a una construcción comprensible por humanos.

Según esta definición, debería parecer sensato para cuestiones de NLP tratar las puntuaciones de atención como un vehículo de transparencia.

Los mecanismos de atención proporcionan una mirada al funcionamiento interno de un modelo, ya que producen una ponderación fácilmente comprensible de los estados ocultos.

Rudin [9] define la explicabilidad como simplemente una reconstrucción plausible (pero no necesariamente fiel) del proceso de toma de decisiones, y Riedl clasifica los fundamentos explicables como valiosos en el sentido de que imitan lo que nosotros como humanos hacemos cuando racionalizamos acciones pasadas: crear una historia que justifique de manera plausible nuestras acciones, aunque no sea una reconstrucción completamente precisa de los pasos que llevaron al comportamiento en un momento dado.

Al distinguir entre interpretabilidad y explicabilidad como dos nociones separadas, Rudin [9] sostiene que la interpretabilidad es más deseable pero más difícil de lograr que la explicabilidad, porque requiere presentar a los humanos una comprensión general de la relación correlación entre entradas y salidas.

En trabajos anteriores, Lei et.al.[7] entrenan un modelo para generar simultáneamente fundamentos y predicciones a partir del texto de entrada, utilizando fundamentos de etiqueta dorada para evaluar su modelo. Generalmente, muchos aceptan la noción de métodos extractivos como Lei et.al.[7], en el que las explicaciones provienen de funciones aplicadas a una entrada (como en la atención), como plausible. Trabajos como Mullenbach et al. (2018) y Ehsan et. al. (2019) utilizan la evaluación humana para evaluar explicaciones; el primero se basa en puntuaciones de atención sobre la entrada, y el segundo se basa en sistemas con capacidad funcional de generación de fundamentos. Los autores muestran que los fundamentos generados de manera post-hoc aumentan la confianza del usuario en un sistema. Citando a Ross et. al. (2017), el requisito de Jain y Wallace[6] para que las distribuciones de atención se utilicen como explicación es que solo debe existir una o unas pocas explicaciones correctas estrechamente relacionadas para la predicción de un modelo. Sin embargo, Doshi-Velez y Kim (2017) advierten contra la aplicación de evaluaciones y terminología de manera amplia sin aclarar las necesidades de explicación específicas de la tarea. Si aceptamos que las definiciones de explicabilidad de Rudin y Riedl proporcionan una justificación plausible, pero no necesariamente fiel, para la predicción del modelo, entonces el argumento en contra

Sundararajan et.al.[13] analizaron el problema de atribuir la predicción de una red profunda a las características de entrada e introdujeron un método de atribución de predicción llamado 'gradientes integrados' el cual cumple con 2 axiomas introducidos en su articulo *'Axiomatic Attribution for Deep Networks'*. Gracias a estos axiomas, y el cumplimiento de ellos por los gradientes integrados, podemos tener certeza en el análisis de atribución de características realizado por este método, el cual intuitivamente atribuye las decisiones correctamente a aquellas características más relevantes para una predicción dada.

¿Qué es el juego compulsivo?

El juego compulsivo, trastorno del juego o ludopatía, es un trastorno de conducta problemática asociada con la depresión, abuso de sustancias, violencia doméstica, estado financiero de quiebra y altas tasas de suicidio. En la quinta edición del Manual diagnóstico y estadístico de los trastornos mentales (DSM-V, por sus siglas en ingles) se categorizo dentro de *Desordenes Relacionados a Substancias y Adicciones* para reconocer que las investigaciones sugieren que el juego patológico con el alcohol y la adicción a las drogas están relacionadas.

El juego compulsivo tiene una presencia cada vez mayor en las redes sociales, tanto en términos de contenido promocional para usuarios, lugares y actividades de juegos de azar, tanto como en forma de entretenimiento interactivo-social.

Las tendencias y problemáticas relacionadas al juego compulsivo se pueden desglosar de la siguiente manera:

- **Aumento en la Accesibilidad**
  Las redes sociales han permitido a la practica de apuestas ser más accesibles, cada vez con ventanas menos restrictivas para públicos susceptibles.

- **Publicidad Personalizada**
  Los usuarios de sitios de apuestas son objetivos de anuncios en redes sociales, lo que ha llevado a un incremento en la exposición.

- **Apuestas en Juegos**
  Algunas plataformas de redes sociales permiten apuestas en juego o apuestas en tiempo real, creando un entorno donde los usuarios pueden participar de forma rápida e impulsiva.

- **Juegos de casino sociales**
  Los juegos de casino sociales en plataformas como Facebook simulan experiencias de juego sin apuestas con dinero real. Si bien no son intrínsecamente problemáticos, pueden contribuir a normalizar las conductas de juego.

Se han implementado regulaciones para los anuncios de juegos de casino y medidas de seguridad en redes sociales, que prevén acceso a usuarios menores, pero la preocupación latente de los avances tecnológicos se mantiene presente en los foros de investigación.

Metodología

Limpieza de los Datos y Análisis Exploratorio

Los datos fueron parte de una competencia eRisk del ano 2022 por parte de la asociación CLEF. El reto consistió en realizar una tarea de detección temprana de la adicción por las apuestas patológica.

El desafío esta en procesar secuencialmente evidencias y detectar lo antes posible rastros tempranos de adicción patológica, también conocido como juego compulsivo o juego desordenado. La tarea se ocupa principalmente de evaluar soluciones de Text Mining y, por tanto, se concentra en textos escritos en redes sociales.

Los textos deben procesarse en el orden en que fueron creados. De esta manera, se podrían aplicar sistemas que realicen eficazmente esta tarea para monitorear secuencialmente las interacciones de los usuarios en blogs, redes sociales u otros tipos de medios en línea.

Para el análisis exploratorio, se visualizaron los datos diferentes maneras

- Diagrama de linea de palabras más frecuentes: Figura 1.
- Nube de palabras de palabras todos los textos: Figura 2. Nubes de palabras para aquellos textos que fueron clasificados de forma positiva y negativa: Figura 5 Y Figura 4 respectivamente.
- Histograma de las longitud de textos: Figura 2.
- Histograma de cantidades de tokens en cada muestra: Figura 3.

La distribución de las palabras más frecuentes es la siguiente:

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

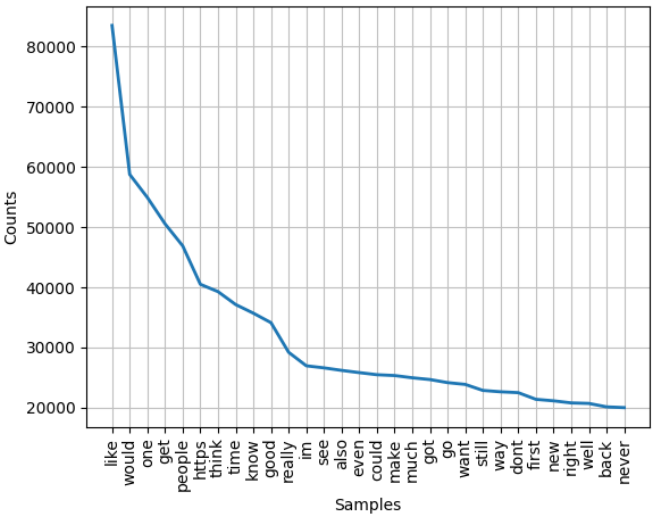

Figura 1. Frecuencia de palabras en los comentarios

Figura 2. Nube de palabras del texto general

Podemos observar como los datos son diversos en contenido y volumen.

BERT y DistilBert

La red pre-entrenada BERT (*Bidirectional Encoder Representations from Transformers* [3]) es un transformer que ha tenido un enorme impacto en NLP y es el estado-del-arte en múltiples tareas. Este se compone de una secuencia de bloques de encoders.

El modelo ha sido entrenado con dos tarea especificas

- Masked Language Modeling (MLM). Esta parte del entrenamiento consiste en esconder un porcentaje de los tokens,

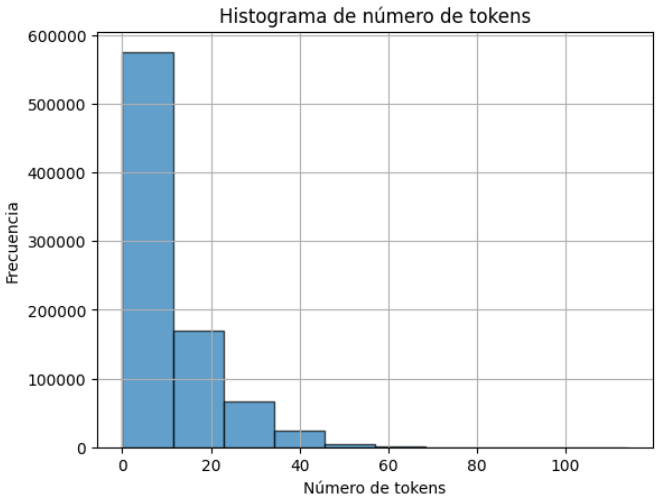

Figura 3. Histograma representante del numero de tokens presentes

y el modelo entrenado predice esas palabras escondidas.

- Next Sentence Prediction (NSP) El modelo recibe tokens pareados como entrada y aprende a predecir si el segundo token va después de la primero.

El resultado es un modelo que al ser entrenado para alguna tarea especifica (fine-tuning) logra obtener un desempeño impresionante, desde análisis de sentimientos hasta tareas de pregunta-respuesta.

Cada bloque encoder se compone de una capa de 'Autoatención' y una capa densamente conectada, además de una capa normalizadora después de cada una.

La secuencia de entrada se envía a la función embedding de entrada y de codificación de posición, produciendo una representación para cada palabra en la secuencia de entrada que captura el significado y la posición de cada palabra. Esto se envía a los tres parámetros: Query, Key y Value (denominados Q, E y V respectivamente), en la capa de Autoatención se produce una representación codificada para cada palabra en la secuencia de entrada, que ahora también incorpora las puntuaciones de atención para cada palabra. A medida que esto pasa por todos los codificadores del modelo, cada módulo de Autoatención también agrega sus propios puntajes de atención a la representación de cada palabra.

BERT es usado para todo relacionado a lenguaje natural, desde tareas de análisis de sentimientos hasta Q&A. La demanda de modelos BERT más compactos aumenta para poder utilizar BERT en entornos computacionales más limitados, como teléfonos móviles y computadoras personales. En marzo de 2020 se lanzaron 23 modelos BERT más pequeños. **DistilBert** es entrenado por destilación por parte de un modelo BERT, lo que significa que es entrenado para predecir las mismas probabilidades. DistilBERT ofrece una versión más ligera de BERT que funciona un 60 % más rápido y mantiene más del 95 % del rendimiento de BERT[10], a través de una técnica de aprendizaje llamada 'Destilación de conocimiento'.

Destilación de conocimiento

La Destilación de conocimiento (Bucila et.al., 2006 [1], Hinton etl.al., 2015 [4]) es una técnica de compresión en la que un modelo (o ensamble de modelos) entrena a otro, logrando reproducir el conocimiento del primero. Esta técnica es usada para el pre-entrenamiento del modelo DistilBert y es una técnica usada para la construcción de modelos de tareas especializadas.

Durante un entrenamiento supervisado de un modelo de clasificación, se busca maximizar la probabilidad estimada de las etiquetas verdaderas, al minimizar alguna función de perdida. i.e. Cross-Entropy sobre una capa *softmax*.

Un modelo que funcione bien en el conjunto de entrenamiento predecirá una distribución de salida con alta probabilidad en la clase correcta y con probabilidades cercanas a cero en otras clases. Pero algunas de estas probabilidades çercanas a cero"son mayores que otras y reflejan, en parte, las capacidades de generalización del modelo y qué tan bien funcionará en el conjunto de prueba.

El modelo 'alumno' aprende con una función de perdida durante la destilación de conocimiento sobre las probabilidades del 'profesor': $L_{ce} = \sum_i t_i \times log(s_i)$, donde t_i (resp. s_i) son las probabilidades estimadas por el profesor, donde se usa una capa '*softmax-temperature*': $p_i = \frac{exp(z_i/T)}{\sum_j exp(z_i/T)}$, donde T controla la suavidad de la distribución de salida y z_i es el puntaje de la clase i. Se aplica la misma temperatura T al estudiante y profesor durante el entrenamiento, mientras que al inferir se establece como $T =$, tomando la forma usual de *softmax*.

Adicionalmente, se usa una combinación lineal de la función de perdida L_{ce} con la función de perdida del entrenamiento supervisado, *masked language modeling loss* L_{mlm} y la función de perdida *cosine embedding loss* L_{cos}, lo que ayuda a alinear las direcciones de estados ocultos entre el estudiante y profesor[5].

Wordpiece

Antes de introducir los tokens al modelo, estos se procesan por un Word embedding. Esto mejora el desempeño de la clasificación, así como capturar conocimiento semántico particular del corpus particular de la tarea, entre muchos otros beneficios. Los algoritmos Word embeddings buscan principalmente generar una representación numérica que mantenga importancia semántica, es decir capture contexto y/o significado de las palabras.

Wordpiece[12] es un algoritmo de segmentación de palabras utilizado frecuentemente en el campo de procesamiento de lenguaje natural. Esta técnica de tokenización fue descrita en el articulo 'Búsqueda de voz japonesa y coreana', por Schuster[11]. Wordpiece comienza con un vocabulario limitado con tokens especiales para iniciar el alfabeto, para después agregar iterativamente las combinaciones de caracteres más frecuentes.

Parecido a BPE, se definen nuevos tokens al añadir prefijos '##', a todos los caracteres dentro de una palabra. Por ejemplo, "word"se divide de la siguiente manera

```
w  ##o  ##r  ##d
```

De manera que el alfabeto inicial contiene a todos los caracteres iniciales, y todos los otros caracteres de las palabras predichos por el prefijo

Wordpiece aprende reglas para converger las palabras. La diferencia principal con BPE, es la manera en la que selecciona la convergencia. En lugar de seleccionar los pares más frecuente, Wordpiece computa un puntaje para cada par, usando la

Figura 4. Nube de palabras de comentarios de personas no diagnosticadas.

Figura 5. Nube de palabras de comentarios de personas diagnosticadas. Podemos observar alto uso de palabras relacionadas al uso del dinero y del tiempo, como los son 'will', 'money', 'time'.

siguiente formula.

$$\text{score} = \frac{\text{Freq. of pair}}{\text{Freq. of first element} \times \text{Freq. of second element}}$$

A diferencia de BPE, WordPiece no elige el par de símbolos más frecuente, sino el que maximiza la probabilidad de que los datos de entrenamiento se agreguen al vocabulario.

Se ha encontrado diversos beneficios en el uso de Wordpiece en el campo de procesamiento de lenguaje natural, en forma de

- Manejo de palabras fuera-del-vocabulario (OOV)
  Al dividir las palabras en tokens de subpalabras, WordPiece mejora el manejo de palabras OOV que no están presentes en el vocabulario. Esto es especialmente útil para manejar términos raros o especializados.
- Codificación de texto eficiente
  La tokenización de subpalabras reduce el tamaño del vocabulario en comparación con la tokenización a nivel de palabra, lo que resulta en una codificación más eficiente de los datos de texto.
- Agnóstico del idioma
  WordPiece es independiente del idioma y se puede aplicar a varios idiomas, lo que lo hace versátil para aplicaciones multilingües.
- Adaptabilidad a los requisitos de la tarea
  La granularidad de los tokens de subpalabras permite que los modelos capturen variaciones morfológicas, manejen términos fuera del dominio y generalicen mejor a palabras nuevas.

Este algoritmo es usado frecuentemente para el pre-entrenamiento de BERT, DistilBERT y Electra. Particularmente el implementado en la distribución de HuggingFace `DistilBertTokenizerFast` que fue utilizada para la experimentación.

Fine-Tuning

Utilizar un modelo pre-entrenado reduce costos computacionales al reducir el computo necesario para abordar tareas especificas.

El modelo DistilBert entrenado tiene 6 bloques encoder, con 12 cabezales de atención, seguidas de 2 capas densamente conectadas las cuales fueron entrenadas para la tarea de clasificación

Para la experimentacion see uso Python 3.10.12, `PyTorch v2.1.0`, `transformers v4.36`, en la plataforma de Google Colab Pro.

Y se entreno la distribución de DistilBert de por 10 épocas usando minibatches de 1303 muestras usando una función de costo de logaritmo de verosimilitud negativo con un optimizador de Adam con tasa de aprendizaje inicial de $lr = 0{,}0001$.

Gradientes Integrados

[13]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.

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:

- **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.

Métricas de evaluación

Al realizar el fine-tuning del modelo sobre el conjunto de entrenamiento por 300 épocas, se obtuvieron las siguientes métricas de rendimiento:

	precision	recall	f1-score	support
Ludopatía	0.79	0.74	0.76	5222
No ludopatía	0.75	0.80	0.77	5197
accuracy			0.77	10419
macro avg	0.77	0.77	0.77	10419
weighted avg	0.77	0.77	0.77	10419

El desempeño y tamaño de los datos es comparable a los usados por Jain y Wallace en sus experimentos, siendo una diferencia enorme el modelo RNN implementado. Donde ellos realizan análisis sobre un modelo LSTM. Creemos que este modelo sera suficientemente certero para realizar interpretación sobre el mismo.

Resultados y Discusión

Resultados generales

Al observar las ponderaciones de atención, concordamos que muchas cabezas prestan atención desbalanceada hacia el token [SEP] y otras se especializan en relaciones semánticas en acorde a lo hallado por Clark[2]. Por otra parte muchas otras ponen peso de manera arbitraria al siguiente o previo token en la entrada. Por otra parte la tecnica de gradientes integrados provee transparencia a un nivel de entrada, lo que nos permite ver dentro de la caja negra del modelo DistilBert.

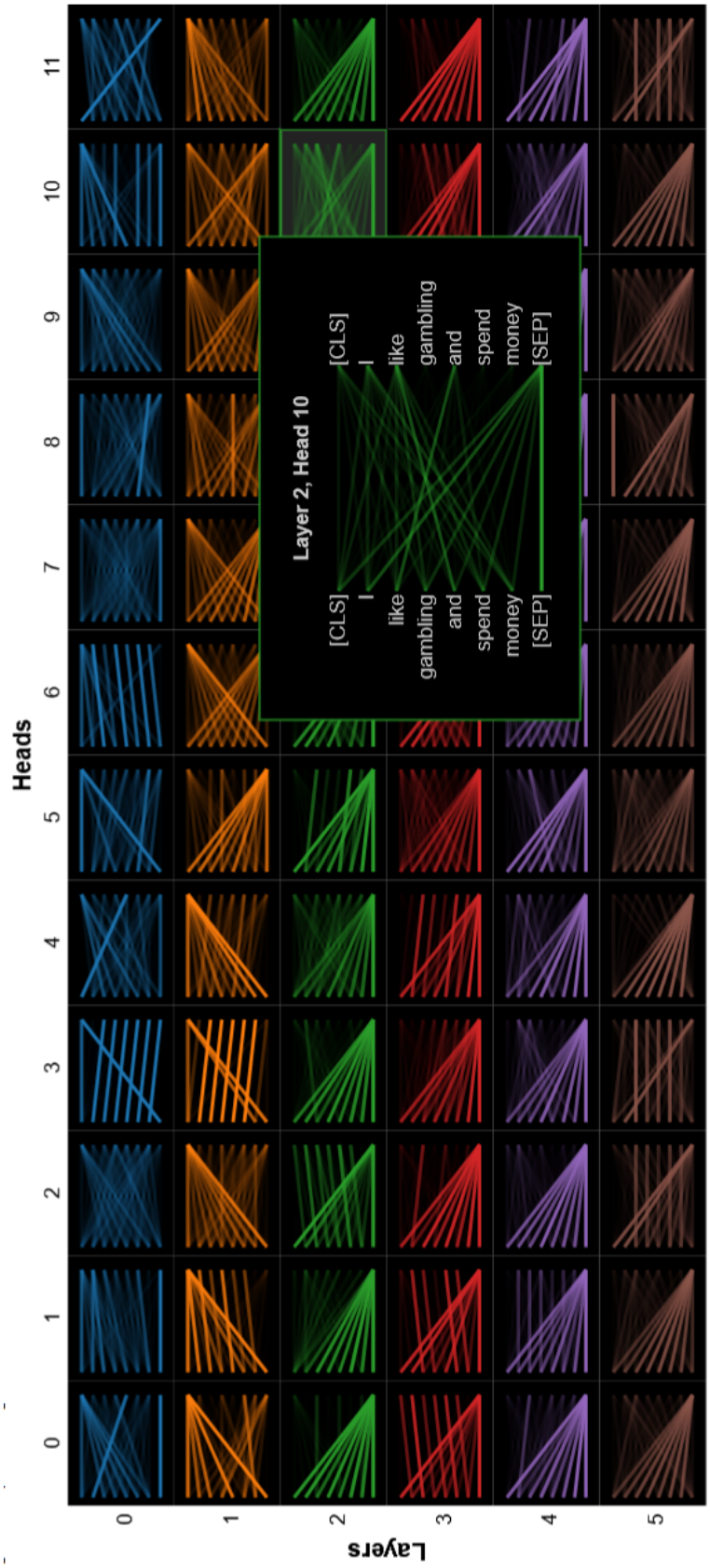

Figura 6. Atención entre los distintos tokens por capas y cabezales

Conclusión

Los mecanismos de atención porque no son exclusivos, como afirman Jain y Wallace[6], no son válidos, y la evaluación humana es necesaria para evaluar la plausibilidad de los fundamentos generados al realizar un análisis sobre la interpretación de un modelo profundo.

El hecho de que exista otra explicación no significa que la proporcionada sea falsa o carezca de sentido y la existencia de múltiples explicaciones diferentes no es necesariamente indicativa de la calidad de una sola.

Jain y Wallace definen la atención y la explicación como medir la responsabilidad"que tiene cada token de entrada en una predicción. Esto se alinea más estrechamente con la definición más rigurosa de transparencia de Lipton[8], o la definición de interpretabilidad de Rudin[9]: comprensión humana del modelo como un todo en lugar de sus partes respectivas. La pregunta objetivo planteada inicialmente de '¿Es la atención es una explicación?' parece tener de respuesta otra cuestión diferente: '¿Son las ponderaciones de atención sobre ciertos elementos en la entrada suficientes para llevar a un modelo a hacer su predicción?'

Según la definición dada de transparencia, el requisito de exclusividad de Jain y Wallace está bien definido y encontramos valor en su marco contrafactual como concepto: si un modelo es capaz de producir múltiples conjuntos de pesos de atención para la misma predicción, entonces la relación entre las entradas y las salidas no es comprendida por el análisis de la atención.

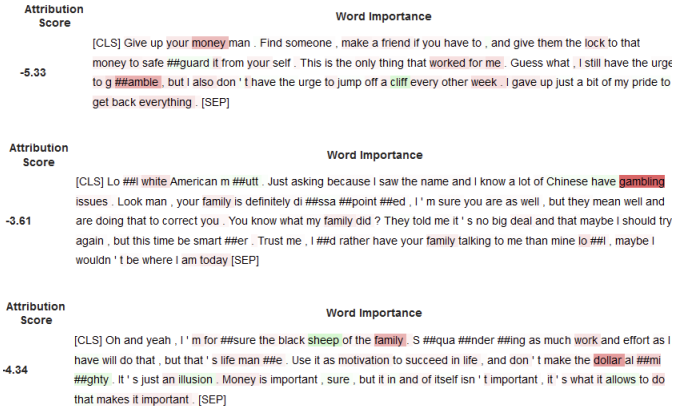

Figura 7. Importancia de las características del texto usando gradientes en personas diagnosticadas

Referencias

[1] Cristian Bucila, Rich Caruana y Alexandru Niculescu-Mizil. "Model compression". En: vol. 2006. Ago. de 2006, págs. 535-541. DOI: 10.1145/1150402.1150464.

[2] Kevin Clark et al. *What Does BERT Look At? An Analysis of BERT's Attention*. 2019. arXiv: 1906.04341 [cs.CL].

[3] Jacob Devlin et al. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. 2019. arXiv: 1810.04805 [cs.CL].

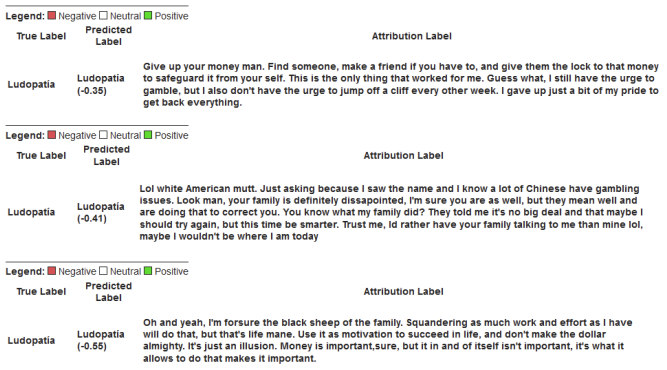

Figura 8. Importancia de las características del texto usando gradientes en personas diagnosticadas

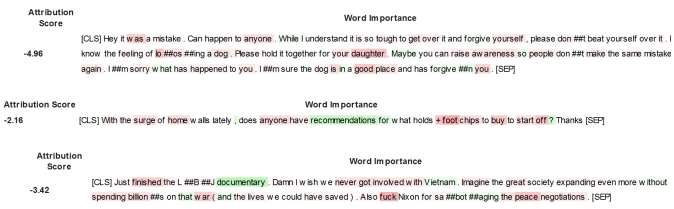

Figura 9. Importancia de las características del texto usando gradientes en personas no diagnosticadas

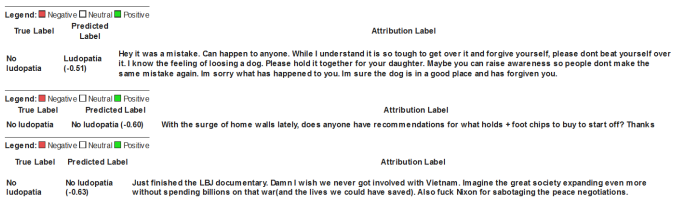

Figura 10. Importancia de las características del texto usando gradientes en personas no diagnosticadas

Figura 11. Tópicos encontrados en individuos diagnosticados

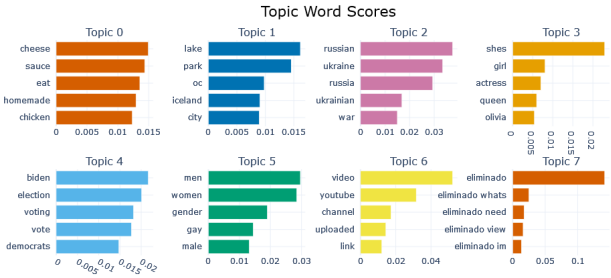

Figura 12. Tópicos encontrados en individuos no diagnosticados

[4] Geoffrey Hinton, Oriol Vinyals y Jeff Dean. *Distilling the Knowledge in a Neural Network*. 2015. arXiv: 1503.02531 [stat.ML].

[5] Geoffrey Hinton, Oriol Vinyals y Jeff Dean. *Distilling the Knowledge in a Neural Network*. 2015. arXiv: 1503.02531 [stat.ML].

[6] Sarthak Jain y Byron C. Wallace. *Attention is not Explanation*. 2019. arXiv: 1902.10186 [cs.CL].

[7] Tao Lei, Regina Barzilay y Tommi Jaakkola. "Rationalizing Neural Predictions". En: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Ed. por Jian Su, Kevin Duh y Xavier Carreras. Austin, Texas: Association for Computational Linguistics, nov. de 2016, págs. 107-117. DOI: 10.18653/v1/D16-1011. URL: https://aclanthology.org/D16-1011.

[8] Zachary C. Lipton. *The Mythos of Model Interpretability*. 2017. arXiv: 1606.03490 [cs.LG].

[9] Cynthia Rudin. *Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead*. 2019. arXiv: 1811.10154 [stat.ML].

[10] Victor Sanh et al. *DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter*. 2020. arXiv: 1910.01108 [cs.CL].

[11] Mike Schuster y Kaisuke Nakajima. "Japanese and Korean Voice Search". En: *International Conference on Acoustics, Speech and Signal Processing*. 2012, págs. 5149-5152.

[12] Mike Schuster y Kaisuke Nakajima. "Japanese and Korean voice search". En: mar. de 2012, págs. 5149-5152. ISBN: 978-1-4673-0045-2. DOI: 10.1109/ICASSP.2012.6289079.

[13] Mukund Sundararajan, Ankur Taly y Qiqi Yan. *Axiomatic Attribution for Deep Networks*. 2017. arXiv: 1703.01365 [cs.LG].

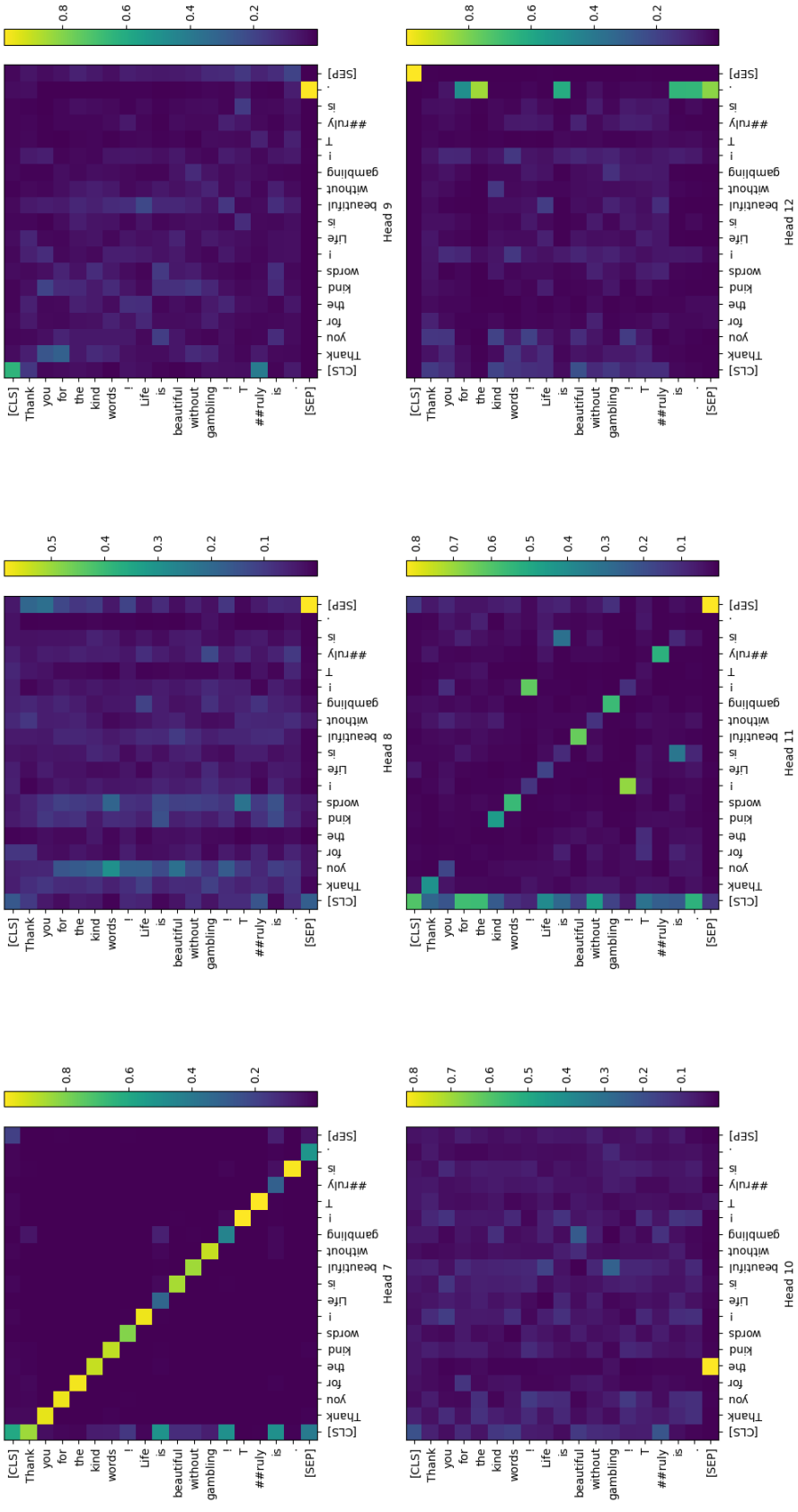

Figura 13. Hola. Hey

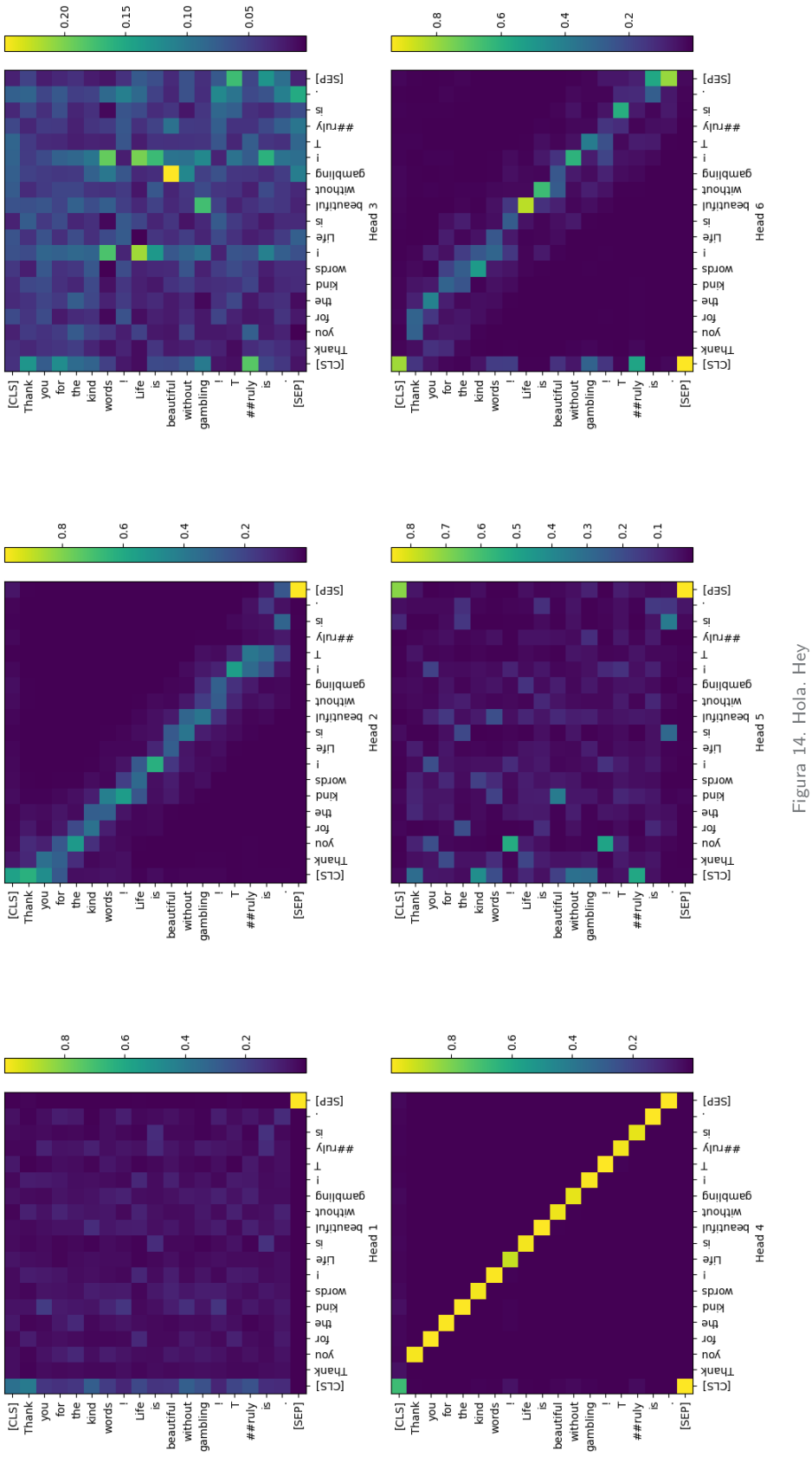

Figura 14. Hola. Hey

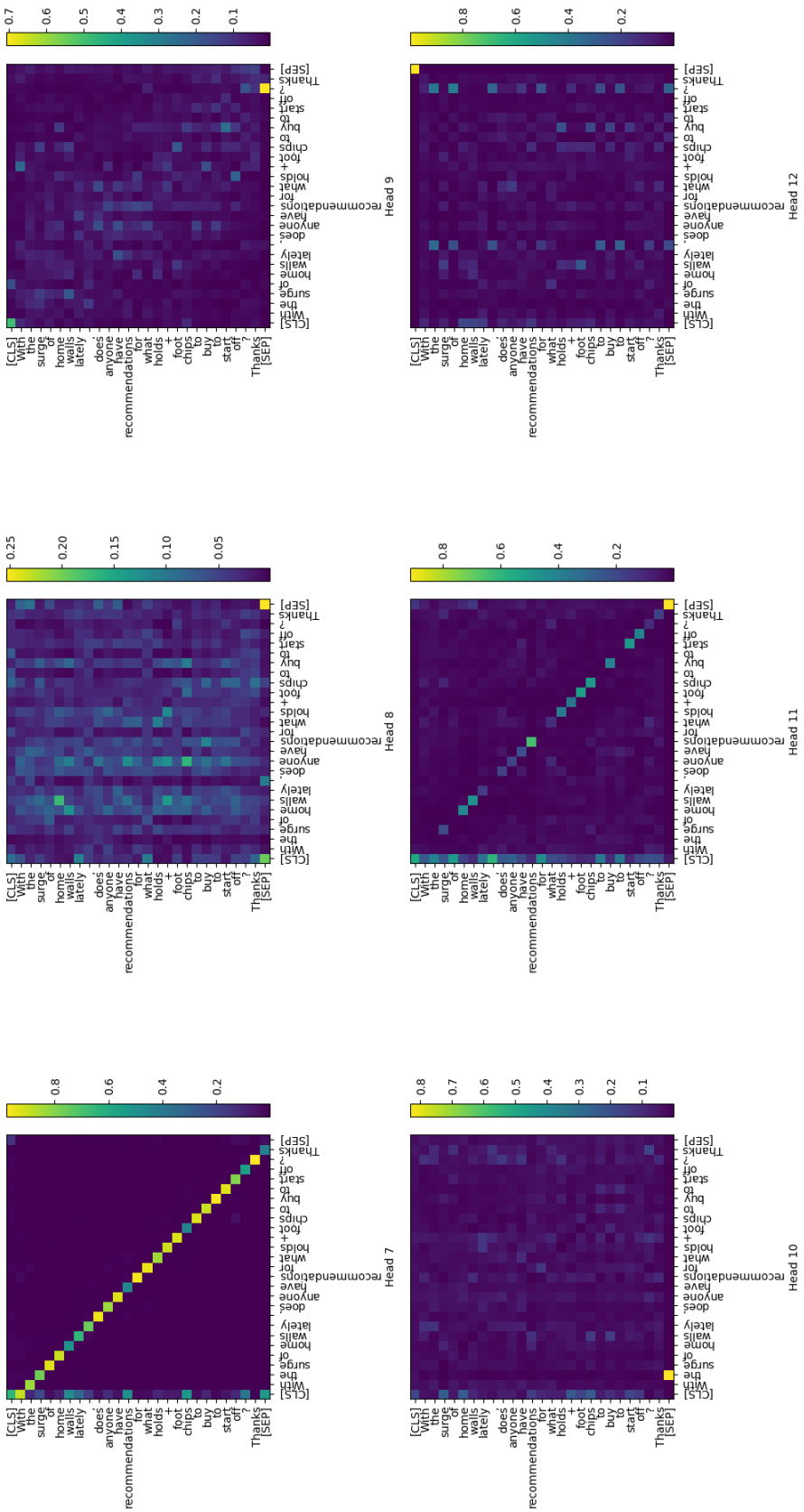

Figura 15. Hola. Hey

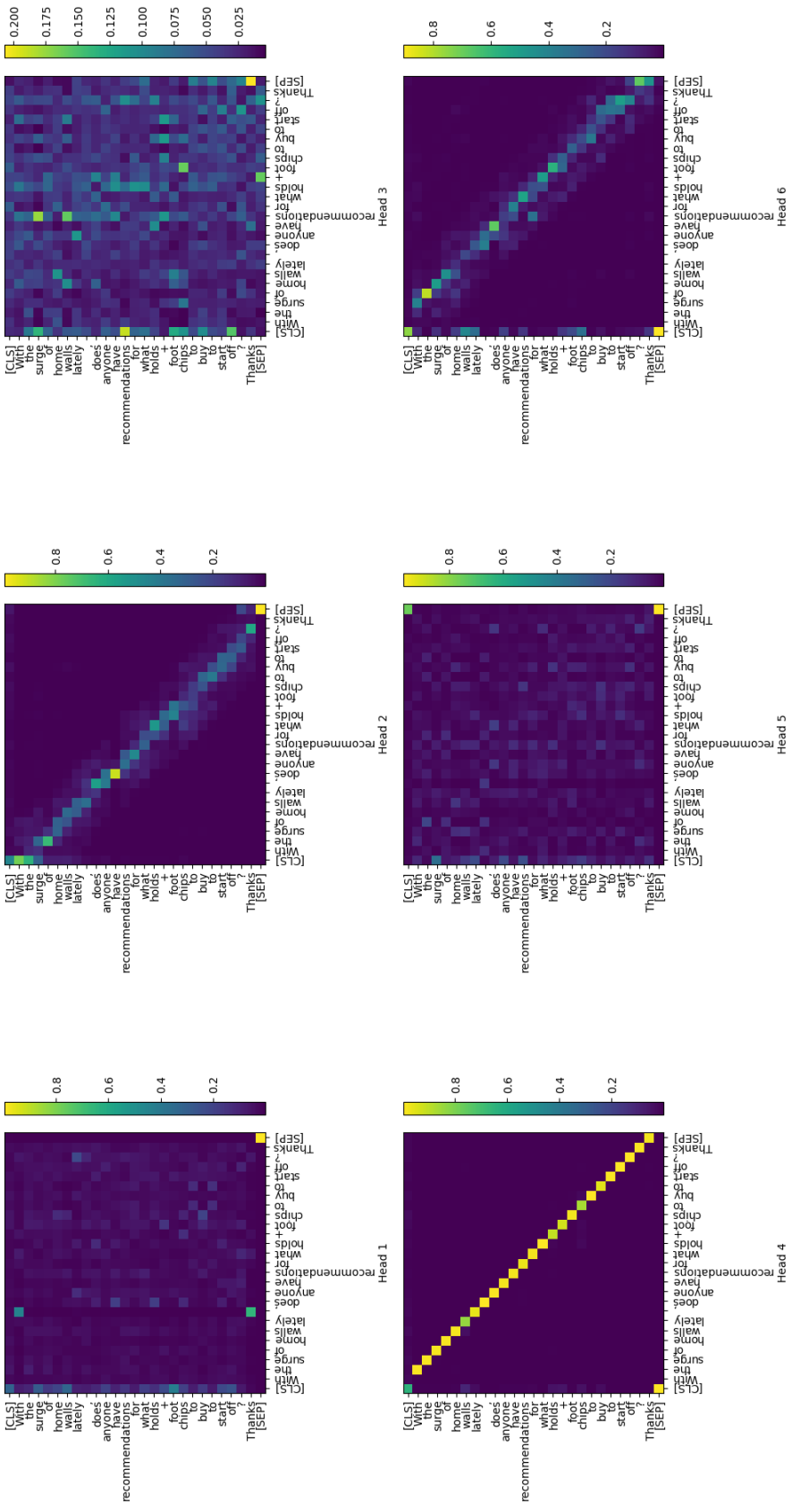

Figura 16. Hola. Hey