ISSN 1870-4069

Análisis de las características más importantes para la detección de noticias falsas en español

Sergio Damián, Hiram Calvo, Alexander Gelbukh

Instituto Politécnico Nacional, Centro de Investigación en Computación, México

{sdamians2019, hcalvo}@cic.ipn.mx, gelbukh@gelbukh.com

Resumen. El uso de modelos de lenguaje basados en Transformers para tareas del procesamiento del lenguaje natural (PLN) es cada día más común gracias a su potencial, su generalidad y gracias a que sus resultados superan de manera notoria otro tipo de estrategias y modelos. El principal problema de éstos es su gran número de parámetros y la complejidad que esto representa para la explicabilidad del modelo durante el proceso de resolución de una tarea en específico. El presente trabajo busca la explicabilidad para el funcionamiento del modelo BETO en la tarea de detección de noticias falsas utilizando un método alternativo al uso de pesos de atención encontrados en los mecanismos de atención de los Transformers. Se observa que las categorías gramaticales como la adposición, los determinantes, los sustantivos y los nombres propios representan los tokens más importantes que el modelo analiza para obtener las estimaciones o resultados. A su vez, el trabajo expone que para el corpus analizado, el uso de mayúsculas y el ignorar el entrenamiento de la capa de embeddings del modelo, mejora la comprensión de la tarea, presentando un valor F1 de 0.8653.

Palabras clave: Noticias falsas, explicabilidad, BETO, transformers.

Feature Importance Analysis for Fake News Detection in Spanish

Abstract. Using Transformer-based language models to solve Natural Language Processing (NLP) tasks is more common due to their potential, generalization, and the results that can surpass any other types of strategies and models. The main issue when using them is their huge amount of parameters and the complexity when trying to provide explainability of the model behavior. The present work analyses an alternative method to search explainability of BETO model behavior for a Fake News Detection task. It is observed that some part-of-speech subsets are the most relevant features in this case of study, like adposition, determinant, noun and proper noun labels. Besides, this work demonstrates uppercased tokens are relevant for this specific corpus and also freezing embedding layers can improve the metric scores, getting a F1 Score of 0.8653.

Keywords: Fake news, explainability, BETO, transformers.

1. Introducción

Los modelos de lenguaje basados en la arquitectura de Transformers han surgido como el actual estado del arte en múltiples tareas de procesamiento de lenguaje natural (PLN) desde su presentación en [9]. El primer modelo de lenguaje desarrollado con esta arquitectura fue BERT (Representación de Codificador Bidireccional de Transformadores por sus siglas en inglés), el cual fue lanzado un año después de la presentación de la arquitectura Transformer y cambió la forma en la que eran abordadas las tareas de PLN [3].

La gran mayoría de los mejores resultados en las tareas de PLN eran obtenidos mediante modelos basados en BERT. Después se elaboraron múltiples variantes del modelo, considerando diferentes arquitecturas, conjuntos de datos, procesos de entrenamiento e idiomas.

Una de las variantes fue BETO presentado por [1], cuya diferencia con BERT consiste en que fue entrenada para el idioma español. La arquitectura de BERT y BETO toma en cuenta solamente la parte del codificador de la arquitectura de Transformers para obtener representaciones de los textos (también llamadas embeddigs) de las secuencias de texto utilizadas como entradas para los modelos.

BETO y BERT en su versión base cuentan con una arquitectura de doce capas, donde a su vez, cada una de ellas cuenta con doce mecanismos de atención o heads, siendo estos la parte central de la arquitectura de los Transformers. Los embeddings o representaciones de los tokens de entrada son actualizados en cada una de las capas a partir de los parámetros del mecanismo de atención, proporcionando doce diferentes representaciones adicionales.

La figura 1 representa cómo funciona el mecanismo de atención. Consiste en una serie de operaciones de producto punto entre vectores, además de la construcción de una matriz de pesos de atención, cuya principal tarea es asignar un valor de peso o importancia para cada uno de los tokens (palabras o signos en los que es dividido el texto) de entrada con respecto a los demás. Es a partir de este mecanismo donde cada token obtiene en su representación un cierto contexto con respecto a los demás.

Uno de los principales retos de utilizar BERT o cualquiera de sus variantes es explicar lo que el modelo es capaz de aprender, tanto del idioma como de la tarea que se está resolviendo. Una forma de explicabilidad consiste en encontrar las propiedades más importantes que el modelo considera dentro de su arquitectura para su aprendizaje en la tarea a resolver.

Este trabajo presenta un análisis de la arquitectura de BETO y cuáles son las características o tokens más importantes para un caso de estudio en particular: la detección de noticias falsas en español, como un primer paso para exponer una posible interpretación de las estimaciones del modelo.

El presente trabajo se estructura de las siguientes secciones: La sección 2 presenta los trabajos relacionados, explicando cómo algunos autores estudian el comportamiento de modelos de lenguaje basados en BERT. La sección 3 describe el conjunto de datos utilizado en este trabajo y la descripción de la arquitectura de los experimentos. La sección 4 presenta los resultados obtenidos y un análisis de éstos.

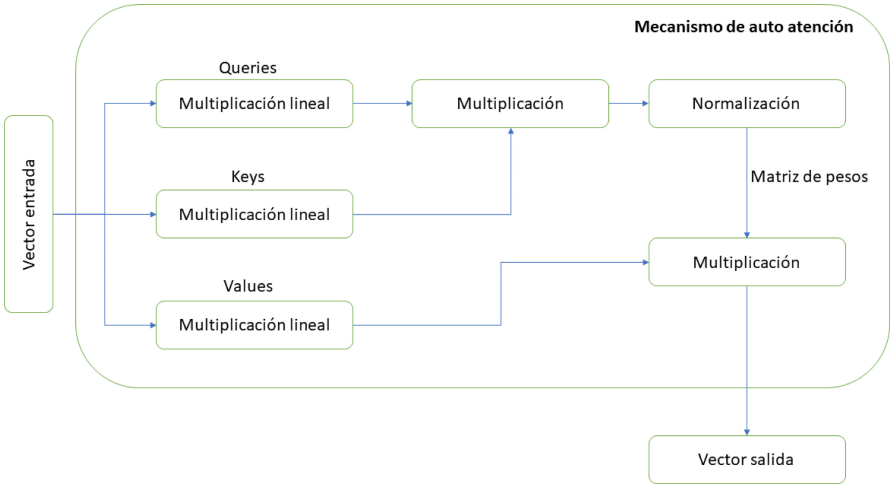

Fig. 1. Diagrama básico del mecanismo de auto atención que utiliza la arquitectura de Transformer. Se observa el punto donde se obtienen los pesos de atención, los cuales son generalmente utilizados para la explicabilidad del modelo.

Y finalmente la sección 5 concluye el presente trabajo, provee algunas conclusiones para el análisis de los modelos de lenguaje para la tarea de la detección de las noticias falsas y describe un posible trabajo a futuro.

2. Trabajo relacionado

Los valores de los pesos de atención encontrados en los mecanismos de atención son comúnmente utilizados para proporcionar la explicabilidad del modelo, ya que permiten presentar la relación que tienen los diferentes tokens respecto a los otros. [2] estudiaron el comportamiento de cada cabeza de atención por capa en BERT y observaron que existen múltiples características lingüísticas intrínsecas en los pesos de atención del modelo. Utilizando tareas de clasificación para diferentes tipos de sintaxis en el idioma inglés, demostraron que tokens especiales del modelo como [CLS] y [SEP] juegan un papel importante en el rol de detectar características sintácticas del lenguaje.

Sin embargo, trabajos posteriores han estudiado que no es suficiente utilizar únicamente los pesos de atención como forma de explicar el comportamiento de este tipo de modelos. [4] compararon las características más importantes encontradas por los pesos de atención contra las características encontradas mediante otros métodos de explicabilidad más tradicionales, como el uso de métodos de selección de características usando valores Chi^2, ganancia de información o information gain, entre otros.

Las características encontradas por los pesos de atención difieren de las encontradas por otros métodos de selección, lo que sugiere que los pesos de atención no proveen la suficiente información como para denotar realmente las características más importantes del modelo. [5] también demostraron que existe una correlación muy débil entre los pesos de atención y las medidas de importancia de características como aquellas basadas en el gradiente descendente.

Tabla 1. Información del corpus de noticias falsas y veridícas a utilizar.

Conjunto de datos	Noticias falsas	Noticias verídicas	Total
Entrenamiento	1036	1444	2480
Validación	133	185	318
Pruebas	311	434	745
Total	1480	2063	3543

Por otro lado, en [10] se concluye que encontrar las características más importantes por medio de los pesos de atención es solo una forma de explicar el comportamiento del modelo, y que simplemente no se debería de considerar como la única ni la mejor manera de proveer dicha información. Con esto presente, se presentaron algunos otros trabajos que proveen diferentes metodologías para explicar el comportamiento de los modelos.

El presente trabajo toma inspiración en el desarrollo presentado por [6], quienes ejecutan una nueva función a partir de los pesos de atención, la norma de los vectores values y una capa de pesos densa que permite la conexión de los valores de los doce mecanismos de atención en una capa.

Demostraron que sus resultados presentan una mejor claridad para entender el comportamiento del modelo en lugar de utilizar únicamente los pesos de atención. La ecuación 1 representa la nueva forma de representar el valor de un vector de un token de entrada, utilizando la nomenclatura de [9]:

$$f(x) := (xW^V + b^V)W^O. \tag{1}$$

En [6] se explicó que en primer lugar se calculan los pesos de atención α, posteriormente los nuevos valores de cada vector x utilizando $f(x)$, después realiza la suma de los pesos y finalmente realiza el cálculo de la norma de $\mathrm{sum}(\alpha f(x))$. La ecuación 2 representa la transformación final del valor que se utiliza en sustitución del solo uso de los pesos de atención presentada en el trabajo anteriormente mencionado:

$$\mathrm{katt}_i = \left\| \sum_{j=1}^{n} \alpha_{i,j} f(x_j) \right\|, \tag{2}$$

donde cada token x de una entrada es transformado por la función $f(x)$ presentada en la ecuación 1, posteriormente se multiplica por su correspondiente peso de atención $\alpha_{i,j}$, luego los valores de todos los tokens son sumados y finalmente representan la atención katt_i obteniendo la norma de ese vector, donde i representa la i-ésima secuencia de entrada de tokens para el modelo.

2.1. Contribuciones de este trabajo

En la actualidad no existe una metodología única o precisa para la explicabilidad de los resultados producidos por modelos de lenguaje. Existen trabajos como el de [6] o el de [4], donde se explica que los pesos de atención no han sido una métrica suficiente para presentar explicabilidad del modelo, ya que no tienden a presentar resultados

Tabla 2. Tipos de modelos utilizados en este trabajo.

Nombre	Freezing embeddings	Uso de mayúsculas
all_layers	No	No
all_layers_cased	No	Sí
all_layers_no_emb	Sí	No
all_layers_no_emb_cased	Sí	Sí

equiparables con otros métodos de explicabilidad o de selección de características más convencionales. [6] requirieron modificar el código base de BERT para que pudieran aplicar los cálculos necesarios para su experimentación. El presente trabajo presenta un análisis inspirado en el trabajo de [6]. En lugar del análisis del modelo BERT, se utiliza BETO, ya que el caso de estudio a analizar se encuentra en el idioma español.

La tarea en cuestión es la detección de noticias falsas, la cual se aborda como una tarea de clasificación binaria. La hipótesis es que mediante la relación del uso de los valores de vectores con los pesos de atención, es suficiente para encontrar una mejor interpretación del modelo que la presentada por el solo uso de pesos de atención.

3. Desarrollo de la solución

3.1. Corpus

El corpus consiste en un conjunto de documentos (textos de noticias) etiquetados de manera binaria como contenido de noticias falsas y verídicas. El corpus consiste en una combinación de los documentos recolectados por [7] y por [8]. La tabla 1 muestra información cuantificada sobre el corpus a utilizar.

3.2. Preprocesamiento del texto

La longitud máxima para cada secuencia de entrada en BETO es de 512 tokens. Significa que incluso cada símbolo de puntuación es convertido en un token y la longitud máxima puede ser alcanzada fácilmente en textos relativamente largos. [2] demostraron que símbolos de puntuación como puntos y comas pueden ser relevantes para el modelo de lenguaje BERT, por lo que se determinó no removerlos.

Además dentro del corpus, el conjunto presentado por [7] contiene enmascarados todos los caracteres numéricos con la etiqueta *NUMBER*, algo que BETO no considera de forma nativa y termina por transformar dicha etiqueta en 3 tokens diferentes. Por lo tanto, estas etiquetas fueron cambiadas por la etiqueta num para simplificar la cantidad de tokens.

3.3. Arquitectura del experimento

Se consideró un proceso de entrenamiento tipo ajuste fino o fine tuning con diez épocas, para dos tipos de modelos: BETO cased y BETO uncased. El modelo BETO cased trabaja con tokens en mayúsculas y minúsculas, mientras que el modelo BETO uncased transforma todos los caracteres a minúsculas.

Tabla 3. Resultados de los experimentos en el corpus de validación.

Modelo	Precisión	Recall	Exactitud	Valor F1
all_layers	0.8377	0.8113	0.8270	0.8243
all_layers_cased	0.8377	0.8113	0.8270	0.8243
all_layers_no_emb	0.8431	0.8113	0.8302	0.8269
all_layers_no_emb_cased	**0.8506**	**0.8239**	**0.8396**	**0.8371**

Tabla 4. Resultados de los experimentos en el corpus de pruebas.

Modelo	Precisión	Recall	Exactitud	Valor F1
all_layers	0.8592	0.8221	0.8443	0.8402
all_layers_cased	0.8595	0.8571	0.8591	0.8583
all_layers_no_emb	0.8508	0.8302	0.8430	0.8404
all_layers_no_emb_cased	**0.8736**	**0.8571**	**0.8671**	**0.8653**

También se experimentó con una variante donde se utiliza el proceso freezing o congelamiento de capas, donde evitamos la actualización de los párametros en ciertas capas de la arquitectura.

En este trabajo se presenta una variante de los modelos donde no se actualizaron los valores de los parámetros en las capas utilizadas para el ajuste de los embeddings, para analizar solamente los valores de las capas que incluyen los mecanismos de atención. La tabla 2 muestra en resumen los tipos de variantes del modelo BETO y la nomenclatura utilizada en las secciones posteriores del trabajo.

3.4. Explicabilidad de los resultados

Para la observación y explicabilidad de los modelos entrenados, se determina la estrategia de encontrar las características o tokens más importantes que el modelo considera en cada una de sus doce capas.

Para la búsqueda de las características más importantes, se obtienen los valores de los pesos de anteción y los vectores values. La ecuación 3 representa el cálculo para la obtención los pesos de atención de acuerdo con [9], lo cual se mantiene en este trabajo:

$$\alpha = \frac{QK^T}{\sqrt{\dim_k}}, \tag{3}$$

donde las matrices Q y K representan los módulos de la multiplicación lineal de Queries y Keys presentadas en la figura 1 y $\dim_k$ es la dimensión o cantidad de tokens k. Entonces se obtienen los vectores values para cada token x y son multiplicados por la matriz de pesos de atención calculada en la ecuación anterior, como se observó en [6]. En este trabajo, se experimenta simplificando la ecuación 1 por la ecuación 4, donde la matriz W^O era la principal causante de recodificar la arquitectura del modelo del lenguaje:

$$g(x) := xW^V + b^V, \tag{4}$$

donde x representa un token de entrada, W^V y b^V son las matrices de pesos y bias de la multiplicación lineal en el módulo values del mecanismo de atención respectivamente.

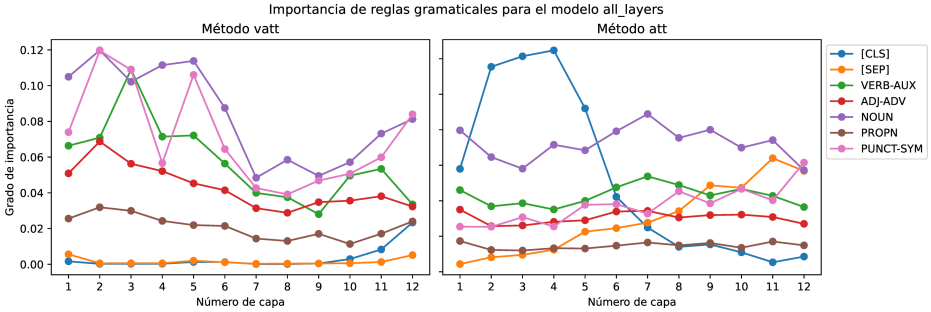

Fig. 2. Comparación de la importancia de categorías gramaticales en los métodos **vatt** y att para el modelo all_layers.

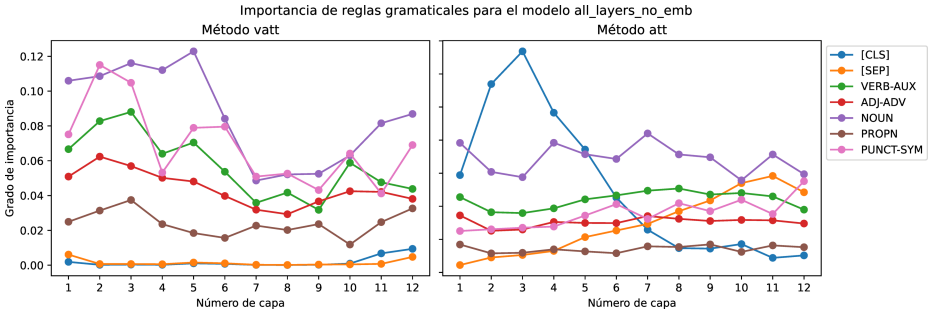

Fig. 3. Comparación de la importancia de categorías gramaticales en los métodos **vatt** y att para el modelo all_layers_no_emb.

Cada capa del modelo tiene doce diferentes mecanismos de atención o heads, cuyos valores son promediados y finalmente son multiplicados por los vectores de $g(x)$, de tal forma que la ecuación 5 presenta la forma en la que este trabajo analiza la explicabilidad del modelo, donde $\alpha'_{i,j}$ representa el valor promedio del peso de atención:

$$\text{vatt}_i = \left\| \sum_{j=1}^{n} \alpha'_{i,j} g(x_j) \right\|. \tag{5}$$

Durante las siguientes secciones, al método convencional de interpretabilidad que hace uso de los pesos de atención, se le nombra como **método att** y al método propuesto se le nombra como **método attv** para su comparación. La interpretabilidad es presentada mediante subconjuntos de los tokens organizados por categorías gramaticales, extrayéndolas con la librería SpaCy.

4. Experimentos y resultados

La tabla 3 y la tabla 4 muestran los resultados obtenidos para el conjunto de validación y pruebas de los experimentos para cada tipo de variante del modelo BETO.

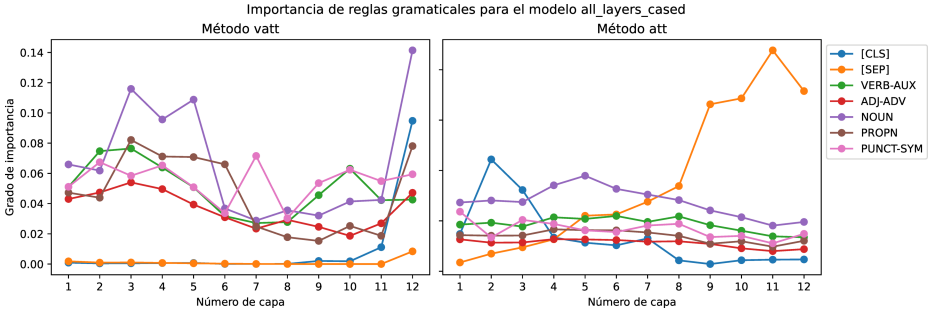

Fig. 4. Comparación de la importancia de categorías gramaticales en los métodos **vatt** y att para el modelo all_layers_cased.

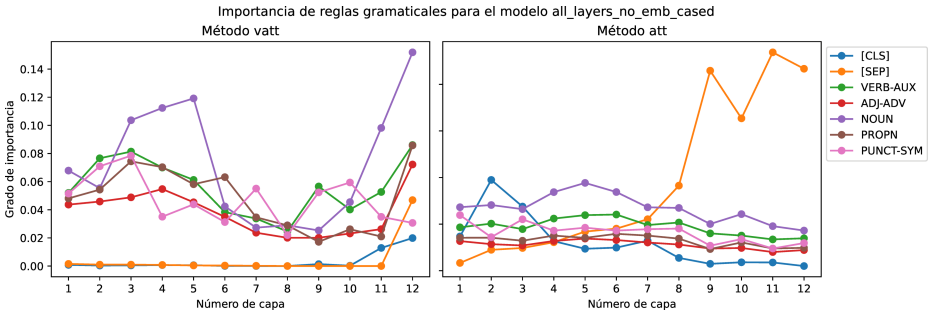

Fig. 5. Comparación de la importancia de categorías gramaticales en los métodos **vatt** y att para el modelo all_layers_no_emb_cased.

El modelo all_layers_no_emb_cased es el que obtuvo los mejores resultados, lo que significa que el uso de mayúsculas y el mantener los valores preentrenados de los embeddings mejora el proceso de detección de noticias falsas para el corpus analizado en este trabajo. Las figuras 2, 3, 4 y 5 muestran una comparación de categorías gramaticales y su importancia en cada una de las doce capas de la arquitectura del modelo BETO. Se presenta el promedio de los resultados obtenidos para cien entradas del conjunto de pruebas.

En adición a las categorías gramaticales se consideran de manera individual los tokens especiales [CLS] y [SEP], ya que para el método **att** estos suelen obtener importancia significativa en ciertas capas del modelo, incluso superando el valor de importancia de las categorías gramaticales.

Sin embargo, se puede observar que los tokens especiales carecen de importancia en el método **vatt**, proporcionando una interpretación del comportamiento del modelo completamente diferente, pero a su vez, conserva coherencia, ya que categorías como sustantivos (NOUN), verbos (VERB-AUX) y signos de puntuación (PUNCT-SYM) presentar un alto nivel de importancia para el modelo.

En particular, la figura 5 muestra el análisis del mejor modelo presentado en este trabajo, donde se aprecia que la categoría gramatical de sustantivos fue de suma importancia para las capas 3-6 y las capas 11-12.

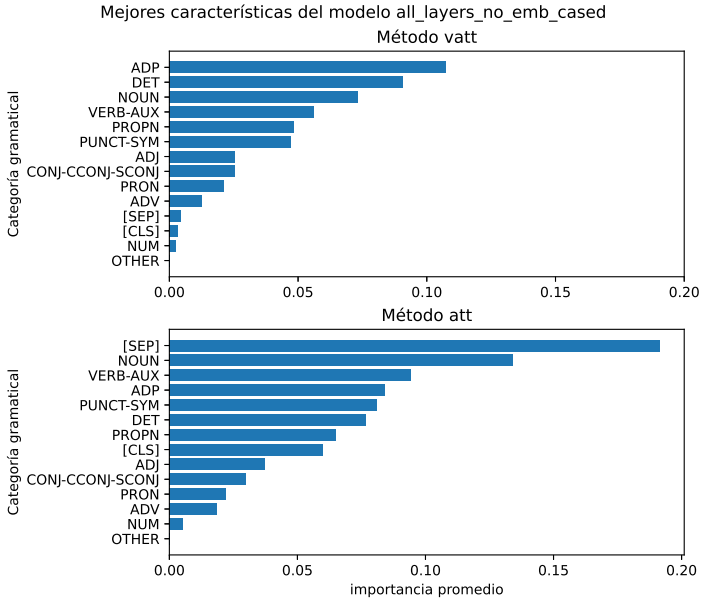

Fig. 6. Comparación de la importancia de todas las categorías gramaticales presentes en el modelo all_layers_no_emb_cased para los métodos **vatt** y att.

Por el contrario, en la figura 2 se observa que no hubo una diferencia tan clara para la categoría gramatical de sustantivos, por lo que puede considerarse como una forma de explicar el por qué ese modelo no obtuvo los mejores resultados. Una vez observada la diferencia entre ambos modelos, lo siguiente es encontrar explicabilidad con el método **vatt** para el mejor modelo presentado en este trabajo. La figura 6 presenta un análisis general de las categorías gramaticales presentes en el modelo all_layers_no_emb_cased, representando el promedio de todas las capas. Para el método **vatt** se concluye que las categorías adposición (ADP) y determinante (DET) son las más importantes en promedio para el modelo.

Ambas categorías cumplen el objetivo de contribuir a la semántica de la oración, por lo que se intuye que el modelo considera que encontrar el contexto de los textos es la parte más importante para la solución a esta tarea. No obstante, para el método **att**, el token especial [SEP] tiene la mayor relevancia, por lo que se puede generar otro tipo de explicabilidad completamente diferente. En este caso, se puede considerar que [SEP] mantiene un contexto de la entrada importante, por lo que el modelo determina que debe tener un peso de atención alto en la mayor parte de la arquitectura.

5. Conclusiones y trabajo a futuro

El presente trabajo presenta un análisis de cómo identificar de manera general las características o tokens más relevantes para la tarea de detección de noticias falsas en textos en español cuando se utilizan modelos de lenguaje basados en Transformers (específicamente en el modelo BETO).

La mejor variante del modelo BETO que se analizó fue aquella que maneja tokens en mayúsculas y minúsculas y no considera actualizar los pesos de los embeddings previamente entrenados. El valor F1 encontrado es de 0.8653 para la clase de noticias falsas. El análisis de un subconjunto de los resultados finales mostró que categorías gramaticales como los sustantivos (NOUN), la adposición o partículas gramaticales (ADP) y determinantes (DET) son las características más importantes para el modelo BETO en la tarea de detección.

El método utilizado para encontrar las características está basado en el trabajo de [6], simplificando en parte la función que se presenta, pero manteniendo resultados en donde los tokens especiales [CLS] y [SEP] presentan importancia insignificante en comparación con tokens más específicos para la tarea.

Como trabajo a futuro, se pretende analizar otros tipos de modelos basados en Transformers, con la finalidad de observar el comportamiento de la metodología propuesta. Además, es posible el análisis de otros tipos de tareas de PLN además de la clasificación, como la generación de textos, entre otras. También, es de interés el experimentar con diferentes idiomas además del español, ya que las categorías gramaticales pueden ser variadas y los modelos puedan atender a diferentes características para una misma tarea.

Referencias

1. Cañete, J., Chaperon, G., Fuentes, R., Ho, J. H., Kang, H., Pérez, J.: Spanish pre-trained BERT model and evaluation data. In: Practical ML for Developing Countries Workshop and International Conference on Learning Representations, pp. 1–10 (2020)
2. Clark, K., Khandelwal, U., Levy, O., Manning, C. D.: What does BERT look at? An analysis of BERT's attention (2019) doi: 10.48550/ARXIV.1906.04341
3. Devlin, J., Chang, M. W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding (2018) doi: 10.48550/ARXIV.1810.04805
4. Garcia-Silva, A., Gomez-Perez, J. M.: Classifying scientific publications with BERT - is self-attention a feature selection method? In: Lecture Notes in Computer Science, pp. 161–175 (2021) doi: 10.1007/978-3-030-72113-8_11
5. Jain, S., Wallace, B. C.: Attention is not explanation (2019) doi: 10.48550/arxiv.1902.10186
6. Kobayashi, G., Kuribayashi, T., Yokoi, S., Inui, K.: Attention is not only a weight: Analyzing transformers with vector norms. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (2020) doi: 10.18653/v1/2020.emnlp-main.574
7. Posadas-Durán, J. P., Gómez-Adorno, H., Sidorov, G., Moreno-Escobar, J. J.: Detection of fake news in a new corpus for the Spanish language. Journal of Intelligent and Fuzzy Systems, vol. 36, no. 5, pp. 4869–4876 (2019) doi: 10.3233/jifs-179034
8. Tretiakov, A.: Noticias falsas en español (2020)
9. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in Neural Information Processing Systems, vol. 30 (2017)
10. Wiegreffe, S., Pinter, Y.: Attention is not not explanation (2019) doi: 10.48550/arXiv.1908.04626