

TRANSFORMERS APLICADO A LA DETECCIÓN Y ANÁLISIS DE AGRESIVIDAD EN REDES SOCIALES

Victor Manuel Gómez Espinosa

Asesores:

Dr. Victor Muñoz Sánchez

Dr. Adrián Pastor López Monroy

1. OBJETIVOS

- General
- **Identificación y análisis de agresividad** en redes sociales mediante métodos de aprendizaje profundo.



1. OBJETIVOS

- Particulares

1. **Adaptación de un modelo** en el estado del arte basado en la arquitectura **Transformer** (Vaswani et al. 2017) a la tarea específica de **identificación de agresividad en tweets en español de México (MEX-A3T 2020)**.
2. **Incorporar diferentes canales de información** (Sentimiento, estilo, sintáctico, etc.), **en la arquitectura Transformer** para modelar diferentes características relacionadas con los tweets (Tsai et al. 2019).

2. JUSTIFICACIÓN

- En los últimos años han proliferado distintas **redes sociales** y su uso ha ido en aumento, pero esto consecuentemente también trae **problemas** como lo son las diferentes **manifestaciones lingüísticas** por parte de los usuarios como lo son el **ciber acoso, racismo u odio**. En el contexto de esta problemática, se define al **lenguaje agresivo como aquel que busca causar daño y puede incitar a la violencia**, lo cual es a lo que se exponen los usuarios en redes sociales, y esto pueden **causar daños a largo plazo, en algunas ocasiones llevando al suicidio**, de ahí la importancia de buscar formas de identificar este tipo de manifestaciones (Aragón et al. 2020).



2. JUSTIFICACIÓN

- **Identificar la agresividad no es un problema fácil**, puesto que no sólo depende de la presencia o ausencia de palabras.

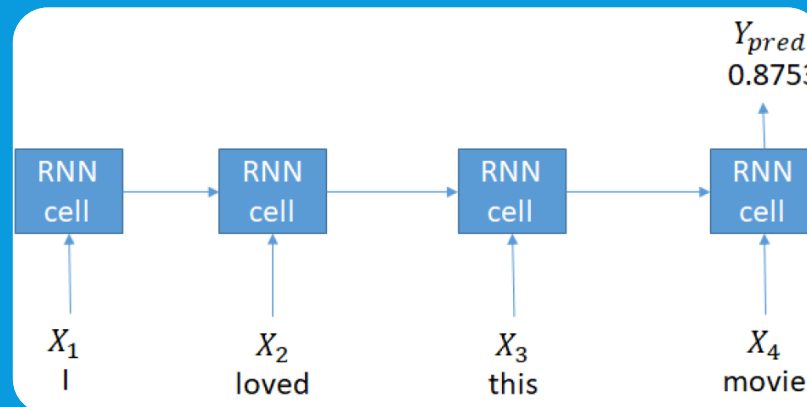
Categoría predicha: Agresivo
Mugrosa gorda jajaja
Categoría predicha: No agresivo
Hablar no sirve para ni madre
Categoría predicha: Agresivo
Todavía no es Halloween y ya estoy disfrazada de tu pendeja

*pero estas gorda... aprovecha tu fin
pendeja que el lunes te violo*

Ejemplos del corpus MEX-A3T 2020: a) tweets con lenguaje ofensivo o vulgar pero no agresivo. b) tweet agresivo

2. JUSTIFICACIÓN

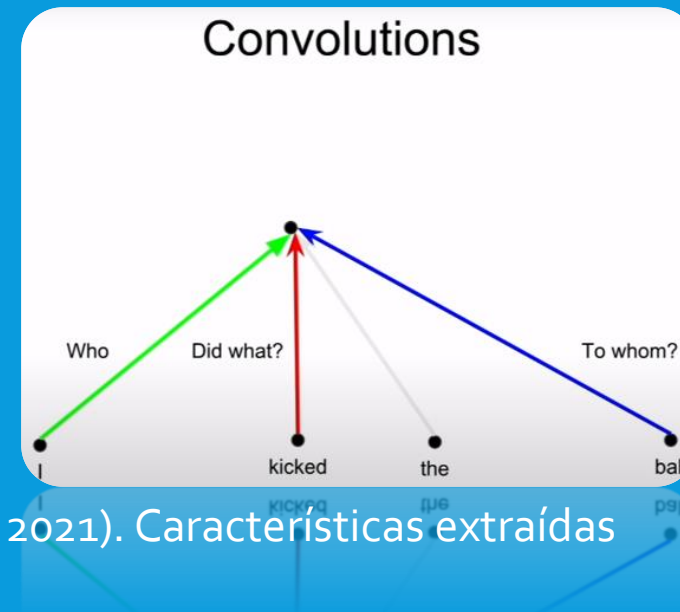
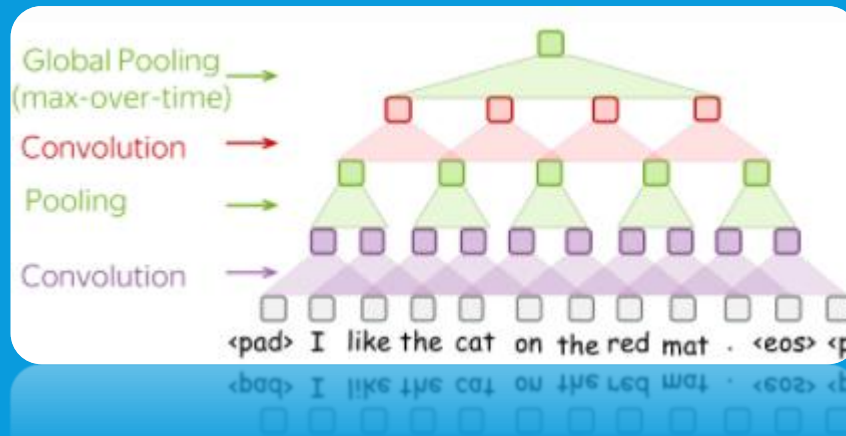
- Hasta hace algunos años las **redes neuronales recurrentes** eran el **enfoque estándar para modelado de secuencias** en general, y de textos en particular, como sucede con algunas tareas en el área de procesamiento de lenguaje natural, pero estos enfoques tienen la **desventaja de que son dependientes del paso de tiempo** anterior lo que causa **limitantes** en el entrenamiento ya que **no permite la paralelización** además de **perdida de información para secuencias largas** debido al **vanishing gradient**.



Modelado de secuencias con RNN
(Cecchini, datacamp)

2. JUSTIFICACIÓN

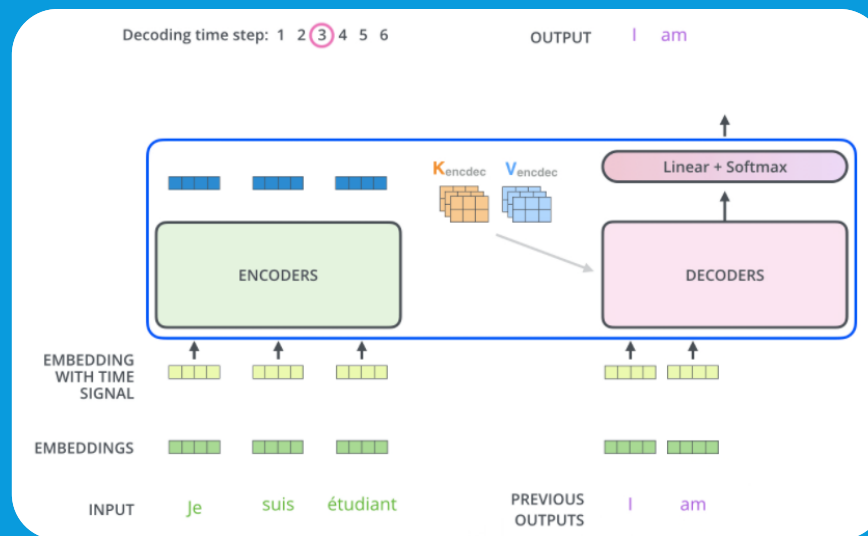
- Se ha explorado también, el uso de **redes neuronales convolucionales** para encontrar dependencias entre posiciones distantes del texto, pero estas tienen la **desventaja de que se requieren crecer en profundidad** (más capas) haciéndolas **más complejas**.



Representación de la convolución en textos (Voita 2021). Características extraídas mediante convolución (Stanford, 2019)

2. JUSTIFICACIÓN

- La arquitectura del **Transformer** aparece como una alternativa a estos problemas, **utilizando únicamente mecanismos de auto atención**, permitiendo la paralelización y **reduciendo los tiempos en el entrenamiento** (Vaswani et al. 2017).






Representación del Transformer
(Alammar, 2018)

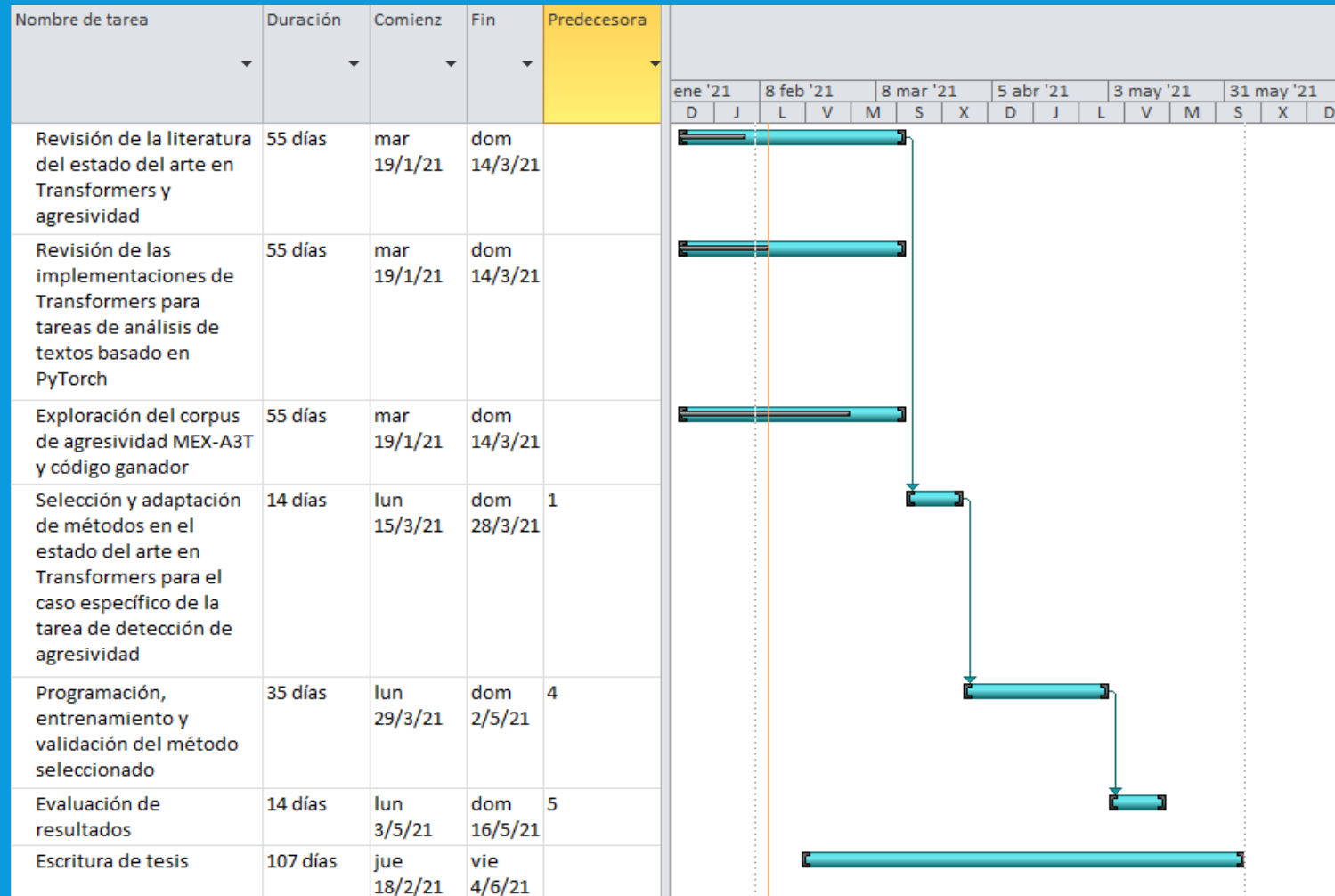
2. JUSTIFICACIÓN

- Aunque la arquitectura del Transformer originalmente fue propuesta para la tarea de traducción, **se han explorado diversas adaptaciones** y en la actualidad **existen diversos modelos en el estado del arte** basados en su arquitectura para diversas tareas **en el área de NLP** como lo son **GPT-2 y GPT-3** (Radford et al. 2018) para generar textos, **BERT** (Devlin et al. 2019) para embeddings, **T5** (Raffel et al. 2020) para distintas tareas como Q&A, clasificación, traducción y resumen automático, lo que representa al Transformer como una notable área de investigación vigente.

3. METODOLOGÍA

-  1. **Revisión de la literatura** del estado del arte en **Transformers** y **agresividad**
-  2. **Revisión de las implementaciones** de Transformers para tareas de análisis de textos basado en **PyTorch**
-  3. **Exploración del corpus** de agresividad **MEX-A3T** y código ganador
- 4. **Selección y adaptación de métodos** en el estado del arte en Transformers para el caso específico de la tarea de detección de agresividad
- 5. **Programación, entrenamiento y validación** del método seleccionado
- 6. **Evaluación de resultados**
- 7. **Escritura de tesis**

4. CRONOGRAMA



5. REFERENCIAS CITADAS

- Aragón, M. E., Jarquín-Vasquez, H., Montes-Y-Gómez, M., Escalante, H. J., Villasenõr-Pineda, L., Gómez-Adorno, H., Posadas-Durán, J. P., and Bel-Enguix, G. (2020), “Overview of mex-a3t at iberlef 2020: Fake news and aggressiveness analysis in mexican Spanish,” *CEUR Workshop Proceedings*, 2664, 222–235.
- Devlin, J., Chang, M.-W., Lee, K., Google, K. T., and Language, A. I. (2019), *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*.
- Guzman-Silverio, M., Balderas-Paredes, Á., and López-Monroy, A. P. (2020), “Transformers and data augmentation for aggressiveness detection in mexican Spanish,” *CEUR Workshop Proceedings*, 2664, 293–302.
- Radford, A., Narasimhan, K., Salimans, T., and Sutskever, I. (2018), *Improving Language Understanding by Generative Pre-Training*.
- Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. (2020), *Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer*, *Journal of Machine Learning Research*.
- Tsai, Y. H. H., Bai, S., Liang, P. P., Zico Kolter, J., Morency, L. P., and Salakhutdinov, R. (2019), “Multimodal transformer for unaligned multimodal language sequences,” *Association for Computational Linguistics*, 6558–6569. <https://doi.org/10.18653/v1/p19-1656>.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017), “Attention is all you need,” *arXiv*.

6. REFERENCIAS FIGURAS

- Alammam, J. (2018), “The Illustrated Transformer – Jay Alammam – Visualizing machine learning one concept at a time.,” Github, Available at <http://jalammar.github.io/illustrated-transformer/>.
- Cecchini, D. (n.d.). “Recurrent neural networks for language modeling in Python,” *datacamp*.
- Stanford (2019), “Stanford CS224N: NLP with Deep Learning | Winter 2019 | Lecture 14 –Transformers and Self-Attention - YouTube,” standfordonline, Available at https://www.youtube.com/watch?v=5vcj8kSwBCY&list=PLakWuueTN59e7ck3fB5lvy_aHphUvMfgA&index=6&t=854s.
- Voita, L. (2021), “Convolutional Models for Text,” *Github*, Available at https://lena-voita.github.io/nlp_course/models/convolutional.html.