Modelo Text GCN para la clasificación de texto

Text GCN Model for text classification

DOI: 10.46932/sfjdv5n5-005

Received on: Apr 02nd, 2024
Accepted on: Apr 22nd, 2024

Moisés García Villanueva
Maestria en Ingeniería Eléctrica
Institución: Universidad Michoacana de San Nicolás de Hidalgo, Facultad de Ingeniería Eléctrica
Dirección: Av. Francisco J. Mújica S/N, C.P. 58030, Morelia Michoacán
Correo electrónico: moises.garcia@umich.mx

Salvador Ramírez Zavala
Maestria en Ingeniería Eléctrica
Institución: Universidad Michoacana de San Nicolás de Hidalgo, Facultad de Ingeniería Eléctrica
Dirección: Av. Francisco J. Mújica S/N, C.P. 58030, Morelia Michoacán
Correo electrónico: salvador.ramirez@umich.mx

RESUMEN
El problema de clasificación de texto es una actividad fundamental en el área de Procesamiento de Lenguaje Natural (PLN). Recientemente las redes neuronales de grafos (GNN) han mostrado ser de las mejores técnicas en diferentes conjuntos de datos representativos en este problema de clasificación. Las redes convolucionales de grafos son aplicados para decodificar estructuras sintácticas en los documentos o frases y entonces son aplicados a la clasificación de texto considerando la construcción del grafo mediante dos enfoques: a) un grafo por cada documento o frase; y b) un grafo completo del conjunto de datos. En las investigaciones reportadas los mejores resultados se obtienen empleando el enfoque de grafo completo de conjuntos de datos en inglés. En este trabajo se presentan los resultados preliminares de emplear está nueva técnica de clasificación de documentos en un conjunto de datos en español, lo que permitirá contar además del modelo de red neuronal de grafo para el conjunto, con una representación vectorial de palabras.

Palabras clave: Redes Neuronales de Grafos, Clasificación de Texto, Procesamiento de Lenguaje Natural, GCN.

ABSTRACT
The text classification problem is a fundamental activity in the area of Natural Language Processing (NLP). Recently, graph neural networks (GNN) have shown to be one of the best techniques on different representative data sets in this classification problem. Convolutional graph networks are applied to decode syntactic structures in documents or sentences and are then applied to text classification considering the construction of the graph using two approaches: a) a graph for each document or sentence; and b) a complete graph of the data set. In the reported research, the best results are obtained using the complete graph approach of English data sets. This work presents the preliminary results of using this new document classification technique in a Spanish data set, which will also allow the graph neural network model for the set to have a vector representation of words.

Keywords: Graph Neural Networks, Text Classification, Natural Language Processing, GCN.

1 INTRODUCCIÓN

Los estudios de clasificación de texto mediante el aprendizaje profundo (DL por sus siglas en Inglés de Deep Learning) puede ser categorizado en dos grupos: el primero se refiere al grupo de estudios enfocado en los modelos basados en encajes de palabras (Mikolov *et al*. 2013; Pennington, *et al*., 2014). Algunos autores fusionan el aprendizaje de palabras y documentos etiquetados como encajes (Tang *et al*., 2015; Wang *et al*. 2018). La diferencia entre estos métodos principalmente se refiere a la forma en como se construye la representación del texto después del aprendizaje de encajes de palabras, otros enfoques aprenden encajes de palabras y documentos en forma simultanea para la clasificación de texto. El segundo grupo de estudios emplean redes neuronales profundas. Dos de las arquitecturas representativas son redes neuronales convolucionales (CNN de Convolutional Neural Network) y redes neuronales recurrentes (RNN de Recurrent Neural Network) (Kim, 2014). Aún cuando estos métodos son efectivos y ampliamente utilizados, principalmente se enfocan en secuencias de palabras consecutivas locales, es decir, no hacen uso explicito de las palabras globales y de la información de coocurrencia en el conjunto de datos de estas palabras. Dentro de estos tipos de redes y con la expectativa de considerar información global en un conjunto de datos, se le ha dado recientemente al tópico de Redes Convolucionales de Grafos (GCN por sus siglas en Inglés de Graph Convolutional Networks) una gran atención. En el trabajo inicial de Kipf y Welling 2017 se presentó el modelo de red neuronal de grafos, al cual se le denominó GCN, en el cual se obtuvieron resultados del estado del arte en la clasificación de texto empleando diferentes conjuntos de datos representativos en esta tarea, todos ellos en el idioma Inglés. GCN maneja el etxto como un conjunto de aristas que relacionan palabras, consoideradas como nodos en el grafo y entonces procede a aprender las características de los nodos en una forma específica. En 2016, Defferrard *et al*., utilizó primero un GCN para la tarea de clasificación de texto y alcanzó una eficiencia innovadora. A partir de entonces, una gran cantidad de investigadores han aplicado GCN a la clasificación de texto (Peng *et al*., 2024). En este trabajo se presenta la definición de un GCN y los resultados de una clasificación de texto binaria en un conjunto de datos en español del algoritmo Text GCN.

Proporcione una descripción de la contextualización, tema de investigación y justificación de la investigación utilizando la fuente Times New Roman tamaño 12, con un interlineado de 1,5. El máximo número de autores permitidos es ocho; si el artículo excede este límite, debe comunicarse con la revista para consultar sobre el cargo adicional por agregar otro autor.

En cuanto a la longitud del manuscrito, debe constar de un máximo de 20 páginas, incluidas las referencias. Los trabajos pueden estar escritos en inglés o español.

Al final de la introducción se deben delinear claramente los objetivos del trabajo, de forma específica y medible. Si lo deseas, puedes crear un subelemento exclusivo para el objetivo. Además, es

esencial que estén formulados de manera alcanzable, asegurando que el lector comprenda completamente el alcance del estudio y lo que se cubrirá y evaluará.

2 RED CONVOLUCIONAL DE GRAFO

Un GCN (Kipf y Welling 2016) es una red neuronal multicapa que opera directamente sobre un grafo que induce vectores de encajes de nodos basados en las propiedades de sus nodos vecinos. Formalmente un grafo se define como G = (V, E), donde V y E son conjuntos de nodos y aristas respectivamente. Cada nodo se asume que está conectado así mismo, representando está conexión como (v, v) ∈ E para cualquier v. Sea $X \in R^{n \times m}$ que representa una matriz conteniendo todos los n nodos con sus características, mientras que m es la dimensión del vector de características, cada renglón en la matriz, $x_v \in R^m$, es el vector de caraterísticas para el nodo v. Se establece una matriz de adyacencia A para el grafo G y su matriz de grado D, donde $D_{i,i} = \sum_j A_{i,j}$. Los elementos de la diagonal de la matriz A tienen el valor de 1, por su conexión así mismo. GCN puede capturar solemente información acerca de los vecinos inmediatos por medio de una capa convolucional. Cuando múltiples capas en un GCN son apiladas, la información acerca de vecinos más alejados es integrada. Para una sola capa en el GCN, la nueva matriz de características de los nodos es de dimensión k, $L^{(1)} \in R^{n \times k}$ y se calcula mediante (1).

$$L^{(1)} = \rho(\hat{A} \times W_0) \qquad (1)$$

dónde:

$\hat{A} = D^{-1/2} A D^{-1/2}$ es la matriz de adyacencia simétrica normalizada

$W_0 \in R^{m \times K}$ es una matriz de pesos.

ρ función de activación, por ejemplo la función ReLU que se define como ρ(x) = max(0, x).

La función de activación ρ, incorpora información de orden superior de vecinos cercanos, implica apilar múltiples capas en el GCN, esto se representa por (2).

$$L^{(j+1)} = \rho\left(\hat{A} L^j W_j\right) \qquad (2)$$

en donde

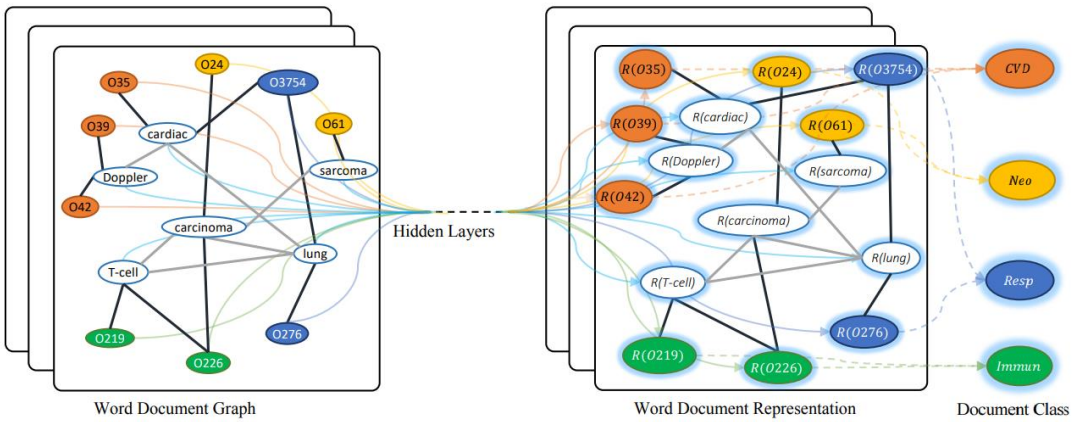
el subindice j representa el número de capa y $L^{(0)} = X$.

El marco teórico en un estudio comprende un análisis crítico y organizado de la literatura relevante al tema, proporcionando una contextualización teórica y definiendo los conceptos clave. Debe contener de manera integral teorías, modelos e investigaciones previas, identificando vacíos, contradicciones y consensos en la literatura que sean importantes para el enfoque del trabajo que se desarrolla.

2.1 TEXT GCN

El trabajo de Yao (*et al*., 2019) describe Text GCN, un método que consiste en construir un grafo grande y de texto heterogeneo, el cual contiene palabras y documentos como nodos, de tal forma que la coocurrencia de palabras globales puede ser explicitamente modelada y la convolución en el grafo puede ser adaptada fácilmente. El número de nodos en el grafo de texto es el número de documentos (tamaño del conjunto de datos) más la cantidad de palabras únicas (tamaño del vocabulario) en el corpus. Las aristas contruidas entre los nodos se establecen en base a la coocurrencia de palabras en los documentos (aristas entre palabras). Se asignan los pesos de una arista entre un nodo del tipo documento y un nodo que representa una palabra mediante el valor TF-IDF (Term Frequency Inverse Document Frecuency) de la palabra en el documento. TF o término frecuencia se refiere al número de veces que la palabra aparece en el documento, mientras que IDF es la escala logaritmica de la fracción inversa del número de documentos que contienen la palabra que representa el nodo. Una medida alternativa para obtener los pesos que asocian las palabras es PMI (Point-wise Mutual Information). La Figura 1 es un diagrama esquemático del método Text GCN. Los nodos que inician con “O” son nodos con documentos, el resto son palabras. Las aristas de color negro son para indicar aquellas aristas entre un documento y una palabra, las grises y más delgadas son aristas entre palabras. R(x) significa la representación vectorial (encajes) de un documento o una palabra. Los colores señalan diferentes clases de documentos.

Figura 1. Diagrama esquemático del método Text GCN.



Fuente: Yao (*et al*., 2019).

3 RESULTADOS Y DISCUSIONES

La pregunta que se plantea en esta sección al evaluar el modelo GCN, se refiere especificamente para determinar lo siguiente:

¿Pueden obtenerse resultados satisfactorios en la tarea de clasificación de texto con documentos en español, empleando un conjunto de datos pequeño?

La clasificación de texto se realizó en la tarea de noticias falsas, este es un problema de interés en PLN. Debido a que las redes sociales e internet les ha dado a los individuos el poder de crear y compartir su propio contenido, este generalmente es parcial y no verificado. En el concurso FakeDeS (Gómez-Adorno *et al*., 2021) se produjo un conjunto de noticias falsas en español de dominio público. El conjunto consiste de un total de 971 noticias, de las cuales 491 son verdaderas y 480 son falsas. Los tópicos involucrados en este conjunto de datos son 9: ciencia, deportes, economía, Educación, entretenimiento, política, salud, seguridad y sociedad. Con este conjunto de datos se entrenó el GCN utilizando un 70% del total de documentos y el 30% restante se utilizó como conjunto de datos de prueba. La Tabla 1 nos muestra los mejores resultados obtenidos al utilizar diferentes cantidades de unidades ocultas en la capa 1 del GCN y el número de epocas para obtener esos resultados. Se estableccieron los parámetros por default del método Text GCN a excepción de los establecidos con sus respectivos valores para el proceso de entrenamiento del GCN y que fueron: Dropout = 0.5, taza de aprendizaje = 0.01.

Tabla 1. Resultados de clasificación de noticias falsas a través de Tect GCN

Unidades Ocultas capa 1	Epocas	Precisión	Macro F1-Score
12	50	0.7601	0.7603
14	45	0.7534	0.7531
15	40	0.7545	0.7521
20	39	0.7471	0.7471
80	14	0.7672	0.7095
150	23	0.7578	0.7541
200	21	0.7591	0.7574

Fuente: Elaboración propia.

Los resultados son competitivos con los reportados para la competencia FakeDeS (2021) por Gómez-Adorno (*et al*., 2021), en donde el mejor equipo obtuvó una precisión de 0.7657 y una Macro F1 de 0.7666. La línea base establecida para la competencia para la precisión fue de 0.7290, se consideró el uso de un clasificador de máquinas de soporte vectorial como clasificador de texto para establecerla.

4 CONCLUSIONES

Se presentó el uso de un clasificador de texto, utilizando y describiendo a grandes rasgos el método Text GCN. Yao (*et al*. 2019) quien fue el que propuso Text-GCN y quien aplicó por primera vez el uso de está técnica en la clasificación de texto (Peng *et al*., 2024). Los resultados para documentos en español son competitivos a los reportados por los creadores del conjunto de datos de noticias falsas en español. Se observa la necesidad de crear recursos en el idioma español e implementar las herramientas de las nuevas técnicas que se han generado en el área de Inteligencia Artificial.

REFERENCIAS

Defferrard, M., Bresson, X. & Vandergheynst, P. (2016). Convolutional neural networks on graphs with fast localized spectral filtering. *Advances in neural information processing systems*, *29*.

Gómez-Adorno, H., Posadas-Durán, J. P., Enguix, G. B. & Capetillo, C. P. (2021). Overview of FakeDeS at IberLEF 2021: Fake News Detection in Spanish Shared Task. *Procesamiento del Lenguaje Natural*, *67*, 223-231.

Kim, Y. 2014. Convolutional neural networks for sentence classification. *EMNLP*, 1746–1751.

Kipf, T. N. & Welling, M. (2016). *Semi-supervised classification with graph convolutional networks*. arXiv preprint arXiv:1609.02907.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. 2013. Distributed representations of words and phrases and their compositionality. *NIPS*, 3111–3119.

Peng, Y., Wu, W., Ren, J. & Yu, X. (2024). Novel GCN Model Using Dense Connection and Attention Mechanism for Text Classification. *Neural Processing Letters*, *56*(2), 1-17.

Pennington, J., Socher, R. & Manning, C. (2014). Glove: Global vectors for word representation. *EMNLP*, 1532– 1543.

Tang, J., Qu, M & Mei, Q. (2015). Pte: Predictive text embedding through large-scale heterogeneous text networks. *KDD*, 1165–1174. ACM.

Wang, Y., Huang, M., Zhao, L. et al. (2016). Attention-based lstm for aspect-level sentiment classification. *EMNLP*, 606–615.

Yao, L., Mao, C., & Luo, Y. (2019, July). Graph convolutional networks for text classification. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 33, No. 01, pp. 7370-7377).