Polo del Conocimiento

Pol. Con. (Edición núm. 68) Vol. 7, No 3
Marzo 2022, pp. 1282-1295
ISSN: 2550 - 682X
DOI: 10.23857/pc.v7i3.3791

Clasificación y etiquetado de tweets de Ecuador para determinar qué tema tratan, utilizando un modelo Transformer

Classification and labeling of tweets from Ecuador to determine what topic they are about, using a Transformer model

Classificação e rotulagem de tweets do Equador para determinar de qual tópico eles tratam, usando um modelo Transformer

César Humberto Espin-Riofrio [I]
cesar.espinr@ug.edu.ec
https://orcid.org/0000-0001-8864-756X

Kerlly Vera-Guamán [II]
kerlly.verag@ug.edu.ec
https://orcid.org/0000-0001-5019-5465

Ricardo Yela-García [III]
ricardo.yelag@ug.ecu.ec
https://orcid.org/0000-0001-8423-4645

Correspondencia: cesar.espinr@ug.edu.ec

Ciencias Técnicas y Aplicadas
Artículo de Investigación

***Recibido:** 30 de enero de 2022 ***Aceptado:** 25 de febrero de 2022 * **Publicado:** 21 marzo de 2022

I. Magister en Sistemas de Información Gerencial, Ingeniero en Electricidad Especialización Electrónica, Universidad de Guayaquil, Guayaquil, Ecuador.
II. Universidad de Guayaquil, Guayaquil, Ecuador.
III. Universidad de Guayaquil, Guayaquil, Ecuador.

Resumen

El presente artículo tiene como uno de sus objetivos el estudio y establecimiento del estado de arte del Procesamiento de Lenguaje Natural, así como también identificar los métodos más utilizados para la tarea de clasificación y etiquetado de textos basados en el idioma español a través de la revisión y comparación de diferentes artículos científicos de relevancia y trabajos académicos relacionados. Se procederá a experimentar con el modelo Transformer Selectra-Medium para clasificación de textos cortos, utilizando mensajes de la red social Twitter de usuarios de Ecuador como fuente de datos en idioma español, los mismos serán almacenados, procesados, clasificados y finalmente etiquetados para poder identificar de qué temas tratan de forma automática. A través de la utilización del modelo se establecen categorías previamente definidas como sociedad, economía, entretenimiento, salud, deportes y delincuencia sobre las cuales procede la clasificación. Se busca obtener una proyección de los temas de interés que tratan los usuarios agilitando tareas de análisis de textos, dichos resultados podrán ser beneficiosos como aporte a las investigaciones sobre el tema.

Palabras Clave: Procesamiento de Lenguaje Natural; Transformers; Selectra; Clasificación de textos.

Abstract

The present article has as one of its objectives the study and establishment of the state of the art of Natural Language Processing, as well as to identify the most used methods for the task of classification and labeling of texts based on the Spanish language through the review and comparison of different relevant scientific articles and related academic papers. We will proceed to experiment with the Transformer Selectra-Medium model for the classification of short texts, using messages from the social network Twitter of Ecuadorian users as a source of data in Spanish language, which will be stored, processed, classified and finally labeled in order to be able to identify the topics they deal with automatically. Through the use of the model, previously defined categories such as society, economy, entertainment, health, sports and crime are established on which the classification proceeds. The aim is to obtain a projection of the topics of interest that users deal with, thus speeding up text analysis tasks, and these results may be beneficial as a contribution to research on the subject.

Keywords: Natural Language Processing; Transformers; Selectra; Text Classification.

Resumo

Este artigo tem como um de seus objetivos o estudo e estabelecimento do estado da arte do Processamento de Linguagem Natural, bem como identificar os métodos mais utilizados para a tarefa de classificar e rotular textos baseados na língua espanhola através da revisão e comparação de diferentes artigos científicos relevantes e trabalhos acadêmicos relacionados. Prosseguiremos experimentando o modelo Transformer Selectra-Medium para classificar textos curtos, usando mensagens da rede social Twitter de usuários equatorianos como fonte de dados em espanhol, eles serão armazenados, processados, classificados e finalmente rotulados para poder identificar com quais tópicos eles lidam automaticamente. Por meio do uso do modelo, são estabelecidas categorias previamente definidas como sociedade, economia, entretenimento, saúde, esporte e delinquência sobre as quais procede a classificação. Busca-se obter uma projeção dos temas de interesse com os quais os usuários lidam agilizando as tarefas de análise de texto, esses resultados podem ser benéficos como contribuição para pesquisas sobre o assunto.

Palavras-chave: Processamento de Linguagem Natural; transformadores; Seleção; Classificação de texto.

Introducción

El Procesamiento del Lenguaje Natural (PLN) es una rama de la Inteligencia Artificial (IA) que se encarga de que un ordenador sea capaz de entender lo que dice un ser humano en su lenguaje natural, dándole sentido e interpretando el significado de las palabras (Khurana et al., 2017). El PLN ha ayudado a facilitar el trabajo de muchas tareas entre ellas los resúmenes automáticos, predicción de textos, traductores automáticos, respuestas a preguntas, etc., debido a la relación que tiene con la IA y el Big Data (Sánchez, 2020), dando a surgir modelos inspirados en el PLN como Alexa un asistente virtual controlado por voz desarrollado por Amazon, Sophia un robot humanoide creado por la Cía. Hanson Robotics, diseñado para imitar o simular los movimientos de un ser humano, esto y muchas otras tecnologías que surgen con los grandes avances de la ingeniería informática. Por otro lado actualmente, se evidencian abundantes cantidades de textos que son generados en la web a través de usuarios, que en su mayoría estos datos no son procesados ni estructurados, por lo que tomaría grandes cantidades de esfuerzo y tiempo para un ser humano, pero gracias al PLN es conveniente poder analizar y clasificar dicha información de forma

automática a través de sus técnicas y métodos, convirtiéndola en información útil y necesaria.

Es por ello que el presente artículo tiene como objetivo comparar los métodos de PLN más utilizados para la clasificación y etiquetado de textos cortos en idioma español clasificando tweets extraídos de la red social Twitter de usuarios del Ecuador, para determinar sobre qué tema tratan utilizando Selectra-Medium[1] como modelo Transformers de clasificación en categorías como sociedad, política, educación, salud, economía y entretenimiento

Grandes avances que se han tenido en PLN desde sus inicios, entre ellos se puede mencionar (Turing, 1950) en su prueba que consistía en que un ser humano mantenga una conversación con otra persona y con una computadora, sin que se pudiera identificar quien de los conversadores es la máquina. Por otro lado (Hutchins, 2004) menciona que, con la colaboración de IBM y la Universidad de Georgetown en 1954, realizaron un sistema de traducción automática del ruso al inglés, en el cual se utilizó 250 palabras y seis reglas de gramática. Otros avances se tuvieron en el área de la sintaxis (Chomsky, 1965), fue el de poder clasificar en dos niveles, nivel superior constituidos por el reconocimiento de voz y el nivel inferior correspondiente al lenguaje natural. En 1980 los algoritmos de Aprendizaje Automático (Machine Learning) llegan al PLN, agilitando tareas entre ellos la clasificación de documentos y detección de personas, aumentando el enfoque en datos a través de técnicas computacionales. (Hochreiter & Schmidhuber, 1997) propusieron Long Short Term Memory (LSTM) redes neuronales recurrentes capaces de aprender la dependencia del orden en problemas de predicción de secuencias, utilizada en problemas complejos como la traducción automática, reconocimientos de voz y más. (Gers et al., 2000) a través del LSTM obtuvieron grandes avances, incluso se creó un sistema en 2009, en el cual identifica un texto escrito a mano y lo convierte a un texto digital, dejando atrás todas las técnicas que había en reconocimiento de imagen. Es así que también surge el Aprendizaje Profundo (Deep Learning) en el que (Hinton et al., 2006) desarrollaron una manera más rápida para entrenar estos modelos, su objetivo principal es realizar tareas como identificación de imágenes, predicción de palabras o reconocimiento de voz aprendiendo a través de patrones y convirtiendo a las redes neuronales (Neural Networks) cada vez más complejas y capaces de entrenarse. (Mikolov et al., 2013) presentaron los Word2Vec, que son una red neuronal en el cual asigna cada palabra a un espacio o lista de números llamados vectores, en el que tenga tantas posiciones posibles como palabras queramos tener, usando técnicas de Word Embedding, los cuales buscan aquella codificación

[1] https://huggingface.co/Recognai/zeroshot_selectra_medium

semántica y la relación que representa las palabras como vectores de números reales (García, 2018). Los modelos Transformers, introducidos en el artículo "*Attention is All you need"* (Vaswani, 2017), que son una de las arquitecturas dominantes para el PLN, superando modelos neuronales convolucionales y recurrentes, con un mayor rendimiento en tareas de comprensión y de generación de lenguaje natural que requieren de menor tiempo de entrenamiento (González & Centenera, 2020). Uno de los primeros modelos fue GPT - Generative Pretrained Transformer creado por (Openai et al., 2018), el cual comprende el entrenamiento de un conjunto de datos en el que el modelo es capaz de aprender tareas como respuesta a preguntas, traducción automática o resumen de textos sin ninguna supervisión explícita (Beltrán & Rodríguez, 2021). Bidirectional Encoder Representations from Transformer - BERT desarrollador por GoogleAI (Devlin et al., 2018), este modelo se diseñó para entrenar representaciones bidireccionales profundas sin etiquetas para todas las capas (Beltrán & Rodríguez, 2021). GPT-2 (Radford et al., n.d.) una versión mejorada de GPT, sus resultados generan textos con coherencia y han tenido mejoras en los chatbots, GPT-3 (Brown et al., 2020), aprendizaje de disparo cero (Zero shot), es autorregresivo y usa aprendizaje profundo, entrenados con conjuntos de datos de lenguaje general de gran tamaño como Wikipedia Corpus y Common Crawl, (Arámbula Cosío et al., 2021). Otro modelo como BART (Lewis et al., 2019) y T5 (Raffel et al., 2019) son algunos de los muchos modelos basados en Transformers. En la Figura 1, se puede apreciar la evolución de los modelos más importantes.

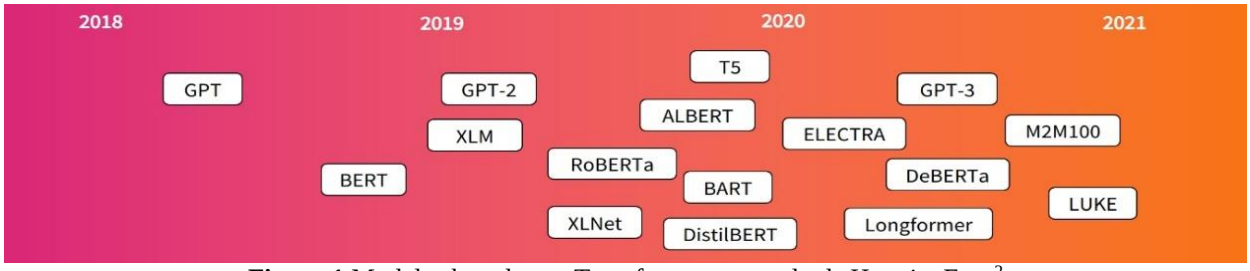

Figura 1 Modelos basados en Transformers, tomado de HuggingFace[2]

Metodología

Se utilizó la metodología de investigación bibliográfica analizando información relevante de diferentes artículos científicos y trabajos académicos en temas sobre clasificación y etiquetado de

[2] https://huggingface.co/course/chapter1/4?fw=pt

Pol. Con. (Edición núm. 68) Vol. 7, No 3, Marzo 2022, pp. 1282-1295, ISSN: 2550 - 682X

textos cortos. A continuación, se describe los métodos utilizados en diferentes trabajos.

Existen métodos que nos permiten saber de diferentes maneras como se realiza la clasificación y etiquetado de textos. (Vapnik, 1995) desarrolló el método de clasificación Máquina de Soporte Vectorial (Support Vector Machine SVM), (Chamorro, 2018) en su trabajo que consiste en un sistema para la detección de eventos relacionados con el tráfico y la contaminación a partir del análisis de datos de Twitter, en el que concluye que el mejor método evaluado es el de SVM, debido a que representa un mayor valor de exactitud para la clasificación de eventos de tráfico y no tráfico. (Cuenca & León, 2008) comenta que los algoritmos de aprendizaje supervisado se utilizan para solucionar problemas de clasificación y regresión. Otro método son los árboles de decisiones presentado por (Quinlan, 1996) estos son un método de clasificación supervisado que consiste en construir un árbol con múltiples caminos donde cada nodo tiene el atributo que proporciona la mayor utilidad de información para cada una de sus clases. (Elgueta, 2017) en su trabajo hace el uso de Árboles de decisión, Naive Bayes y Support Vector Machine, para saber qué técnica funciona mejor en clasificación de textos de los titulares de periódicos en el idioma español. (Vidal Ruiz, 1986) propuso un método de clasificación supervisada denominado algoritmo del vecino más próximo (Nearest Neighbour), que consiste en ser uno de los más sencillos de implementar, la idea del algoritmo es cuando se calcula la similitud entre el documento a clasificar y cada documento de entrenamiento, la puntuación más próxima indica en qué clase o categoría debe clasificarse el documento (Sancho Caparrini, n.d.). (Pérez et al., 2017) como solución a diferentes problemas de técnicas lineales y no lineales emerge un nuevo método llamado “aprendizaje basado en kernell” en donde las representaciones del kernel brindan una alternativa, aumentando el poder computacional del aprendizaje automático lineal, al proyectar los datos en un espacio de características multidimensional. (López, 2018) destaca que el algoritmo de “kernelización” sustituye el producto escalar de la representación original al evaluar la función del núcleo. A través de la revisión y análisis de estos otros numerosos trabajos relacionados se determinó que los métodos más utilizados en tareas de clasificación de texto son Support Vector Machine (SVM), Naive Bayes Classifiers (NBC), Decision Trees (DT) y K-Nearest-Neighbor (KNN).

Para la investigación experimental se hace el uso del modelo SELECTRA-MEDIUM preentrenado de Transformers, el cual clasifica y etiqueta textos cortos en español, con el fin de obtener resultados sobre los temas de interés de los usuarios del país. Se detallan los pasos llevados a cabo para la experimentación y análisis de los datos en el que se realizó la extracción de datos,

preprocesamiento de datos, clasificación, etiquetado y finalmente se presenta los resultados obtenidos.

Extracción de datos

Se realizó la extracción de 500 tweets, en donde se instaló la librería Twint para extraer los datos y se procedió a configurar los parámetros necesarios tales como la localización geográfica que es Ecuador, que los tweets sean en idioma español, el límite de tweets a extraer que son 500 y finalmente se estableció el rango de la fecha qué serán extraídos, recibiendo como parámetros el username, tweet y el lenguaje como se muestra en la Figura 2.

username	tweet	language
xsavageal	@Combo_Ex Gracias combo no eres un buen negociante con la gente de activision que fue a tu casa	es
siwonka09	@Marosas13 @SJofficial Ya con lo que dijo kyu en el video de los "10 paises" ya me hago a la idea que Ecuador no estará incluido en las proyecciones dla peli, de alguna manera me da rabia (no con ellos obvio) sino xq al menos pa el boleto d cine comparado con1entrada al super show si me alcanzaba :(	es
drianrogelio	La histórica selección de los #PaisesBajos NL regresa a una Copa del Mundo, tras perderse #Rusia2018, los tres veces finalista del orbe (1974, 1978, 2010), una vez más al mando de Louis Van Gaal sellan su boleto a #Qatar2022 tras vencer 2-0 a Noruega NO en el Stadion Feijenoord!! https://t.co/yx6NTEH6VT	es
clickradioudla	La inseguridad desde la comunicación: La entrevista junto a las voces de #periodismo y #EcuadorAtento EC https://t.co/C6FI5xkzhJ	es
educacionz6_ec	.@MeryVicua, directora del Distrito Cuenca Norte, mantuvo una reunión con delegados de la Universidad del Azuay, con el objetivo de articular acciones para definir proyectos a aplicarse en función de convenio con la universidad del Azuay. #EncontrémonosPorLaEducación https://t.co/2lgt0xlIRd	es
leonelaa23		und
j_villarreals	@SofikaVikinga2 Agüita de arroz a de votar.	es
radiodiblufm	¡INVITADOS DE LUJO! YA LOS TENEMOS a estos cracks en nuestra previa del partido de #LaTriEC contra #ChileCL . Conéctate a FB Live: https://t.co/DL9Fzr0P8D #EliminatoriasPorDiblu https://t.co/cQPa3Cea6L	es

Figura 2 Extracción de los 500 Tweet.

Preprocesamiento de datos

En el preprocesamiento de datos, se usó la librería Tweet-preprocessor para eliminar información no útil en su respectiva clasificación y etiquetado, como son los emojis, hashtag, menciones @, enlaces, los mismo que fueron localizados y eliminados a través del uso de librerías como Emoji, también se normalizó los textos convirtiéndolos de mayúsculas a minúsculas, eliminar tweets vacíos y finalmente son almacenados los tweets preprocesados en un archivo csv. En la Figura 3, se presentan los tweets preprocesados en un dataframe.

index	tweet
0	gracias combo no eres un buen negociante con la gente de activision que fue a tu casa
1	ya con lo que dijo kyu en el video de los "10 paises" ya me hago a la idea que ecuador no estará incluido en las proyecciones dla peli, de alguna manera me da rabia (no con ellos obvio) sino xq al menos pa el boleto d cine comparado con1entrada al super show si me alcanzaba
2	la histórica selección de los regresa a una copa del mundo, tras perderse , los tres veces finalista del orbe (1974, 1978, 2010), una vez más al mando de louis van gaal sellan su boleto a tras vencer 2-0 a noruega en el stadion feijenoord!!
3	la inseguridad desde la comunicación: la entrevista junto a las voces de y
4	., directora del distrito cuenca norte, mantuvo una reunión con delegados de la universidad del azuay, con el objetivo de articular acciones para definir proyectos a aplicarse en función de convenio con la universidad del azuay.
6	agüita de arroz a de votar.

Figura 3 500 Tweet preprocesados

Clasificación y etiquetado

Los tweets preprocesados, pasan al modelo seleccionado Selectra-Medium de Transformers, se realizó cambios al modelo original, ya que este permite realizar una clasificación y etiquetado de un solo texto, se adaptó el modelo para que permita evaluar una gran cantidad de textos cortos en el idioma español recibiendo una cantidad de tweets almacenados en un solo archivo, así clasifica y etiqueta mensajes cortos de usuarios de Twitter en Ecuador, con etiquetas previamente definidas como ámbitos sociales, entretenimiento, salud, deportes, economía, delincuencia, a su vez, muestre el porcentaje más alto por cada texto clasificado según la categoría. Como se puede visualizar en la Figura 4, la clasificación de cada tweet mostrando un score de relación entre el tweet y la etiqueta.

```
Tweet: xfavor escriban solo interesados en comprar mi contenido o deseen pagar x la salida. si empiezan a preguntar x hacer perder el tiempo los bloqueo. le
 Labels: entretenimiento     Score: 0.6989607214927673
Tweet: ¡victoria que beneficia a la tri! 3-0 los del altiplano golearon a los charrúas que no levantan cabeza y peligran clasificar a qatar.
 Labels: deportes     Score: 0.5718531012535095
Tweet: anabell salinas: "el programa de capacitación para operación minera graduó a 306 alumnos de la provincia de zamora chinchipe como operadores de mina
 Labels: sociedad     Score: 0.3343869149684906
Tweet: | para garantizar el retorno progresivo y seguro a clases presenciales, , en coordinación con su par de , desarrollaron una jornada de vacunación mas
 Labels: salud     Score: 0.7805933356285095
Tweet: los influencers de  ya empiezan a ventilar q lo que propuso en campaña no se pude hacer, ahhh por cierto y será culpa de correa,
 Labels: entretenimiento     Score: 0.30398860573768616
Tweet: ¿hola como estás? no te vuelvas a quedar sin internet, te doy el doble de velocidad x 6 meses, 1 factura gratis, instalación sin costo en 2 horas, en
 Labels: entretenimiento     Score: 0.41386619210243225
Tweet:  nuestro rector recibió hoy la visita de carlos loaiza, presidente de , quien presentó los proyectos de la comisión de la ciudad. este espacio ciudad
 Labels: sociedad     Score: 0.9700083868026733
```

Figura 4 Aplicación del modelo Selectra medium

Resultados

A través de la ejecución de los pasos mencionados anteriormente, se obtuvo los siguientes resultados de promedio en cada categoría como se muestra en la Tabla 1.

Tabla 1 Tweets clasificados según su categoría

Categoría	**Promedio de análisis de tweets**
Economía	10.9%
Salud	11.4%
Deportes	14.3%
Delincuencia	17.4%
Entretenimiento	21.4%
Sociedad	24.6%

Se aprecia que, de los 500 tweets extraídos, preprocesados y clasificados, la categoría con más afluencia fue sociedad con 24.6% seguido de entretenimiento con 21.4% y en último lugar economía con 10.9%.

Se realiza una función random, en donde de los 500 tweets se tomarán únicamente 10 de forma aleatoria para realizar una prueba del modelo en la clasificación de los tweets, Figura 5 muestra los resultados obtenidos

	tweet	labels	scores
309	lee mi bro, dice "por".	economia	0.3424
281	el tipo d camisa azul d la esquina miren la ca...	sociedad	0.2489
287	ahí te hablan a ti lo que te recibirán algún d...	delincuencia	0.297
298	amigos espero hoy no sea de esos días que le ...	deportes	0.2885
214	| esta mañana compartimos nuestros conocimient...	salud	0.4063
252	[] fausto cobo, director encargado del ingresó...	deportes	0.3137
329	este tipo de noticias en tendría que estar aco...	delincuencia	0.925
195	'existirán los votos para destituir a eckenner...	sociedad	0.3471
6	agüita de arroz a de votar.	delincuencia	0.5051
102	mi suerte la usé en encontrarla a ella	salud	0.8878

Figura 5 Tweets aleatorios clasificados

Pol. Con. (Edición núm. 68) Vol. 7, No 3, Marzo 2022, pp. 1282-1295, ISSN: 2550 - 682X

En la Tabla 2, se muestra los promedios obtenidos por cada categoría tomando tweets de manera aleatoria.

Tabla 2 Promedio de las categorías de tweets aleatorios

Categoría	**Promedio de análisis de tweets**
Entretenimiento	0%
Economía	10%
Salud	20%
Deportes	20%
Sociedad	20%
Delincuencia	30%

Se observa, de los 10 tweets tomados aleatoriamente, la categoría con más afluencia es delincuencia con 30% seguido de sociedad con 20% y al final entretenimiento con 0%.

Discusión

La mayoría de los tweets son escritos de diferentes formas en donde la parte semántica y la sintaxis juegan un papel importante en la comprensión de los textos, resultando una tarea difícil al momento de interpretar un texto, por eso es importante llevar a cabo un buen preprocesamiento de datos.

Por otro lado la clasificación de tweets va a depender conforme susciten acontecimientos en el país, por ejemplo, si se toma una muestra de los tweets del mes de noviembre 2021 sus resultados se posicionarían en la categoría deporte por tratarse de fecha de eliminatorias sudamericana al mundial Qatar 2022, o si son de fecha abril 2020 el tema seguramente sería salud al estar en plena pandemia Covid, es así que se podrá observar e identificar que el modelo funciona correctamente proporcionando resultados que se pueden verificar acorde a la fecha en que los tweets fueron extraídos.

Existen trabajos relacionados para la clasificación y etiquetado de texto que en su mayoría son para el idioma inglés, en el cual en este artículo se basó en textos cortos en idioma español debido a que

es de amplio uso en Iberoamérica.

Conclusiones

Con el análisis de contribuciones científicas de impacto sobre el tema, haciendo énfasis en investigaciones realizadas para el idioma español, se estableció el estado de arte y se identificó los métodos más utilizados para la clasificación y etiquetado de textos cortos, entre ellos los algoritmos de Naive Bayes Multinomial, Desicion Tree, KNN y Support Vector Machine. En la experimentación e implementación del modelo SELECTRA-MEDIUM de Transformers, se logró clasificar exitosa y automáticamente gran cantidad de tweets, concluyendo que el modelo es capaz de clasificar textos cortos según las temáticas que se le definan. Los resultados obtenidos y presentados en el presente artículo contribuyen a las investigaciones realizadas sobre la clasificación de textos cortos y el uso de modelos Transformers como herramienta de gran avance en el estudio de Procesamiento de Lenguaje Natural.

Referencias

1. Arámbula Cosío, F., Emmanuel Maqueda Bojorquez, D., Luis Morales-Reyes, J., Gabriel Acosta-Mesa, H., Nora Aquino-Bolaños, E., Herrera-Meza, S., Cruz-Ramírez José Luis Chávez-Servia, N., Hevia-Montiel, N., Mota Antonio Neme, S., Arámbula Cosío, F., Torres Robles, F., Velásquez-Rodríguez, G., Galicia Gómez, E., Escalante-Ramirez, B., Olveres, J., Pérez, J. L., Medina Bañuelos, V., Camargo Marín Guzmán Huerta, L. M., Fanti, Z., … Hazan Lasri Arámbula Cosío, E. F. (2021). *De redes neuronales recurrentes a modelos de lenguaje: la evolución del PLN en la generación de textos Clasificación de poblaciones nativas de frijol utilizando visión artificial Las anomalías: ¿qué son?, ¿dónde surgen?, ¿cómo detectarlas? Aprendizaje com*.
2. Beltrán, N. C., & Rodríguez, E. C. (2021). Procesamiento del lenguaje natural (PLN) - GPT-3.: Aplicación en la Ingeniería de Software. *Tecnología Investigación y Academia*, *8*(1), 18–37. https://revistas.udistrital.edu.co/index.php/tia/article/view/17323
3. Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., … Amodei, D. (2020).

Language Models are Few-Shot Learners. *Advances in Neural Information Processing Systems*, *2020-December*. https://arxiv.org/abs/2005.14165v4

4. Chamorro, V. (2018). *CLASIFICACIÓN DE TWEETS MEDIANTE MODELOS DE APRENDIZAJE SUPERVISADO*.
5. Chomsky, N. (1965). *Aspects of the theory of syntax*. 251.
6. Cuenca, D., & León, D. (2008). *SUPPORT VECTOR MACHINE*.
7. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, *1*, 4171–4186. https://arxiv.org/abs/1810.04805v2
8. Elgueta, J. (2017). *Comparación de rendimiento de técnicas de aprendizaje automático para análisis de afecto sobre textos en español*. http://repobib.ubiobio.cl/jspui/handle/123456789/1772
9. García, I. (2018). *Estudio de word embeddings y métodos de generación de meta embeddings*. https://addi.ehu.es/handle/10810/29088
10. Gers, F. A., Schmidhuber, J., & Cummins, F. (2000). Learning to forget: Continual prediction with LSTM. *Neural Computation*, *12*(10), 2451–2471. https://doi.org/10.1162/089976600300015015
11. González, S., & Centenera, C. (2020). *Estudio del rendimiento de BERT frente a métodos clásicos de procesamiento de lenguaje natural*.
12. Hinton, G. E., Osindero, S., & Teh, Y. W. (2006). A Fast Learning Algorithm for Deep Belief Nets. *Neural Computation*, *18*(7), 1527–1554. https://doi.org/10.1162/NECO.2006.18.7.1527
13. Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, *9*(8), 1735–1780. https://doi.org/10.1162/NECO.1997.9.8.1735
14. Hutchins, W. J. (2004). *The Georgetown-IBM experiment demonstrated in January 1954*.
15. Khurana, D., Koli, A., Khatter, K., & Singh, S. (2017). *Natural Language Processing: State of The Art, Current Trends and Challenges*. https://arxiv.org/abs/1708.05148v1
16. Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., & Zettlemoyer, L. (2019). *BART: Denoising Sequence-to-Sequence Pre-training for*

Natural Language Generation, Translation, and Comprehension. 7871–7880. https://doi.org/10.18653/v1/2020.acl-main.703

17. López, A. (2018). *Fundamentos Matemáticos de los Métodos Kernel para Aprendizaje Supervisado*. 73.
18. Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. *1st International Conference on Learning Representations, ICLR 2013 - Workshop Track Proceedings*. https://arxiv.org/abs/1301.3781v3
19. Openai, A. R., Openai, K. N., Openai, T. S., & Openai, I. S. (2018). *Improving Language Understanding by Generative Pre-Training*. https://gluebenchmark.com/leaderboard
20. Pérez, S. A., Profesor Guía, V., Alfaro, R., Profesor Co-Referente, A., Héctor, :, & Cid, A. (2017). *"Análisis y Clasificación de Textos con Técnicas Semi Supervisadas Aplicado a Área Atención al Cliente."*
21. Quinlan, J. R. (1996). Learning decision tree classifiers. *ACM Computing Surveys (CSUR)*, *28*(1), 71–72. https://doi.org/10.1145/234313.234346
22. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (n.d.). *Language Models are Unsupervised Multitask Learners*. Retrieved March 2, 2022, from https://github.com/codelucas/newspaper
23. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2019). Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research*, *21*, 1–67. https://arxiv.org/abs/1910.10683v3
24. Sánchez, J. (2020). *Análisis del estado del arte de la generación de texto con redes neuronales mediante modelos de Transformer*.
25. Sancho Caparrini, F. (n.d.). *Aprendizaje Inductivo: Árboles de Decisión*.
26. Turing, A. (1950). *Maquinaria computacional e Inteligencia*.
27. Vapnik, V. N. (1995). The Nature of Statistical Learning Theory. *The Nature of Statistical Learning Theory*. https://doi.org/10.1007/978-1-4757-3264-1
28. Vaswani, A. (2017). Attention Is All You Need arXiv:1706.03762v5. *Advances in Neural Information Processing Systems*, *2017-Decem*(Nips), 5999–6009. https://www.autoriteitnvs.nl/onderwerpen/stralingsincidenten/inhoud/incidenten-met-a--

en-of-b-objecten

29. Vidal Ruiz, E. (1986). An algorithm for finding nearest neighbours in (approximately) constant average time. *Pattern Recognition Letters*, *4*(3), 145–157. https://doi.org/10.1016/0167-8655(86)90013-9