UNIVERSIDAD DEL VALLE DE GUATEMALA
Data Science
Sección 10
Ing. Lynette García Pérez

Proyecto 2

Procesamiento del lenguaje natural con tweets de desastres naturales

Alexa Carolina Bravo Vicente (18831)
José Eduardo López Gómez (181045)
Karina Alejandra Valladares Díaz (18005)

Guatemala, noviembre 19 de 2021

Introducción

Es innegable el crecimiento que han tenido los alcances de la computación y la ciencia de datos en la época actual. Además de resolver problemas computacionales y retratables por medio de modelos matemáticos, la ciencia actual y los procesos de estudio de datos nos permite explorar mucho más allá de esto. Uno de dichos casos es el creciente interés que ha experimentado el procesamiento del lenguaje natural, por medio del cual es posible entrenar a una máquina a cómo entender a los humanos y comunicarse con ellos. Dicho avance supone un puente más de comunicación y puede ser crucial en ciertos escenarios.

Por otro lado, en la actualidad las redes sociales se han convertido en una parte natural de la comunicación y una fuente de información rápida y accesible. Son además una forma de comunicación masiva que puede ser utilizada en escenarios imprevistos. Por su parte, la red social Twitter tiene potencial para ser una buena herramienta en caso de emergencia. Dado que es posible aprovecharse de la información estructurada que proporciona la red, analizando los textos escritos cortos de Twitter, los conocidos *tweets*. A pesar de que son textos de baja calidad, el límite en la longitud permite fácilmente otorgar un mensaje contundente y rápido.

Es especialmente conveniente los efectos de Twitter y su límite de caracteres cuando se intenta realizar reportes. Por otro lado, incluso el ecosistema de Twitter permite realizar análisis rápidos y acertados. En este trabajo, el enfoque es la clasificación de *tweets*, y la evaluación de los mismos como un reporte directo de un evento de desastre natural. Se aplicaron tres modelos distintos para la clasificación de los mismos y se evaluó su rendimiento. Además, se probó directamente con *tweets* recientes, por lo que fue posible evaluar los modelos con los datos proporcionados y con datos completamente nuevos.

Objetivos del proyecto

Objetivo general

El objetivo general del proyecto es la creación de un modelo de aprendizaje automático que sea capaz de analizar y evaluar los datos presentados en *tweets*, para determinar si la información presentada en ellos es verdadera o falsa.

Objetivos específicos

- Definir un problema científico real y determinar un método de solución para el mismo.
- Determinar qué técnicas usadas dentro del Procesamiento del Lenguaje Natural (NLP, por sus siglas en inglés) son pertinentes para el estudio en cuestión.
- Realizar un análisis detallado del problema planteado y crear una solución teórica para el mismo.
- Realizar un proceso de limpieza para los datos, puesto que son lenguaje natural.
- Generar una aplicación replicable para la detección de desastres naturales presentados en *tweets*.

Marco teórico

Procesamiento del lenguaje natural

El procesamiento de lenguaje natural, o NLP, por sus siglas en inglés, es una rama de la ciencia de computación enfocada en otorgar a las computadoras la habilidad de comprender y entender palabras habladas y escritas, de forma muy similar a la que los humanos lo hacemos. Para ello, se combina el modelaje del lenguaje humano con procesos estadísticos, *machine learning* y *deep learning*. Dada la ambigüedad del lenguaje humano, existen múltiples técnicas y prácticas dentro del NLP, con el fin de volverlo más sencillo de procesar para las máquinas (IBM Cloud Education, 2020).

Dada la complejidad del lenguaje humano, en general existen técnicas preliminares para desglosar el lenguaje, de forma que colabora con la computadora en la comprensión de lo que está analizando. Entre estas se incluyen:

- Reconocimiento de habla: es la habilidad en la que el texto hablado se convierte en un texto escrito. Dicha tarea es necesaria para cualquier aplicación que sigue con comandos de voz o responde a preguntas orales.
- Clasificación de categorías gramaticales: en este proceso se determinan qué categorías gramaticales están presentes dentro de cada oración y se van clasificando de acuerdo a estas.
- Desambiguación del sentido de las palabras: es la selección del sentido y significado de las palabras con varios significados a partir de un análisis semántico.
- Análisis sentimental: intenta la extracción de cualidades subjetivas del texto

(IBM Cloud Education, 2020).

Ali Sit, Koylu y Demir (2018) introdujeron en su estudio un espacio de trabajo analítico encargado de *tweets*. Este fue diseñado para identificar y categorizar detalles sobre un desastre natural, como cantidad de afectados, infraestructura dañada, servicios interrumpidos, áreas de impacto diferenciadas y periodos de tiempo y su relevancia. Para ello, el modelo planteado sigue la identificación de los *tweets* relacionados con desastres naturales al crear un dataset de entrenamiento, consistente en etiquetas generadas por humanos y siguiendo con la experimentación continua de métodos de *deep learning* y *machine learning* para la clasificación binaria.

Para este análisis con las herramientas mencionadas anteriormente, los autores resaltan el uso de redes de Long Short-Term Memory (LSTM). Específicamente hablando, el uso de los LSTM es para la tarea de la clasificación, puesto que supera en practicidad y tiempo de ejecución a los demás métodos para este propósito, puesto que considera la construcción textual completa con semántica y dependencias. Luego, se aplicó una

clasificación no supervisada con múltiples etiquetas de *tweets* utilizando Latent Dirichlet Allocation (LDA), que es un modelo de tópicos generativo. Este modelo asume que cada palabra en un documento es generada a partir de un tópico que es tomado de una distribución de tópicos para cada documento, por lo que así se identifican categorías latentes incluidas en cada texto (Ali Sit, Koylu y Demir; 2018).

Por último, los autores mencionados comentan que luego de estos procesos, utilizan suavizado espacial del núcleo y suavizado basado en la densidad para agrupar como forma de identificación. Es decir, por medio de estos procesos, se determina la prominencia relativa y el área de impacto para cada categoría de información. Dado que utilizando el huracán Irma como un estudio base, se tomaron en cuenta más de 500 millones de colecciones de texto y locaciones antes, durante y después del desastre. Los resultados obtenidos resaltan áreas con alta densidad de afectados y daños en la infraestructura a lo largo de la progresión del desastre (Ali Sit, Koylu y Demir; 2018).

Por su parte, Periñal-Pascual y Arcas-Túnez (2018) en su estudio *The Analysis of Tweets to Detect Natural Hazards* indican que la detección de eventos trágicos o fenómenos naturales es abordada como un problema de clasificación. Este se compone de dos aspectos principales: la categorización de tópicos y el análisis de sentimientos. Para ello, los investigadores señalan que existen dos distintos abordajes, por medio de *machine learning,* implementado a través de un proceso de aprendizaje supervisado o un acercamiento simbólico, que se basa, principalmente, en una base de conocimiento. Un método de aprendizaje supervisado, como Naïve Bayes o máquinas de vectores de soporte (SVM, por sus siglas en inglés) requiere un dataset de entrenamiento, que consiste en una colección de texto que ha sido manualmente anotada como positiva o negativa, en referencia al evento objetivo. Este set de entrenamiento debe estar cuidadosamente catalogado, pero también debe ser suficientemente grande y representativo, lo que genera conflictos con el desarrollo de un sistema. Por lo descrito anteriormente, los autores decidieron entonces seguir un un acercamiento basado en conocimiento previo.

Por su parte, Gupta, Negi, Vishwakarma, Rawat, Badhani y Tech (2017) realizaron un estudio de sentimientos utilizando algoritmos de Python de *machine learning*. Para el análisis, el objetivo principal es la clasificación de los *tweets* en diferentes categorías. Para ello, se han desarrollado varias formas de abordaje, en la que se proponen los métodos y el entrenamiento. Entre estos métodos, se encuentran

- Regresión logística bayesiana: selecciona características y provee optimizaciones para la categorización de textos. Anterior a este proceso, utiliza Laplace para evitar el sobreajuste y produce modelos predictivos ligeros para el texto. La estimación tiene la forma paramétrica

$$P(C|F) = \frac{1}{z(f)} exp((\Sigma \lambda_{i,c} F_{i,c}(f, c)))$$

en donde *z(f)* es una normalización, λ es un vector de pesos paramétricos y $F_{i,c}$ es una función binaria que toma una característica y una etiqueta de clase.

- Clasificador bayesiano ingenuo: es un clasificador probabilístico con un fuerte supuesto en la independencia condicional, que es óptima para las clases de clasificación con una alta cantidad de características dependientes. La adherencia de las clases de sentimientos están dadas por el teorema de Bayes.
- Algoritmo de máquinas de vectores de soporte: son modelos supervisados con algoritmos asociados de aprendizaje que analizan los datos usados para la clasificación y análisis de regresión.

Por otro lado, el modelo *long short-term memory* (LSTM) es un tipo especial de redes recurrentes. Se caracteriza al tener la información y ser capaz de reproducir bucles en el diagrama de la red, de forma que "recuerda" los estados previos. Dicha capacidad de memoria del modelo le permite entonces utilizar la información para decidir cuál será el siguiente estado. El objetivo principal de las redes LSTM es la creación de un modelo de mantenimiento predictivo, lo cual puede ser utilizado para alertar sobre la existencia de desastres naturales según el texto analizado. Para obtener una predicción adecuada, esta red se debe entrenar (Colah, 2015).

El entrenamiento es un proceso de aprendizaje iterativo en el que hay instancias de datos en las que se conoce la salida correcta. Para ello, se utiliza el mecanismo de *train-test split*, que es un procedimiento rápido y sencillo. Los resultados de este permiten realizar comparaciones respecto al comportamiento y el éxito de los algoritmos de aprendizaje de máquinas. El procedimiento involucra tomar un dataset y dividirlo en dos subconjuntos. El primero es utilizado para adecuarse al modelo y se denomina dataset de entrenamiento. El segundo subconjunto es entonces el elemento de entrada para el modelo, y las predicciones son realizadas y comparadas con los valores esperados (Brownlee, 2020).

Metodología del proyecto

Solución del problema

A partir de la información recolectada y de las investigaciones realizadas, se determinó un plan a seguir para la solución del problema. Dado que el objetivo era crear un modelo de predicción de desastres naturales, basado en *tweets*, se procedió de forma inicial con un análisis exploratorio de los datos disponibles para su comprensión. A partir del mismo, se determinó la naturaleza del dataset, así como los procesos necesarios para su uso posterior.

Luego de los hallazgos, se determinó una función encargada del preprocesamiento de la información. Es decir, se decidió crear una función para limpiar el texto –en el escenario, los *tweets*–, que luego será utilizado para las predicciones. Ya con un texto limpio, el mismo se convierte en una serie de tokens, los cuales serán utilizados por los modelos seleccionados.

Selección de datos de entrenamiento y prueba

Los datos seleccionados fueron obtenidos de la página Kaggle, puesto que forman parte de una competencia abierta. De estos, a pesar de que se presentaban tres distintos dataframes, se seleccionó únicamente el de nombre *train.csv*, puesto que dichas fueron las instrucciones dadas. Es importante destacar que dicho dataset contiene una variable llamada *target*, que es binaria, e indica si el tweet en cuestión sí corresponde a un desastre natural.

Luego de limpiar los datos que se obtuvieron de la página, se usaron únicamente los contenidos en el dataframe de *train.csv*. Para la selección de los datos de prueba y los de entrenamiento, se utilizó la herramienta conocida como *train test split*, que se detalla en el marco teórico. Se utilizó un 33 % de los datos como prueba, y un 67 % de los datos fueron entrenados.

Selección de los algoritmos

A partir de los objetivos iniciales y luego de la investigación realizada sobre los métodos a aplicarse, se seleccionaron tres. Estos fueron tomados con base en aspectos como: rendimiento, funcionalidad, adaptabilidad y su reproducción. Dado que los mismos fueron utilizados en investigaciones y proyectos con alcances similares al nuestro, fue una selección rigurosa y curada para concluir en qué se trabajaría.

El primero, Naive Bayes, es un clasificador probabilístico con un fuerte supuesto en la independencia condicional, que es óptimo para las clases de clasificación con una alta cantidad de características dependientes. La adherencia de las clases de sentimientos están dadas por el teorema de Bayes.

Luego, se determinó seguir con un algoritmo de redes neuronales, el modelo long short-term memory (LSTM). Dada su naturaleza proveniente de las redes neuronales, tiene un buen rendimiento en el aprendizaje de secuencias, lo que lo hace atractivo y aplicable en el escenario actual.

Por otro lado, se seleccionó realizar una regresión lineal con un conteo de los vectores con los datos. Este modelo es discriminatorio, lo que implica que parte de una probabilidad condicional. Puesto que es un modelo con un promedio ponderado, dependiendo del peso de cada token o palabra, consideramos que puede llegar a otorgar buenos resultados en el análisis de los textos.

Selección de las herramientas utilizadas

Dada la naturaleza del proyecto, fue necesario escoger librerías a utilizar para varios procesos. Dado que los modelos utilizados pertenecen a librerías utilizadas a lo largo del curso, y son modelos que tocamos en clase, no se ahonda en esto. En cambio, resalta la selección de la librería a utilizarse para generar las visualizaciones dinámicas.

A partir de los objetivos del proyecto, se seleccionaron dos herramientas posibles para su uso: Flask y Plotly. Dado que se ha trabajado con el segundo, se decidió realizar una investigación sobre las ventajas y desventajas que presentan ambas herramientas. Es decir, se buscó la herramienta que mejor se adapte a las características y necesidades del proyecto, puesto que es necesaria la creación de una interfaz visual adaptable a los procesos que se llevarán a cabo.

Resultados y análisis de resultados

De forma inicial se llevó a cabo un proceso exploratorio de los datos, para definir qué estrategias se usarán para el modelo final. A partir del análisis inicial, se observó que el dataset en cuestión cuenta con 8544 datos los cuales se dividen en cinco variables:

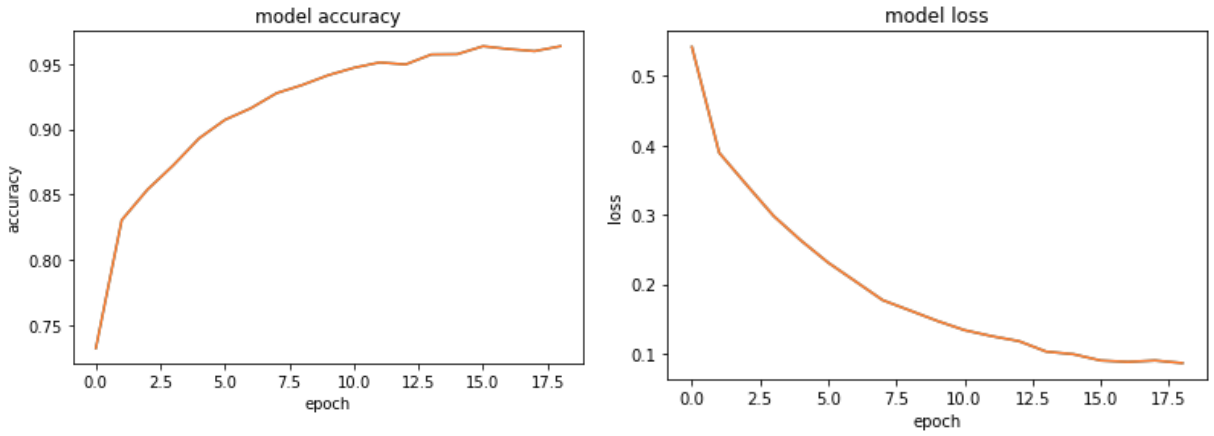
Variable	Tipo	Descripción
id	Numérica	Identificador de cada tweet
keyword	Categórica	Palabra clave del tweet
location	Categórica	Ubicación de donde se envió el tweet
text	Categórica	Texto del tweet
target	Numérica	1 si el tweet es de un caso real, 0 en caso contrario

Se realizó una limpieza de datos, que incluye:

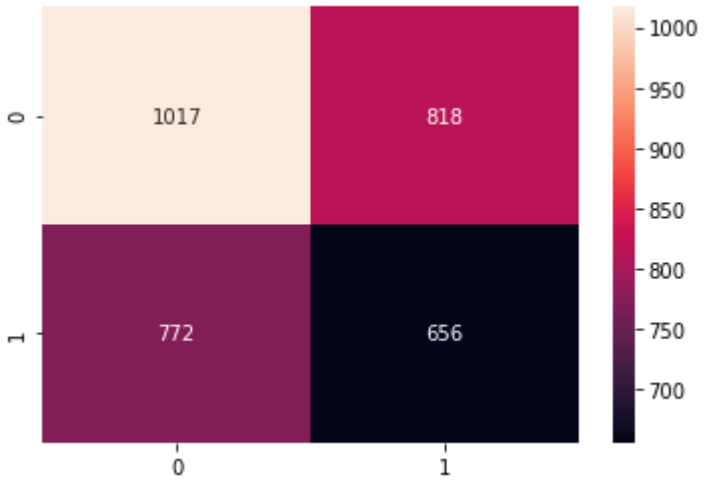
- Convertir todo a minúsculas
- Eliminación de números o caracteres especiales
- Eliminación de referencias a externos, como "@user" o hilos de información
- Eliminación de emojis y signos de puntuación

Posteriormente, se procedió con la aplicación de cada uno de los modelos considerados.

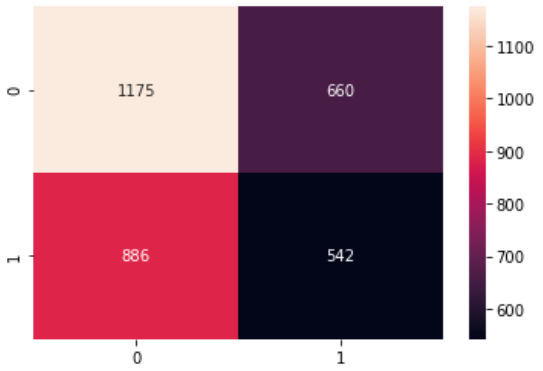
A partir del uso del modelo LSTM, la acertividad del modelo es alta



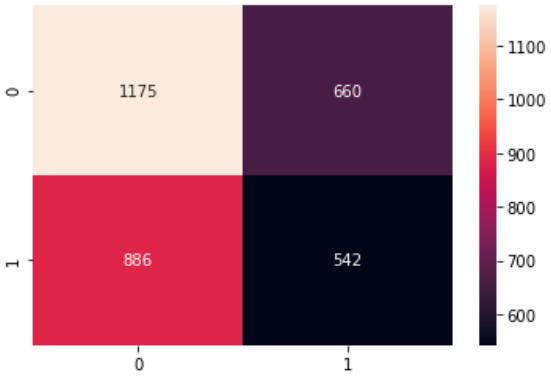
Como se observa en la gráfica, su comportamiento crece de manera exponencial. Si se observa además la pérdida de información, es posible observar que su rendimiento es el ideal para el caso considerado. Si además observamos su desempeño con respecto a la matriz de confusión, es posible concluir que el modelo LSTM es ideal para el escenario.



Ahora bien, la aplicación de la regresión lineal al modelo no resulta tan positiva como se esperaba. Si bien se ve una aceptabilidad del 80 %, la matriz de confusión es un indicador suficiente para señalar que no es la primera opción para el proyecto.



Por otro lado, es importante mencionar que en los procesos de prueba, la probabilidad resultante de la aplicación de este modelo era siempre 0. Esto es un indicador de un sobreentrenamiento en el modelo, por lo que se descartó del producto final.



Por último, la evaluación del desempeño del modelo Naive Bayes probó ser la más efectiva para el escenario en cuestión. Con casi un 80 % de asertividad, es el modelo cuya matriz de confusión probó tener los mejores resultados. A partir de los datos de prueba que se tenían, y los *tweets* extraídos y evaluados de forma manual, los resultados predichos tuvieron un alto rendimiento.

Conclusiones

Se concluye que el modelo con mejor rendimiento es el Naive Bayes, puesto que es el que mejores resultados otorgó.

Se concluye que el modelo menos acertado para la situación es la regresión lineal puesto que tuvo un bajo rendimiento en comparación a los otros modelos.

Se concluye que el proceso previo de limpieza para los datos es crucial en escenarios en los que se trabaje con texto natural.

Bibliografía

Ali Sit, M., Koylu, C. y Demir, I. (2019) Identifying disaster-related tweets and their semantic, spatial and temporal context using deep learning, natural language processing and spatial analysis: a case study of Hurricane Irma, *International Journal of Digital Earth,* 12:11, 1205-1229, DOI: 10.1080/17538947.2018.1563219

Brownlee, J. (2020). *Train-Test Split for Evaluating Machine Learning Algorithms.* Machine Learning Mastery. https://machinelearningmastery.com/train-test-split-for-evaluating-machine-learning-algorithms/

Colah, B. (2015). *Understanding LSTM Networks.* Colah's Blog. http://colah.github.io/posts/2015-08-Understanding-LSTMs/

Gupta, B., Negi, M., Vishwakarma, K., Rawat, G., Badhani, P., & Tech, B. (2017). Study of Twitter sentiment analysis using machine learning algorithms on Python. International Journal of Computer Applications, 165(9), 29-34.

IBM Cloud Education. (2020). *Natural Language Processing (NLP).* IBM Cloud Learn Hub. Recuperado de: https://www.ibm.com/cloud/learn/natural-language-processing

Periñán-Pascual, C., & Arcas-Túnez, F. (2018). *The Analysis of Tweets to Detect Natural Hazards.* In Intelligent Environments 2018 (pp. 87-96). Ios Press.