Revista Innovación y Software
Vol. 4, No. 1, Mes Marzo-Agosto, 2023
ISSN: 2708-0935
Pág. 198-203
https://revistas.ulasalle.edu.pe/innosoft

Tipo de artículo: Artículos originales
Temática: Inteligencia Artificial
Recibido: 25/02/2023 | Aceptado: 23/03/2023 | Publicado: 30/03/2023

Identificadores persistentes:
ARK: ark:/42411/s11/a88
PURL: 42411/s11/a88

Clasificación de texto con NLP en tweets relacionados con desastres naturales

NLP text classification in tweets related to natural disasters

Patrik Renee Quenta Nina [1][0000-0002-6184-1378], Frank Berly Quispe Cahuana [2][0000-0001-5584-1593]

[1] Universidad Nacional de San Agustín. Arequipa, Perú. pquenta@unsa.edu.pe
[2] Universidad Nacional de San Agustín. Arequipa, Perú. fquispecah@unsa.edu.pe

[*] Autor para correspondencia: pquenta@unsa.edu.pe

Resumen

Actualmente existe una gran cantidad de información circula a través de las redes sociales, esta no siempre tiende a ser verídica y tratándose de desastres naturales su falsedad podría llegar a tener bastante consecuencias como histeria colectiva en la población. Para evitar esto se propuso un análisis eficiente para la comprobación de tweets con información falsa utilizando algoritmos de procesamiento de lenguaje natural.

Palabras clave: Desastres naturales, NLP , sentimientos, Twitter.

Abstract

Currently there is a large amount of information circulating through social networks, this does not always tend to be true and in the case of natural disasters its falsity could have quite consequences such as mass hysteria in the population. To avoid this, an efficient analysis was proposed to check tweets with false information using natural language processing algorithms.

Keywords: *Feelings, natural disasters, NLP, Twitter.*

Introducción

Las plataformas de redes sociales como Facebook y Twitter se han convertido en herramientas de comunicación predominantes en la sociedad moderna. Estas plataformas proporcionan un mecanismo para

recopilar datos dinámicos sobre el comportamiento y el sentimiento humanos [1]. Estudios recientes consideran el uso de las redes sociales durante los desastres naturales, estudiando principalmente el estado de ánimo de la población o las diversas reacciones del público durante un incidente específico [1]. Sin embargo es posible captar datos incorrectos debido a que las personas generan percepciones erróneas sobre los peligros y que su conciencia situacional general se vea equivocada también cuando se exponen a información falsa o engañosa [1].

Debido al creciente uso de las redes sociales, la vulnerabilidad de las personas a los rumores falsos está determinada cada vez más por nuevas formas de verificación colectiva de los hechos y de dar sentido a los riesgos y desastres [2]. La existencia de diversas fuentes oficiales y no oficiales hacen que información errónea llegue a las personas durante una crisis. Lo cual puede hacer que juzgar la relevancia y la credibilidad de la información recibida sea una tarea difícil, por lo tanto, no pueden tomar las medidas de protección adecuadas [2]. Ya se han visto trabajos relacionados como AI-SocialDisaster, este es un sistema de apoyo a la toma de decisiones para identificar y analizar desastres naturales como terremotos, inundaciones e incendios forestales utilizando fuentes de redes sociales [3]. Con este software, los estrategas y planificadores de desastres pueden comprender las características de un desastre en un área en particular, además de obtener datos como análisis de sentimientos [3]. Debido a la disponibilidad omnipresente de datos en tiempo real, muchas agencias de rescate monitorean estos datos regularmente para identificar desastres, reducir riesgos y salvar vidas [4]. Sin embargo, es imposible que los humanos verifiquen manualmente la gran cantidad de datos e identifiquen desastres en tiempo real [4]. Con este propósito, se han propuesto muchas investigaciones para presentar palabras en representaciones comprensibles por máquina y aplicar métodos de aprendizaje automático en las representaciones de palabras para identificar el sentimiento de un texto [4]. Un claro ejemplo de la propagación de información falsa es “El caso del tifón Mangkhut en China”. En este caso se realizaron simulaciones de los escenarios reales, de aislamiento y de empotramiento. En esta simulación se probó la relevancia de la Primera Ley de Tobler en las redes sociales [5]. De tres escenarios, un escenario real, un escenario de aislamiento y un escenario de incrustación, se probó que la estrategia de incrustación controlaba mejor la transmisión de información falsa [5] de esta manera se plantearon sugerencias prácticas a la hora de filtrar información falsa en un desastre natural. Basándonos en estos casos podemos afirmar que enseñar a las computadoras cómo entender y hablar lenguajes naturales ofrece una gran cantidad de beneficios

[6]. El subcampo de la IA que hace que las computadoras parezcan inteligentes para comprender y generar lenguajes como los humanos se llama procesamiento de lenguaje natural [6]. NLP se centra en la traducción de lenguaje natural, recuperación de información, extracción de información, resumen de texto, respuesta a preguntas,modelado de temas, y el reciente sobre minería de opiniones [7]. Para resolver el problema de la información falsa generada a través de las redes sociales, nos apoyaremos en las incrustaciones y modelos pre-entrenados, debido a que estos proporcionan una visión de las estrategias fundamentales a la hora de recopilar información [8]. Teniendo en cuenta que Twitter se ha convertido en un importante canal de comunicación en tiempos de emergencia y, la ubicuidad de los teléfonos inteligentes que permite a las personas anunciar una emergencia que están observando en tiempo real. El objetivo del presente trabajo es desarrollar un modelo de deep learning capaz de clasificar si un tweet sobre un desastre natural es real o falso, haciendo uso de algoritmos de procesamiento de lenguaje natural (NLP).

Fundamentación Teórica

Redes neuronales artificiales

Una red neuronal artificial consiste en una red de unidades simples de procesamiento de información, llamadas neuronas. El poder de las redes neuronales para modelar relaciones complejas no es el resultado de modelos matemáticos complejos, sino que surge de las interacciones entre un gran conjunto de neuronas simples. Es normal pensar en las neuronas como un trabajo organizado en capas. [Deep learning]

Procesamiento de Lenguaje Natural (NLP)

El procesamiento del lenguaje natural (NLP) es una técnica de inteligencia artificial que se ocupa de la comprensión del lenguaje humano. El cual implica técnicas de programación para crear un modelo que pueda comprender el lenguaje, clasificar el contenido e incluso generar y crear nuevas composiciones de lenguaje humano. [ML for coders]

Clasificación de texto (Análisis de sentimientos)

El análisis de sentimientos, también conocido como minería de opiniones, se ocupa de inspeccionar estos sentimientos dirigidos hacia cualquier entidad. Liu (2010) utilizó el término objeto para representar la entidad

objetivo mencionada en el texto. Un objeto está constituido por componentes y algún conjunto de atributos. Por ejemplo, se considera la siguiente oración "la pantalla de la computadora portátil está dañada y la duración de la batería es terrible". El objeto aquí es una computadora portátil que tiene pantalla y batería como componentes. La calidad de visualización es un atributo de la pantalla y la duración de la batería es el atributo de la batería. Este texto se puede clasificar en una opinión positiva, negativa o neutral. []

Materiales

- **Google Drive**

  Google Drive proporciona una solución de almacenamiento basada en la nube para archivos de Google Workspace y otros datos de usuario.[]
- **Google Colaboratory**

  Permite escribir y ejecutar Python en el navegador, no se requiere configuración, acceso a las GPU sin cargo y es fácil de compartir
- **Python**
- **Matplotlib**
- **Numpy**
- **Pandas**
- **nltk (Natural Language Toolkit)**
- **Pytorch**
- **TensorFlor**

Metodología

Dataset

El dataset se obtuvo de kaggle el cual contiene los siguientes datos, cada muestra en los datos de entrenamiento y los datos de prueba tiene la siguiente información:

- El “text” de un tweet
- Una “keyword” de ese tweet (¡aunque esto puede estar en blanco!)

- La “location” desde la que se envió el tweet (también puede estar en blanco):

Análisis exploratorio de datos

Consiste en analizar e investigar conjuntos de datos y resumir sus principales características, a menudo empleando métodos de visualización de datos. el cual ayuda a determinar la mejor manera de manipular las fuentes de datos para obtener las respuestas que se necesitan.

Limpieza de datos (Data cleaning)

Es el proceso de detectar, corregir o eliminar registros corruptos o imprecisos de un conjunto de registros, donde pueden ser tablas o bases de datos con información incorrecta, incompleta, mal formateada o duplicada.

Modelado

Para la solución al problema se propuso usar una Red Neuronal Recurrente, debido a que su arquitectura lo permite para el manejo de datos secuenciales como lo es el texto.

Conclusiones

El análisis de los datos que nos brindan las redes sociales es importante debido a que con estos se podría evitar casos de histeria colectiva como se han mencionado anteriormente.

Para evitar la gran acumulación de datos innecesarios que toda red social tiene se destacó la parte de la limpieza de estos datos, de esta manera se podría hacer un mejor y más efectivo análisis con las metodologías de procesamiento del lenguaje natural.

Referencias

[1] S. K. Theja Bhavaraju, C. Beyney y C. Nicholson, "Quantitative analysis of social media sensitivity to natural disasters", International Journal of Disaster Risk Reduction, vol. 39, p. 101251, octubre de 2019. [En línea]. Disponible: https://doi.org/10.1016/j.ijdrr.2019.101251

[2] S. Hansson et al., "Communication-related vulnerability to disasters: A heuristic framework", International Journal of Disaster Risk Reduction, vol. 51, p. 101931, diciembre de 2020. [En línea]. Disponible: https://doi.org/10.1016/j.ijdrr.2020.101931

[3] F. K. Sufi, "AI-SocialDisaster: An AI-based software for identifying and analyzing natural disasters from social media", Software Impacts, p. 100319, mayo de 2022. [En línea]. Disponible: https://doi.org/10.1016/j.simpa.2022.100319

[4] S. Deb y A. K. Chanda, "Comparative analysis of contextual and context-free embeddings in disaster prediction from Twitter data", Machine Learning With Applications, vol. 7, p. 100253, marzo de 2022. [En línea]. Disponible: https://doi.org/10.1016/j.mlwa.2022.100253

[5] Y. Lian, Y. Liu y X. Dong, "Strategies for controlling false online information during natural disasters: The case of Typhoon Mangkhut in China", Technology in Society, vol. 62, p. 101265, agosto de 2020. [En línea]. Disponible: https://doi.org/10.1016/j.techsoc.2020.101265

[6] Raina, V., Krishnamurthy, S., “Natural Language Processing”. In: Building an Effective Data Science Practice.Apress, Berkeley, CA, diciembre de 2021 Disponible. https://doi.org/10.1007/978-1-4842-7419-4_6

[7] K. R. Chowdhary, Fundamentals of Artificial Intelligence. New Delhi: Springer India, 2020.[En línea]. Disponible: https://doi.org/10.1007/978-81-322-3972-7

[8] J. K. Tripathy et al., "Comprehensive analysis of embeddings and pre-training in NLP", Computer Science Review, vol. 42, p. 100433, noviembre de 2021. [En línea]. Disponible: https://doi.org/10.1016/j.cosrev.2021.100433

[9] Kelleher, J. D. (2019). Deep Learning. MIT Press.

[10] Yadav, A. y Vishwakarma, D. K. (2019). Sentiment analysis using deep learning architectures: a review. Artificial Intelligence Review, 53(6), 4335–4385. https://doi.org/10.1007/s10462-019-09794-5

Roles de Autoría

Patrik Renee Quenta Nina: Conceptualización, Curación de datos, Análisis formal, Investigación, Metodología, Software, Validación, Redacción - borrador original. F**rank Berly Quispe Cahuana:** Conceptualización, Curación de datos, Análisis formal, Investigación, Metodología, Software, Validación, Redacción - borrador original.