TECNOLÓGICO NACIONAL DE MÉXICO

Categorización de textos por tema por medio del modelo codificador BERT

Materia:

Tópicos Avanzados de IA

Estudiantes:

Gómez Lara Joshua Israel

García Ríos Sebastián

Horario:

10:00 – 11:00

Docente:

Zuriel Dathan Mora Félix

Resumen

Al momento de recibir noticias o actualizaciones, ya sea por correo, SMS, Messenger u otros mensajeros de texto, podemos contar con información de un millar de temas: Noticias locales, celebridades, deportes, familia, trabajo, esto por nombrar unos cuantos. Una persona puede a veces determinar el contenido de un mensaje a través de distintas características, como los fragmentos en las notificaciones, el nombre del remitente, etc. Sin embargo, a veces simplemente no es posible juzgar el tema del mensaje a primera vista solo con esa información.

A través de algoritmos de categorización de texto, podemos juzgar las distintas cualidades del mensaje en unos instantes: de que tema escribe, si es un mensaje genuino o de spam, o hasta determinar el humor del mensaje para saber si es positivo o negativo.

Introducción

En la era de la mensajería instantánea, el usuario promedio recibe información de innumerables fuentes durante todo momento desde su teléfono, ya sea de redes sociales, correos, mensajes privados, entre un amplio etcétera. Se dice que este interminable flujo de información puede llegar a ser hasta contra productivo para captar la atención de un individuo, que puede aprender a ignorar cualquier intento de capturar su atención entre.

Por esto, empresas en el ámbito de la información como Facebook y Google, han desarrollado distintos algoritmos para generar un *Feed* hecho a la medida para el usuario, basándose en datos como páginas que ha visitado o temas populares en la región donde vive o entre sus demográficas, como edad o género. Estos algoritmos se denominan transformadores de tipo codificador, ya que codifican una entrada de texto natural – una publicación en una red social, por ejemplo – para convertirlas en representaciones vectoriales, que posteriormente procesan el contexto de cada una de estas representaciones.

El modelo BERT, o Representación de Codificador Bidireccional de Transformadores, fue presentado por Google a finales del 2018 como una mejora en las sugerencias de búsqueda y tuvo un gran efecto en la industria de la información, al ser uno de los más grandes avances en los modelos de lenguaje de la época con su capacidad de evaluar el contexto de las palabras vectorizadas. Desde entonces, este se ha utilizado para distintas aplicaciones de predicción y evaluación de texto, particularmente, evaluación de publicaciones en redes sociales.

Con nuestro proyecto, buscamos aplicar este algoritmo para realizar esa misma evaluación y determinar el tema de una publicación.

Marco Teórico

- Anderson, D. (5 de Noviembre del 2019) *Una investigación de BERT: Como BERT llego en cohete al procesamiento del lenguaje natural.* Search Engine Land. https://searchengineland.com/google-bert-used-on-almost-every-english-query-342193

  BERT se describe como una estructura pre-entrenada de Deep Learning para lenguaje natural, que ha logrado excelentes resultados en distintas tareas de proceso del lenguaje natural, como análisis de emociones, determinación de personas nombradas, predicción de oraciones, etc.

- Nayak, P. (25 de Octubre del 2019) *Entendiendo Búsquedas Mejor que Nunca Ántes.* Google's The Keyword. https://blog.google/products/search/search-language-understanding-bert/

  Al aplicar BERT en los rankings de búsqueda y fragmentos destacados, fue más fácil encontrar información útil, particularmente para búsquedas más largas o aquellas con preposiciones como "Quién" o "Porqué" son importantes en el contexto, permitiendo búsquedas más facilidad al buscar.

Objetivos

Objetivo General

Desarrollar un sistema de categorización automática de textos basado en el modelo codificador BERT, con el propósito de identificar el tema principal de cada mensaje de manera eficiente y precisa.

Objetivos Específicos

1. Implementar el modelo BERT para analizar y clasificar textos según su contenido temático.
2. Evaluar el desempeño del modelo en la categorización de textos mediante métricas de precisión, recall y F1-score.
3. Comparar los resultados obtenidos con otros modelos tradicionales de clasificación de texto.
4. Diseñar un conjunto de datos de prueba para evaluar la efectividad del sistema en distintos tipos de mensajes.
5. Analizar las ventajas y limitaciones del modelo en la clasificación automática de texto.