UNIVERSIDAD NACIONAL DE ASUNCIÓN
Facultad Politécnica

Proyecto Final
Diplomado en Inteligencia Artificial Aplicada a Productos y Servicios

---

INFORMACIÓN DEL PROYECTO

Título del proyecto: Analisis de sentimientos para tweets mediante el modelo roBERTa
Nombres de los integrantes: Carlos Velazquez
Modalidad del proyecto: Factibilidad Técnica

1. Problemática

1.1. Problema Práctico

Una empresa cuenta con una fuerte presencia en las redes sociales. Esto se utiliza mayormente para promocionar sus productos y servicios a través de campañas de marketing.

Las campañas de marketing son realizadas por la empresa misma pero también puede contratar influencers para difundir su mensaje. Entonces la empresa desea saber qué tan efectiva es la campaña mediante una red en particular.

Medir correctamente qué tipo de reacciones genera cada campaña podría servir enormemente a la empresa para tomar decisiones sobre cómo poder mejorar las campañas o en qué productos o servicios enfocarse.

1.2. Descripción del producto

Se desea crear una aplicación capaz de obtener publicaciones de la plataforma Twitter y clasificarlas en sentimientos positivos, neutros o negativos.

Estas publicaciones están agrupadas por campañas y se utilizaràn los hashtags para verificar a qué campaña pertenece la publicación.

Una vez clasificadas las publicaciones se podrán generar sumatorias de los sentimientos para determinar en promedio que tipos de reacciones generan las campañas para determinar su éxito.

1.3. Problema Técnico

Realizar análisis de sentimientos sobre tweets puede ser un proceso difícil, ya que son diferentes de otros datos de tipo texto. Esto se debe a que los tweets se escriben normalmente con un lenguaje conversacional y también suelen ser cortos.

Para superar estos problemas, la aplicación realizará la clasificación de las publicaciones mediante el algoritmo llamado BERT[3] desarrollado por Google. Específicamente un algoritmo preentreado y optimizado por el equipo de Facebook AI llamado roBERTa[1].

El usuario podrá especificar el hashtag con el que se buscarán los tweets que la aplicación se encargará de buscar utilizando la API de Twitter.

Una vez obtenido el listado de tweets, la aplicación ya se encarga de clasificar el sentimiento en positivo, neutro o negativo.

El modelo roBERTa[2] ya fue entrenado utilizando el corpus de wikipedia en inglés (16GB) y también fue entrenado con tweets en inglés.

El modelo BERT es un método de redes neuronales que utiliza transformadores. Ya que las redes neuronales puede ser intensas computacionalmentes, idealmente se debe de tener una tarjeta gráfico. De todas maneras como es un algoritmo preentranado y optimizado se puede correr la aplicación utilizando solo la CPU:

Otro requerimiento es contar con una cuenta de desarrollador en la plataforma Twitter. No es necesario para correr el modelo pero es importante para obtener los datos a clasificar.

2. Metodología

2.1. Obtención de datos

Para lograr obtener los datos de Twitter es neceserio autenticarse mediante la API de la plataforma. Para este paso es necesario contar con una cuenta de desarrollador de Twitter y crear una nueva aplicación. Una vez obtenidas las credeciales, la búsqueda de datos se realiza mediante una llamada get a la API de Twiiter[5]. Aquí es donde se indica con que criterios se busca los tweets.

2.2. Preprosecamiento de datos

Para utilizar el modelo roBERTa es necesario realizar una transformación de datos sobre los tweets. Primero los usuarios mencionados son reemplozados por la palabra @user y todos los enlaces son reemplazados pot la palabra http. También es posible remover las palabras que no influyen a la hora de realizar el análisis de sentimientos. Esto puede descargarse desde la misma página donde está disponible el modelo. Por último vale la pena indicar que los tweets deben estar en inglés ya que el modelo viene preentrenado con esta restricción.

2.3. Descarga del modelo preentrenado

La descarga se realiza desde el sitio web Hugging Face[**?**] donde se aloja una implementación opensource del proyecto. Este modelo cuenta la ventaja de estar preentrenado por lo que no es necesario otro proceso para realizar el análisis.

Tabla 1: Métricas obtenidas utilizando el ejemplo para COVID19 (50 items).

Sentimiento	NEGATIVE ,	NEUTRAL.	POSITIVO	OVERALL
NEGATIVE	4	2	0	-
NEUTRAL	1	3	1	-
POSITIVO	0	3	1	-
Precision	-	-	-	0.56
Recall	-	-	-	0.51
Accuracy	-	-	-	0.53

2.4. Evauluación del modelo

Para evaluar los resultados del modelo, se utilizará una matriz de confusión donde se compararán los datos predichos por el programa con un ejemplo de tweets clasificados dentro del mismo curso en la sección de análisis de sentimientos.

3. Evaluación de factibilidad

3.1. Resultados

Utilizando de ejemplo los tweets de ejemplo para COVID19, se tomarón 100 items y se compararón con los valores predichos Como se puede apreciar en la tabla 1, las métricas arrojadas por el modelo son bastante fiables.

3.2. Recomendaciones

- Interfaz gráfica: Una interfaz gráfica mejorada supondría una gran ventaja a la hora de disponibilizar el producto de manera masiva. Una entrada de datos sencilla y resultados en una vista amigable sería muy útil para el usuario final.
- Creación de aplicación como API: Si el servicio puede ser utilizado como una API REST por diferentes aplicaciones tammbién sería de gran ayuda para la disponibilidad y usabilidad final.
- Despliegue de resultados en formatos gráficos o en archivos exportables: Un mejor número de gráficos y capacidad de exportar los resultados a excel o csv podrían ser de ayuda para el usuario.
- Aumento en el número de tweets analizados para determianr métricas: el número de tweets evaluados es aún pequeño por lo que mejorar este aspecto podría mejorar las métricas y

ayudar a captar algún error.

- Idioma nativo al momento de hacer pruebas: Como el conjuento de tweets de pruebas fue traducido del español, se pueden generar bias sobre el resultado final.

3.3. Conclusión

De acuerdo a los resultados obtenidos el modelo parece ser bastante viable, pero se puede mejorar con las recomendaciones. También como es opensource, se facilita mucho la implementación y posible usos comerciales. Una mayor dispobilidad y usabilidad podría hacer de la aplicación un herramienta importante en empresas que desarollen productos nuevos, realizen campañas de marketing o quieran mejorar su imagen con respecto a sus clientes.

Bibliografía

[1] Roberta: An optimized method for pretraining self-supervised NLP systems. Meta AI. https://ai.facebook.com/blog/roberta-an-optimized-method-for-pretraining-self-supervised-nlp-syste Último acceso Enero 11, 2023

[2] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V. (n.d.). Roberta: A robustly optimized Bert pretraining approach - arxiv. RoBERTa: A Robustly Optimized BERT Pretraining Approach. https://arxiv.org/pdf/1907.11692.pdf. Último acceso. Enero 11, 2023

[3] Devlin, J., Chang, M.-W., Lee, K., Toutanova, K. (2019, May 24). Bert: Pre-training of deep bidirectional Transformers for language understanding. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding - arXiv.org. https://arxiv.org/abs/1810.04805. Último acceso Enero 11, 2023

[4] Devlin, J., Chang, M.-W., Lee, K., Toutanova, K. (2019, May 24). Bert: Pre-training of deep bidirectional Transformers for language understanding. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding - arXiv.org. https://arxiv.org/abs/1810.04805. Último acceso Enero 11, 2023

[5] Search Tweets: Standard v1.1 https://developer.twitter.com/en/docs/twitter-api/v1/tweets/search/api-reference/get-search-tweets. Último acceso Enero 11, 2023

[6] Hugging Face https://huggingface.co/cardiffnlp/twitter-roberta-base-sentiment. Último acceso Enero 11, 2023