Análisis de sentimiento a partir de reseñas de películas utilizando el modelo preentrenado BERT

Catriel Bartezaghi, Loreficcio Juan Mauro.

Deep Learning 2023, Ingeniería en Informática, FICH-UNL

Tutor: Leandro Bugnon

***Resumen*—**
En este trabajo, se aborda el análisis de sentimientos en el procesamiento del lenguaje natural utilizando el modelo preentrenado BERT (Bidirectional Encoder Representations from Transformers). BERT se utiliza para obtener representaciones contextualizadas de las palabras. En este estudio, se exploró la integración de BERT con diferentes capas adicionales, incluyendo una LSTM, una GRU y una capa Lineal, para realizar la clasificación de sentimientos y evaluar los resultados de cada caso específico.

***Palabras claves*—Sentimiento, Texto, Transformers, BERT.**

I. Introducción

El análisis de sentimientos es un campo relativamente joven en el procesamiento del lenguaje natural, pero ha experimentado un rápido crecimiento y evolución en las últimas décadas. En sus inicios alrededor de los 2000, se utilizaron enfoques basados en reglas y léxicos que empleaban listas de palabras con etiquetas de sentimiento y se aplicaban reglas predefinidas para determinar el sentimiento general del texto. Pero no fue hasta 2013 con la introducción de *Word2Vec* [1], que se logró realizar un análisis semántico y léxico entre palabras, a partir de una representación de las palabras en un espacio vectorial. Más adelante en 2017 se introdujeron las redes *Transformers* [2], las cuales a diferencia de Word2Vec, no solo consideran el contexto inmediato de una palabra, sino también el contexto global del texto. Esto permite capturar mejor las dependencias a largo plazo y las relaciones entre las palabras.

El análisis sensitivo a partir de un texto puede parecer un tema relativamente fácil de abordar, ya que el estudio de cada palabra individual puede proporcionar una idea clara sobre el sentimiento expresado. Sin embargo, cuando se considera en un contexto más amplio y completo, esta tarea se vuelve mucho más desafiante, especialmente cuando se trata de detectar la ironía en el texto y más aún si el origen del dato proviene del ciberespacio, las redes sociales y foros donde según el ámbito esta cualidad es predominante.

Actualmente los transformers son considerados una de las arquitecturas más efectivas para el análisis de sentimientos, razón por la cual se seleccionaron para este trabajo. Bajo esta línea de aplicación se propone hacer uso de dicha arquitectura para discernir la valoración de una película a partir de las críticas realizadas por la audiencia, haciendo uso del modelo pre-entrenado BERT [3] y proponiendo distintos modelos de fine-tuning para comparar los distintos resultados.

II. Propuesta

Para el presente trabajo se propone abordar un problema de clasificación, más precisamente determinar la valoración de una película a partir de los comentarios críticos de la audiencia plasmados en un reconocido sitio web de reseñas.

Haciendo uso del modelo pre-entrenado (con datos de Wikipedia y BooksCorpus) BERT, se proponen tres tipos de fine-tuning con distintas redes neuronales de salida y evaluar cuál de ellos es el más óptimo.

Para el estudio, se anexará a la salida del modelo BERT distintos casos, una red LSTM (Long Short Term Memory), otra red neuronal recurrente del tipo GRU (Gated-Recurrent-Units) y finalmente una Lineal básica.

III. Dataset

El dataset utilizado en este trabajo consiste en un conjunto de reseñas de películas del sitio web *IMDB* [4]. Los datos fueron obtenidos a partir de la librería *Torchtext* [5] para pytorch. Consiste en reseñas en inglés de películas y las etiquetas de sentimiento asociadas, que representan la reseña otorgadas a cada una. De esta manera se obtiene un conjunto listo para realizar un entrenamiento supervisado. Donde, el comentario de la reseña es el texto a analizar y la puntuación (positiva o negativa) es la etiqueta.

El dataset cuenta con un total de 50.000 reseñas etiquetadas de entre 5 y 2789 palabras, de las cuales la mitad son reseñas positivas y la otra mitad negativas. Se utilizó un 10% del dataset total. Se separaron 1750 para entrenamiento del modelo, 750 para validación y las restantes 2500 para testing.

IV. Modelo

Los tres modelos propuestos toman como base el modelo BERT, luego para cada uno de los casos, se agrega a la salida del modelo otra red de las ya mencionadas para realizar la tarea de clasificación de sentimiento. A continuación se detallan brevemente el funcionamiento del modelo principal y se mencionan los fine-tunings realizados.

A. BERT

BERT (Bidirectional Encoder Representations from Transformers) es un modelo de aprendizaje profundo que

ha revolucionado el procesamiento del lenguaje natural (NLP). A diferencia de los enfoques tradicionales, BERT aprovecha las capacidades de los Transformers y el entrenamiento bidireccional para capturar el contexto contextual de las palabras y mejorar significativamente el rendimiento en diversas tareas de NLP.

El modelo BERT consta de dos etapas principales: pre-entrenamiento y ajuste fino. Durante el pre-entrenamiento, BERT se entrena con grandes cantidades de texto no etiquetado utilizando tareas auxiliares, como la predicción de palabras enmascaradas (*Figura 1)* y la predicción de la siguiente oración (*Figura 2)*. Esto permite que BERT capture representaciones contextualizadas de las palabras.

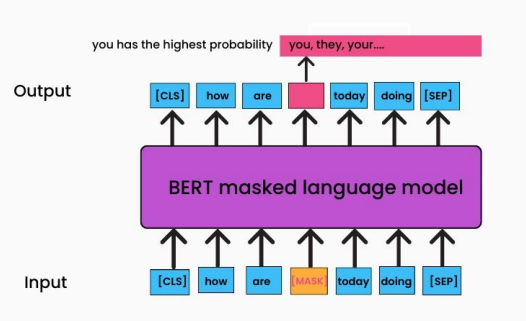

Figura 1: Pre-entrenamiento automático por enmascaramiento

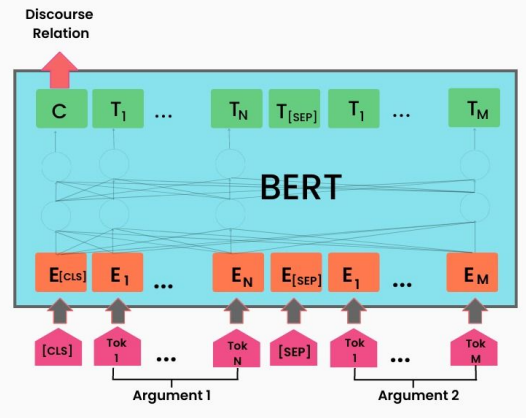

Figura 2: Pre-entrenamiento automático por predicción

En la etapa de ajuste fino, BERT se adapta a tareas específicas de NLP mediante el uso de datos etiquetados. Los parámetros pre-entrenados de BERT se inicializan en el modelo y luego se ajustan utilizando técnicas de aprendizaje supervisado. Cada tarea específica tiene su propio modelo de ajuste fino, aunque comparten los mismos parámetros pre-entrenados.

Una característica destacada de BERT es su arquitectura unificada, que es aplicable a una amplia gama de tareas de NLP. Esto se logra mediante el uso de atención multi-cabezal y codificación bidireccional en los Transformers. Además, BERT logra un rendimiento sobresaliente al capturar el contexto, permitiendo una comprensión más profunda del lenguaje. Este contexto es adquirido a medida que la información se propaga a través de las capas de atención y la representación del token [CLS] se actualiza y evoluciona para capturar información contextualizada de toda la secuencia. Cada capa del modelo ajusta la representación del token [CLS] en función de las interacciones y las relaciones entre los tokens vecinos.

En términos de tamaño del modelo, BERT ofrece dos variantes principales: BERTBASE y BERTLARGE. BERTBASE tiene 12 capas de Transformers, una dimensión oculta de 768 y 12 bloques de atención, mientras que BERTLARGE tiene 24 capas de Transformers, una dimensión oculta de 1024 y 16 bloques de atención.
Para el presente trabajo se seleccionó el modelo BERTBASE, ya que tiene un costo computacional menor y precisión suficiente para la tarea de análisis de sentimiento de breves cadenas de texto.

Si bien el modelo BERT es capaz de capturar información contextualizada de las palabras en un texto, agregar capas adicionales puede ser útil en el análisis de cadenas de textos para capturar más información contextual y secuencial . Esto se debe a que el orden de las palabras es relevante para determinar el sentimiento del texto. Es por ello que se detallan a continuación las diferentes arquitecturas utilizadas.

B. BERTBASE + LSTM

En el caso del modelo BERTBASE + LSTM, se utiliza el modelo BERTBASE como base y se agrega una capa adicional de Long Short-Term Memory (LSTM) a la salida de BERT.

El LSTM es una arquitectura de red neuronal recurrente que permite capturar dependencias a largo plazo en secuencias de datos. Al añadir el LSTM al modelo BERTBASE, se busca aprovechar las representaciones contextuales de las palabras generadas por BERT y, a su vez, capturar patrones de secuencia y dependencias a largo plazo en los datos.

De esta manera, cada secuencia de texto es procesada a través del modelo BERT, el cual da por salida una serie de tokens correspondiente a cada subpalabra. Este conjunto de tokens son procesados de manera secuencial por la red LSTM y cada token va actualizando el estado oculto de la red. Una vez que todos los tokens han sido procesados, se obtiene la última representación oculta la cual es pasada a una capa lineal para finalmente realizar la clasificación de sentimiento.

C. BERTBASE + GRU

Similar al caso anterior, en lugar de procesar las salidas del modelo BERT con una red LSTM se incorpora una GRU (Gated recurrent unit).

Se puede decir que la GRU es una versión simplificada de la LSTM ya que tiene una estructura más simple con menos puertas y parámetros. Por lo tanto tiene una capacidad de almacenamiento más limitada. En este caso, debido a que no se trabaja con cadenas largas de texto, esto no es un problema ya que no hay una gran dependencia a largo plazo.

El procesamiento de los datos es el mismo que el desarrollado en el modelo BERTBASE + LSTM.

D. BERTBASE CLS + Lineal

El último caso planteado consiste en utilizar únicamente el token CLS en lugar de todos los tokens correspondientes a cada subpalabra de la cadena de texto. El token CLS en el modelo BERT es un token especial que se agrega al inicio de cada cadena de texto. Al ser pasado por el modelo BERT, este token obtiene una representación contextualizada que captura información sobre toda la secuencia de tokens en la cadena de entrada.
De esta manera, en lugar de utilizar los tokens correspondientes a cada subpalabra de la cadena de texto, solo se utiliza el token CLS el cual contiene información de todos los demás tokens. Éste es pasado a través de una capa lineal para finalmente obtener una salida numérica entre 0 y 1, que representa el sentimiento de la oración.

V. Resultados

En esta sección se presentan los resultados obtenidos con los tres modelos. Para ello se trabajaron con las definiciones de la tabla 1 y se obtuvo un tiempo aproximado de 50 minutos de entrenamiento para cada uno de los modelos. Las pruebas fueron realizadas utilizando los recursos virtuales de Google Colab.

Parámetro	Valor
Épocas	30
Fusión de pérdida	Binary Cross Entropy with Logits Loss
Optimizador	Adam
HIDDEN_DIM	16
OUTPUT_DIM	1
DROPOUT	0.25

Tabla 1: Parámetros de evaluación

Para cada una de las evaluaciones se midió el Accuracy obteniendo como resultados 79,62%, 76,60% y 76,30% para los casos de LSTM, GRU y LINEAL respectivamente. Obteniendo los mejores resultados con un modelo para fine tuning con una LSTM, con un accuracy cercano al 80%.

Con estos resultados, se realizaron distintas pruebas de performance analizando el resultado según una mirada personal. A modo de ejemplo se ingresaron al modelo dos frases propuestas por el equipo y se analizó si los resultados obtenidos eran razonables.

i. "*Terrible movie! The actors seem to be exceptional but they are horrible! Really s…t movie*"

ii. "*The most peaceful movie I watched in my life. Really interesting the life of the character is so hard and she fight to get the life who want*"

La salidas obtenidas fueron de 0.02 y 0.94 para los casos i y ii respectivamente, denotando un funcionamiento más que adecuado para el tono emocional de las frases planteadas.

VI. Conclusiones

Si bien, aún es necesario realizar más pruebas, se puede notar una gran eficiencia en la clasificación de los sentimientos a partir de las reviews de películas con una muestra de tan solo el 10% del dataset y un número relativamente bajo de épocas.
Analizando cada arquitectura se puede observar un mejor rendimiento en las redes recurrentes, donde la LSTM obtiene un mejor accuracy que la GRU debido a su mayor cantidad de parámetros. En cuanto al modelo que contiene una red lineal, a pesar de que solo se utiliza el token CLS y que la arquitectura es mucho más sencilla, se puede notar un muy buen rendimiento, aunque significativamente menor que en las redes recurrentes. Esto destaca la gran capacidad del modelo BERT para capturar y representar información semántica de la oración completa en el token CLS.

Como trabajo futuro, se puede considerar analizar otros modelos transformers más pequeños como DistilBERT, para evaluar si se pueden obtener resultados similares de una manera más óptima.
Por otro lado, también se puede intentar utilizar modelos más complejos que admitan un número mayor de caracteres para realizar análisis de noticias o informes en lugar de sentencias de texto breves. Esto permitiría abordar tareas de análisis de textos más extensos y complejos, ampliando el alcance y las aplicaciones del modelo.

Agradecimientos

El equipo agradece la asistencia, sugerencias y predisposición por parte de los docentes de la cátedra de Deep Learning de la Facultad de Ingeniería y Ciencias Hídricas de la Universidad Nacional del Litoral, en especial a Leandro Bugnon, por haber ocupado el lugar de tutor del grupo.

Referencias

[1] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention Is All You Need. In Advances in Neural Information Processing Systems (pp. 5998-6008). https://arxiv.org/abs/1706.03762

[2] Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. https://arxiv.org/abs/1301.3781

[3] Devlin, J. et al. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics. https://arxiv.org/abs/1810.04805

[4] IMDb. (2023). IMDb: Ratings, Reviews, and Where to Watch the Best Movies & TV Shows. https://www.imdb.com/

[5] PyTorch. (2023). PyTorch Text Documentation. https://pytorch.org/text/stable/index.html