Proyecto Google QUEST Q&A Labeling

Daniela Andrea Pavas Bedoya

Diana Carolina Huertas Gonzalez

Docente: RAUL RAMOS POLLAN

FUNDAMENTOS DE DEEP LEARNING

UNIVERSIDAD DE ANTIOQUIA

2024-2

1. Resumen

El **Google QUEST Q&A Labeling Challenge** es un desafío que se centra en desarrollar un modelo que pudiera predecir si una pregunta es interesante para una comunidad determinada y si las respuestas proporcionadas son adecuadas, útiles o, por el contrario, irrelevantes o consideradas como spam.

Este informe describe la implementación de modelos avanzados de procesamiento de lenguaje natural (NLP), como LSTM, BERT, XLNet, RoBERTa, y USE, aplicados en un conjunto de tareas de clasificación de texto. A través de diversos experimentos, comparamos los resultados obtenidos al usar modelos pre entrenados, así como las diferentes estrategias de ajuste y preprocesamiento. Además, se analizan las implicaciones de los datos desequilibrados y el rendimiento de los modelos.

2. Descripción de la Estructura de los Notebooks Entregados

Los notebooks entregados se dividen en diferentes secciones para manejar las tareas de preprocesamiento, ajuste y evaluación de modelos.

01. Exploración de datos
02. Preprocesado
03. Arquitectura de línea de base 2 capas. - Modelo LSTM
04. Arquitectura de línea base 20 capas. - Modelo LSTM
05. Modelos. - Implementación otros modelos PNL

Cada notebook sigue una estructura modular, permitiendo la flexibilidad para experimentar con distintas configuraciones de modelos y tareas.

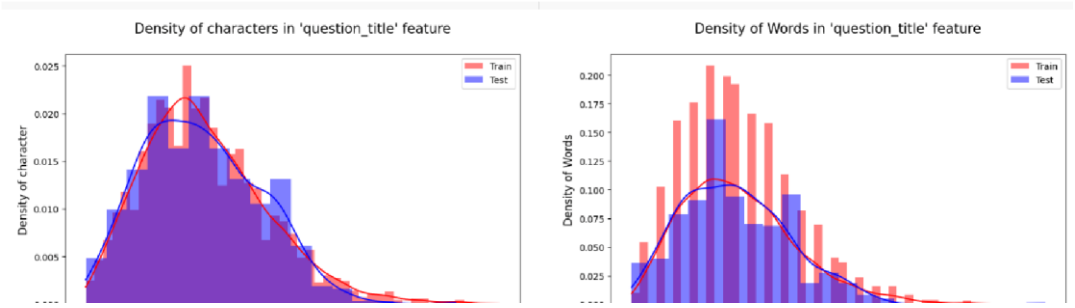
3. Descripción de la Solución

Para este se implementa el siguiente procedimiento:

a. **Carga y Preprocesamiento de Datos**: Conjunto de funciones para cargar datos, aplicar tokenización y realizar preprocesamiento de texto.
b. **Entrenamiento de Modelos**: Implementación de los modelo base LSTM, y modelos **BERT**, **RoBERTa**, **XLNet** y **USE** con parámetros optimizados.
c. **Evaluación de modelos**: Se emplearon métricas como Spearman Rank Correlation y Accuracy para evaluar los modelos entrenados.

d. **Resultados**: presentación de los resultados.

3.1 Descripción de los datos. Análisis



Durante el estudio y análisis de la base de datos es evidente observar que esta se encuentra desequilibrada con una fuerte tendencia hacia la izquierda o lado, lo que podría afectar el entrenamiento y por ende los resultados predictivos de los modelos implementados.

3.1 Arquitectura

El modelo base empleó una arquitectura LSTM (Long Short-Term Memory), diseñada para capturar dependencias a largo plazo en secuencias de texto. Esta arquitectura utiliza tres puertas principales: la de entrada, que regula la nueva información almacenada; la de olvido, que elimina información irrelevante; y la de salida, que selecciona la información utilizada en cada paso, optimizando así el procesamiento de texto para tareas como clasificación y predicción.

La arquitectura de los demás modelos implementados está basada en transformadores, donde los principales modelos probados, **XLNet**, **RoBERTa**, y **USE**. La implementación sigue el enfoque general de BERT pero con algunas modificaciones para mejorar la precisión en tareas específicas.

- **XLNet** utiliza el modelado de permutación (PLM) para capturar el contexto bidireccional. Este modelo, a diferencia de BERT, permite que el modelo prediga en cualquier punto de la secuencia sin requerir una máscara [MASK].
- **RoBERTa** es una versión optimizada de BERT que elimina la tarea de predicción de la siguiente oración (NSP), lo que mejora el rendimiento del modelo en diversas tareas de NLP.
- **USE** (Universal Sentence Encoder) se centró en la similitud semántica, lo que lo hace ideal para tareas de embeddings de oraciones.

3.2 Preprocesamiento

Los datos fueron preprocesados utilizando técnicas estándar de tokenización y vectorización, así como ajustes específicos según las necesidades de cada modelo. El preprocesamiento incluye:

- **Tokenización BPE**: Utilizada por RoBERTa para manejar vocabularios grandes.
- **Enmascaramiento dinámico**: Usado en BERT para permitir un enmascaramiento de tokens más flexible durante el entrenamiento.
- **Reducción de ruido**: Eliminación de palabras irrelevantes y normalización de texto.

Los modelos utilizados en este trabajo fueron entrenados con diferentes bases de datos según su arquitectura. Por un lado, GloVe (glove.6B), -Universidad de Stanford-, proporcionó incrustaciones de palabras preentrenadas basadas en un corpus de 6 mil millones de tokens provenientes de fuentes como Wikipedia y Gigaword. Por otro lado, modelos como BERT, XLNet y RoBERTa emplearon grandes conjuntos de datos textuales, incluidos Wikipedia, BookCorpus, CC-News, OpenWebText y Stories, permitiendo capturar un contexto más rico y bidireccional en sus representaciones lingüísticas.

3.3 Ajuste de Modelos

El ajuste de los modelos incluyó la combinación de diferentes embeddings de oraciones para observar el impacto en las tareas de clasificación de texto. La combinación de modelos como BERT, RoBERTa y XLNet mostró mejoras marginales, siendo **USE** el modelo más eficiente en términos de costos computacionales y rendimiento.

3.4 Manejo de Datos Desequilibrados

En este experimento, se observó que los modelos más complejos como BERT y RoBERTa pueden manejar problemas de desequilibrio de datos hasta cierto punto. Sin embargo, no resuelven el problema completamente. Para obtener resultados óptimos, se debe procurar un conjunto de datos equilibrado que no era este caso como lo vimos en el análisis de datos

4. Descripción de las Interacciones

Durante los experimentos, las interacciones de los modelos fueron evaluadas en términos de:

- **Entrenamiento**: Los modelos fueron entrenados en diferentes combinaciones de datos y ajustados con lotes grandes para mejorar la velocidad de optimización.
- **Evaluación**: Se utilizaron métricas como **Spearman Rank Correlation** para comparar los resultados obtenidos con los valores de referencia (ground truth). Las interacciones entre los modelos también fueron probadas al combinar embeddings de diferentes enfoques, y los resultados mostraron que, aunque hubo mejoras, no fueron significativas.

5. Resultados

La métrica de evaluación **Spearman Rank Correlation** fue seleccionada como principal, ya que evalúa la correlación ordinal, alineándose mejor con el objetivo de predecir etiquetas relacionadas con la calidad y relevancia de preguntas y respuestas. Esto por la necesidad de capturar relaciones entre las etiquetas y las predicciones en problemas multietiqueta, donde métricas como el F1-Score y el Recall no se integraron directamente debido a restricciones de tiempo y recursos computacionales.

5.1 Diferencias entre el Modelo Base y el Modelo Final

El modelo final mostró una mejora moderada en el rendimiento en comparación con el modelo base. La característica objetivo 'answer_well_written' fue mejor aprendida, lo que resultó en una mejora del **val_score**.

- **Spearman Score**: Se observó una mejora leve en la puntuación de correlación en varias características de destino.

5.2 Hallazgos y Resumen en Base a los Resultados

- **Uso de USE**: Las incrustaciones de USE resultaron ser más eficientes para la tarea actual, debido a su bajo costo computacional y su capacidad para capturar similitudes semánticas de manera efectiva.
- **Modelos Combinados**: La combinación de **BERT**, **USE**, **RoBERTa**, y **XLNet** mostró mejoras, pero no lo suficiente como para justificar su uso sobre USE en términos de eficiencia y precisión.
- **Desbalanceo de Datos**: Los modelos complejos aún tienen dificultades para manejar el desbalance de datos de manera efectiva, lo que indica la necesidad de un procesamiento adicional o el uso de datos equilibrados.
- La combinación de 'BERT, USE, RoBERTa, XLNet' ha dado resultados positivos, pero no por un margen significativo. Por lo tanto, es preferible 'USE' ya que a través de 'USE' no es una tarea muy costosa en términos de tiempo y potencia computacional.
- **Datos desequilibrados** : Es notorio que durante este desarrollo, ningún modelo puede manejar datos desequilibrados. Los modelos complejos como BERT pueden manejar problemas de desequilibrio de datos hasta cierto punto, pero no por completo. Por lo tanto, es mejor que obtengamos datos equilibrados y suficientes.
- **Posprocesamiento :** clasificación de características de destino: dado que la métrica es 'spearman_rank_corr', que tiene en cuenta el orden, se volvió importante obtener valores similares a los que (y_true) tiene en los resultados de predicción (y_pred). Y ha mostrado mejoras menores, aunque no muy significativas.

6. Conclusión

Este experimento ha demostrado que, aunque **BERT**, **RoBERTa**, y **XLNet** son poderosos modelos preentrenados, **USE** resulta ser el modelo más eficiente para tareas de clasificación de texto en entornos de producción debido a su bajo costo computacional y alta calidad en las incrustaciones de oraciones. Los modelos como **BERT** y **RoBERTa** ofrecen mejoras marginales, pero no justifican su uso en términos de tiempo y recursos frente a **USE**. Además, se concluye que el manejo de datos desequilibrados sigue siendo un desafío significativo, aunque los modelos más complejos pueden mitigar este problema en cierto grado.

Además, se concluye que el preprocesamiento de datos y el ajuste de hiperparámetros son fundamentales para obtener un modelo de alto rendimiento. Aunque los enfoques combinados pueden mejorar los resultados, es esencial equilibrar la complejidad del modelo con los recursos computacionales disponibles. Algunos retos identificados incluyen el manejo de datos desequilibrados, que afectó el rendimiento predictivo de los modelos más complejos como BERT y XLNet, y la necesidad de hardware avanzado para entrenar modelos de mayor capacidad. Estos aspectos podrían abordarse en trabajos futuros mediante el uso de técnicas de remuestreo, generación de datos sintéticos, y la evaluación de métricas adicionales como F1-Score y precisión, que permitirían una validación más exhaustiva del modelo implementado.

De igual forma, cabe mencionar que durante el desarrollo del proyecto, se identificaron varias limitaciones, como el alto costo computacional de entrenar modelos complejos como BERT, RoBERTa y XLNet, especialmente en hardware limitado. Aunque estos modelos ofrecen resultados prometedores, su implementación en entornos de producción puede ser ineficiente. Por otro lado, la dependencia de bases de datos desequilibradas afectó el rendimiento predictivo, y técnicas de re-muestreo o la generación de datos sintéticos podrían ser exploradas en trabajos futuros.

*** Consideraciones**

- **Uso de USE**: Es recomendable continuar con el uso de **USE** para tareas de clasificación de texto debido a su eficiencia y bajo costo computacional.
- **Equilibrio de Datos**: Se debe procurar utilizar conjuntos de datos equilibrados o aplicar técnicas específicas para manejar el desbalanceo de datos.
- **Optimización de Modelos**: Continuar el ajuste fino de modelos preentrenados para mejorar aún más el rendimiento en tareas específicas.
- **Continuar optimizando el uso de modelos preentrenados ajustándolos a tareas y conjuntos de datos específicos.**

**** Conclusiones finales Vs Objetivos Iniciales.**

Durante el desarrollo del proyecto, se lograron importantes avances en la implementación de modelos de procesamiento de lenguaje natural (PLN) como LSTM, BERT, XLNet, RoBERTa y USE, evaluando sus capacidades para clasificar preguntas y respuestas en el marco del **Google QUEST Q&A Labeling Challenge**. Sin embargo,
Inicialmente, se proyectaba que BERT sería el modelo principal, debido a su capacidad superior para capturar relaciones contextuales profundas y mejorar métricas como el F1-Score y el Recall, según la literatura consultada, pero durante el desarrollo se optó por dar prioridad al uso de **USE**, debido a su menor costo computacional y a que este ofreció un desempeño satisfactorio en el contexto del proyecto.
Aunque BERT y RoBERTa mostraron ventajas teóricas y mejoras marginales en tareas específicas, su costo en términos de tiempo de entrenamiento y recursos no justificó su implementación para todos los experimentos.

De igual manera se tuvo que reevaluar e implementar otra métrica de evaluación diferente a las inicialmente planteadas, debido a su alto costo computacional que hacia irremediablemente imposible su viabilidad durante las etapas de desarrollo del proyecto y que se plantea la posibilidad de nuevas valoraciones con nuevas métricas de acuerdo a la disponibilidad de los recursos.