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. Contexto de aplicación

El **Google QUEST Q&A Labeling Challenge** es un desafío propuesto para mejorar la experiencia de los usuarios al interactuar en comunidades de preguntas y respuestas en línea. Estas plataformas, como Stack Exchange y Quora, contienen millones de preguntas y respuestas con calidad y relevancia variables. Este reto busca abordar la necesidad de clasificar y etiquetar preguntas y respuestas, identificando atributos como calidad, relevancia y claridad, entre otros, que son determinantes para mejorar la experiencia del usuario.

En el contexto del desafío, el objetivo principal es desarrollar modelos capaces de comprender automáticamente preguntas y respuestas complejas. Para ello, se proporciona un conjunto de datos que contiene preguntas y respuestas extraídas de diferentes sitios web de StackExchange, con la intención de entrenar modelos que puedan evaluar la calidad y utilidad de las respuestas, así como la pertinencia de las preguntas en una discusión determinada.

El desafío 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. En otras palabras, el objetivo de la competencia es imitar el criterio humano de evaluación para clasificar el contenido de preguntas y respuestas en la plataforma, lo más cercano posible al juicio que haría un usuario experimentado.

2. Objetivo de machine learning:

El objetivo del desafío se centra en predecir 30 etiquetas que describen características como claridad, relevancia, tema y estructura en preguntas y respuestas. Para ello se utilizan modelos de procesamiento del lenguaje natural (PLN) que son capaces de analizar los datos y comprender las características del texto. Esto permite un análisis detallado de la calidad y relevancia de cada par de preguntas y respuestas en la comunidad en línea.

3. Dataset: Tipo de Datos, Tamaño y Distribución de las Clases:

- **Tipo de Datos**: El dataset está compuesto por pares de preguntas y respuestas con información textual y etiquetas asociadas. Cada entrada incluye el título de la pregunta, su cuerpo, la respuesta y etiquetas que reflejan diferentes atributos
- **Tamaño:** El conjunto de datos tiene unos 60,000 pares de preguntas y respuestas extraídas de 70 sitios web diferentes**.** En total, el dataset tiene un tamaño aproximado de 200 MB en disco.
- **Distribución de las Clases**: El problema es de múltiples etiquetas, con 30 etiquetas diferente. Las etiquetas no están distribuidas uniformemente, ya que algunas aparecen con mayor frecuencia que otras, lo que implica que el modelo debe manejar un desbalance en la distribución de las clases para obtener resultados precisos.
  Etiquetas de Pregunta: Las etiquetas con el prefijo question se relacionan con el título y/o el cuerpo de la pregunta.
  Etiquetas de Respuesta: Las etiquetas con el prefijo answer se relacionan con la respuesta.

4. **Métricas de desempeño (de Machine Learning y Negocio):**

- **Métricas de Machine Learning**:
  - **Coeficiente de correlación de Pearson**: Es la métrica oficial utilizada para evaluar el desempeño del modelo, midiendo la correlación entre las etiquetas predichas y las etiquetas reales para cada par de pregunta-respuesta.
  - **Otras métricas**: Dado que se trata de un problema de multietiquetado, también se pueden emplear métricas como el F1-score, precisión y recall para medir el rendimiento general del modelo, asegurando que se tenga un buen desempeño en todas las etiquetas y no solo en algunas.
- **Métricas de Negocio**:
  - **Mejora en la experiencia del usuario**: Clasificar preguntas y respuestas por calidad y relevancia permite a los usuarios encontrar la información más útil y precisa rápidamente, mejorando la satisfacción general.
  - **Tasa de retención y engagement**: Una mejor clasificación de preguntas y respuestas se traduce en una experiencia de usuario optimizada, lo que potencialmente incrementa la retención y participación en la plataforma.

5. **Referencias y resultados previos:**

Se exploran diversos modelos de deep learning, como CNN, LSTM y BERT

- **BERT (Bidireccional Enconder Representations from Transformers):** BERT ha demostrado ser eficaz en clasificación de texto, gracias a su arquitectura de

atención que comprende el contexto completo de las palabras. Devlin et al. (2018) reportaron mejoras del 5-10% en precisión y recall en comparación con modelos anteriores, lo que lo hace una opción sólida para el etiquetado de preguntas y respuestas.

- **CNN (Convolutional Neural Networks):** Las CNN son efectivas para extraer características locales, pero su capacidad para entender el contexto global es limitada. Kim (2014) demostró que, aunque ofrecen buenos resultados, su rendimiento es inferior al de modelos como BERT en tareas que requieren una comprensión profunda.
- **LSTM (Long Short-Term Memory):** Las LSTM son útiles para manejar secuencias y recordar información a largo plazo. Zhang et al. (2018) encontraron que superaron a modelos tradicionales en precisión al clasificar respuestas, aunque no capturan relaciones semánticas complejas tan bien como BERT.

Resultados Previos

- La literatura indica que BERT generalmente supera a CNN y LSTM en clasificación de texto, especialmente en tareas que requieren comprensión matizada del lenguaje, por lo que inicialmente se plantea la implementación del modelo BERT lo que permitirá en teoría una mejor clasificación del etiquetado analizando como se gestiona las relaciones contextuales y el etiquetado de calidad.
- Se compararán el F1-score, precisión y recall con la expectativa de que BERT logre mejoras del 5-10%, también se tendrá en consideración la velocidad de entrenamiento y el uso de recursos, así como el desempeño en datos no vistos.

Referencias

- Kim, Y. (2014). Convolutional Neural Networks for Sentence Classification. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1746-1751.
- Zhang, Y., Zhao, J., & LeCun, Y. (2018). Text Classification Algorithms: A Survey. *ACM Computing Surveys*, 51(3), 1-36.
- Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint arXiv:1810.04805*.