

ENTREGA 1: Propuesta de Proyecto de Deep Learning

Contexto de Aplicación: Diagnóstico de Experiencia del Cliente en E-commerce

En el sector de E-commerce, el éxito se basa en entender y reaccionar rápidamente al *feedback* del cliente. Una simple puntuación de 5 estrellas no revela la causa de la satisfacción o la frustración. Las empresas necesitan saber **por qué** un cliente está insatisfecho (**¿El producto? ¿La entrega? ¿la posventa?**).

Este proyecto aborda el problema crucial del diagnóstico mediante el **Análisis de Sentimiento Basado en Aspectos (Aspect-Based Sentiment Analysis - ABSA)**. Desarrollaremos un modelo de Deep Learning para analizar una reseña y clasificar su sentimiento no de forma general, sino respecto a los **aspectos clave del negocio**. Esto proporcionará información accionable directamente a los equipos de operaciones y logística.

Objetivo de Machine Learning

El objetivo de Machine Learning es construir un modelo de **clasificación multi-salida** que pueda **predecir el sentimiento (positivo, negativo, neutral)** simultáneamente para **tres aspectos** específicos del proceso de compra, a partir de la reseña de un único producto.

Técnica del Módulo 5: Se utilizarán **Modelos de Secuencia**, específicamente una red **Bidirectional-LSTM (Bi-LSTM)**, ideal para capturar las dependencias de contexto en textos largos.

Definición de Entradas y Salidas

Entrada (X): El texto completo de la reseña del producto (secuencia de palabras).

Salidas (Y): Tres clasificaciones categóricas independientes:

1. **Y1: Sentimiento sobre el Producto** (Calidad, Características, Diseño).
2. **Y2: Sentimiento sobre la Logística** (Velocidad de Envío, Integridad de la Entrega).
3. **Y3: Sentimiento sobre el Servicio/Sopor**te (Atención al Cliente, Empaque, Proceso de Devolución).

Clases por Aspecto: Positivo (1), Neutral (0), Negativo (-1).

Dataset: Synthetic E-commerce Product Reviews Dataset

Se utilizará un conjunto de datos público de Kaggle que provea las bases para esta clasificación.

Atributo	Detalle
Nombre del Dataset	Synthetic E-commerce Product Reviews Dataset (o alternativa de Amazon/Flipkart)
Fuente Sugerida	Kaggle: https://www.kaggle.com/datasets/aryan208/synthetic-e-commerce-product-reviews-dataset
Tipo de Datos	Datos de texto (reseñas) y <i>ratings</i> asociados.
Tamaño (Escala del Proyecto)	Se trabajará con una muestra de ≈100,000 reseñas para mantener la viabilidad computacional.
Tamaño en Disco (Estimado)	La muestra a utilizar será manejable (≈50–100 MB).
Distribución de Clases	La distribución inicial será desbalanceada (típicamente más reseñas positivas), lo que requerirá técnicas de remuestreo o ajustes de pérdida.
Nota de Acceso	El acceso es directo a través de la plataforma Kaggle, garantizando la disponibilidad de los datos .

Nota Importante: La fase inicial del proyecto requerirá un preprocesamiento crucial para **generar las etiquetas de aspecto** (Y1,Y2,Y3). Esto se logrará mediante una combinación de análisis de la puntuación de estrellas (*rating*) y reglas léxicas heurísticas (ej., identificar la presencia de palabras clave como "entrega rápida", "caja rota", "gran calidad") para asignar una polaridad a cada aspecto en cada reseña.

Métricas de Desempeño

Métrica de Machine Learning (ML)

La complejidad del modelo (tres salidas) requiere una métrica agregada. La métrica principal será el **Accuracy Promedio por Aspecto**.

1. **Accuracy Promedio por Aspecto:** Se calcula la Exactitud (Acc) para cada salida (Y1,Y2,Y3) y se promedia el resultado. Un buen rendimiento en esta métrica indica que el modelo es robusto para cada dimensión del problema.
2. **Puntuación F1 con Macro-Promedio:** Esto ayudará a evaluar el rendimiento del modelo en las clases minoritarias (Negativo y Neutral), crucial para los desbalances esperados.

Métrica de Negocio

La métrica más directamente vinculada al negocio y que demuestra el valor del proyecto es la **Tasa de Identificación de Fricción (TIF)**.

TIF=Total de Reseñas Negativas Reales en ese aspecto Reseñas Negativas identificadas correctamente en un aspecto crítico (ej. Logística)

Un TIF alto y bien segmentado significa que el modelo está:

1. **Activando Alertas Precisas:** Los equipos de operaciones (Logística) pueden recibir una alerta instantánea y filtrada por el **motivo exacto** de la queja, permitiendo una acción correctiva focalizada.
2. **Mejorando la Tasa de Conversión:** Al identificar y eliminar cuellos de botella específicos, la empresa reduce las barreras de compra futuras, mejorando las ventas.

Referencias y Resultados Previos

El Análisis de Sentimiento Basado en Aspectos (ABSA) es un problema complejo de NLP.

Modelos de Secuencia (LSTM/Bi-LSTM): Los *benchmarks* en problemas ABSA con múltiples aspectos suelen alcanzar una **exactitud por aspecto de entre el 75% y el 85%**, dependiendo de la calidad del etiquetado.

Primer Objetivo de Iteración: Se busca superar un **Accuracy Promedio de $\geq 70\%$** utilizando la arquitectura Bi-LSTM. Esto valida la compleja implementación de la arquitectura multi-salida y del preprocesamiento de etiquetas.

Este proyecto aplica directamente los conceptos de RNNs, LSTMs y Procesamiento de Texto del Módulo 5 a un problema de alta relevancia empresarial.