

Universidad del Valle de Guatemala

Facultad de ingeniería

Data Science

Catedrático: Luís Furlán



Laboratorio 6. Análisis de Sentimientos

Joaquín André Puente Grajeda 22296

José Antonio Mérida Castejón 201105

Guatemala, septiembre de 2025

1. Características Adicionales y su Relevancia

El modelo original basado únicamente en procesamiento secuencial presenta limitaciones en la captura de información. Debido a esto, se implementó un sistema para extraer características sentimentales de cada palabra individualmente utilizando VADER.

Características Implementadas

Ratio de Palabras Positivas

La proporción de palabras con un sentimiento emocional positivo respecto al total de palabras en la reseña cuantifica la densidad de sentimiento positivo. Esto nos da una métrica adicional y robusta, independiente del orden de las palabras. Mientras el STML capta patrones secuenciales, esta característica nos ayuda a identificar y correlacionar la densidad de palabras positivas con un sentimiento positivo general.

Ratio de Palabras Negativas

Cumple prácticamente la misma funcionalidad que el ratio de palabras positivas, complementa al modelo STML al identificar y correlacionar la densidad de palabras negativas con un sentimiento negativo general.

Longitud Original de la Reseña

Esta característica preserva la información sobre la extensión real de cada reseña antes del proceso de padding y truncamiento a 80 palabras. La longitud original puede correlacionar significativamente con la intensidad y elaboración del sentimiento expresado, ya que reseñas extremadamente cortas o extensas tienden a reflejar diferentes niveles de engagement emocional. Complementa al modelo LSTM al recuperar información contextual que se pierde durante la normalización de longitud, permitiendo distinguir entre reseñas naturalmente concisas y aquellas truncadas por el preprocessing.

2. Arquitectura del Modelo

Principios de Diseño Arquitectural

En esta arquitectura se busca combinar información secuencial y estadística que se procesan en ramas paralelas, antes de integrarse en capas superiores. Adicionalmente, fueron implementadas

diversas técnicas de regularización como dropout relativamente agresivas para evitar el sobreajuste. En el caso de los modelos implementados, este fue un problema bastante severo al aumentar la complejidad de la arquitectura.

Justificación por Componente

Capa de Embedding

Esta capa tiene un vocabulario de 50,000 tokens actualizado para los propósitos del laboratorio, al igual que una dimensionalidad de 128. Se probaron diferentes combinaciones de dimensionalidad con diferentes técnicas de regularización, tras las cuales llegamos a la conclusión que aún con un dropout agresivo los valores de dimensionalidad más grandes llevan a sobreajuste. Adicionalmente, el valor de la capa de dropout utilizada en el modelo final fue de 0.4, tras diferentes intentos evitando el sobreajuste del modelo.

LSTM Bidireccional

Un LSTM unidireccional únicamente considera el contexto precedente, en este caso pierde información valiosa del contexto posterior. En este caso, podemos tomar ejemplos como “This movie as not bad, actually quite good” dónde claramente necesitamos contexto posterior para llegar a una conclusión sobre lo que nos dice el texto. Cuenta con una dimensionalidad de 64 unidades por dirección, principalmente para mantener dimensionalidad consistente con la capa de embedding y prevenir sobreajuste. Adicionalmente, utilizamos dropout de entrada (0.4) y recurrente (0.4) ya que esta fue la técnica más eficiente para evitar el sobreajuste.

Rama de Procesamiento de Características

Para el procesamiento de características, utilizamos una capa densa bastante pequeña de 8 unidades. El objetivo de esta capa densa era aprender combinaciones o interacciones no lineales entre los features adicionales. Por ejemplo, puede ser que un review con un alto ratio de palabras negativas y una longitud sumamente larga sea muy certeramente una review negativa. Esta capa nos ayuda a darle otra dimensionalidad a los features adicionales que agregamos. También decidimos mantener una dimensionalidad baja para evitar sobreajuste, ya que el modelo puede tender a hacer predicciones de manera demasiado certera basado únicamente en los atributos adicionales.

Capa de Fusión

Para la combinación de características utilizamos una capa de fusión para poder concatenar la información aprendida de los features secuenciales y no secuenciales.

Capa de Clasificación

Esta es una capa densa de 32 neuronas y activación ReLU para aprender relaciones no-lineales de las features combinadas. En este caso, elegimos utilizar 32 neuronas ya que es un problema de

clasificación binaria y no debemos ser excesivos. Adicionalmente, agregamos dropout de 0.4 a esta capa, similar a las anteriores, para prevenir el sobreajuste.

Consideraciones de Optimización

Learning Rate Reducido (0.0003)

Los modelos multimodales requieren aprendizaje más cuidadoso para evitar dominancia de una rama sobre otra. Bajo la experimentación empírica del laboratorio, los learning rates más altos resultaban en overfitting y predicciones bastante erróneas. Esto también pudo haberse causado por diferentes ramas aprendiendo a velocidades diferentes, sin embargo, decidimos únicamente disminuir el learning rate para evitar estas dificultades.

Early Stopping y Learning Rate on Plateau

Adicionalmente, implementamos algunos callbacks para evitar que el modelo se siguiera entrenando y sobreajustando a los datos de prueba. Esto nos ayuda a ajustar la tasa de aprendizaje de manera dinámica, al igual que evitar que el modelo siga entrenándose de manera innecesaria.

3. Presentación de Resultados y Comparativa

Resultados del Modelo Mejorado

El modelo híbrido alcanzó una **accuracy final de 84.06%**, mostrando una convergencia estable hasta el epoch 4. El comportamiento del entrenamiento evidencia:

- **Convergencia rápida:** Mejora significativa del 53% al 84% en solo 4 epochs
- **Overfitting controlado:** Early stopping activado en epoch 6 al detectar degradación en validación (84.06% → 83.44%)
- **Regularización efectiva:** Los múltiples dropouts (0.4) y learning rate reducido (0.0003) mantuvieron el modelo generalizable

Comparativa con Modelo Base

Métrica	Modelo Base (LSTM Unidireccional)	Modelo Híbrido (LSTM Bidireccional + Features)
Accuracy	80.00%	84.06%
Mejora	-	+4.06%

Arquitectura	Secuencial simple	Multimodal con feature engineering
---------------------	-------------------	------------------------------------

Conclusiones

La incorporación de features estadísticas (ratios de sentimiento, longitud original) junto con el LSTM bidireccional logró una **mejora del 4%** sobre el modelo base. Este número puede parecer bajo, pero en NLP cada punto porcentual representa miles de clasificaciones correctas adicionales. Este modelo híbrido demuestra que combinar el aprendizaje secuencial profundo con features adicionales puede superar limitaciones de arquitecturas puramente neuronales. Para investigaciones posteriores se sugiere explorar la optimización de la arquitectura LSTM utilizando diferentes dimensionalidades y capas adicionales. También se puede recomendar la expansión del feature engineering, incluyendo métricas como densidad de adjetivos, identificación de intensificadores y análisis de polaridad por oraciones. Por último, la convergencia rápida sugiere que el modelo podría beneficiarse de arquitecturas más profundas siempre y cuando se controle de manera efectiva el sobreajuste y las técnicas de regularización implementadas sean robustas.