

**HAZASGOS CLAVES**

**Presentado por: GRUPO 1  
CINDY JOHANNA ZAPATA ROMERO  
HECTOR GEOVANY BELLO SANTAMARÍA  
MARIO GUERRA GUALY  
LEANDRO REYES JORDÁN**

**Docente:  
CARLOS ISAAC ZAINEA MAYA**

**UNIDAD DE ESTUDIO:  
GERENCIA DE PROYECTOS PARA CIENCIA DE DATOS  
MAESTRÍA - GRUPO 1 - M1V - VIRTUAL - 2025**

**FACULTAD DE INGENIERÍA**

**BOGOTÁ, 01 DE MARZO DE 2025  
UNIVERSIDAD EAN**

## Tabla de contenido

1. Introducción .....	3
2. Hallazgos Claves .....	3
2.1. Desempeño Comparativo de Modelos .....	3
2.2. Problemas de Generalización y Sobreajuste .....	3
2.3. Limitaciones en la Calidad y Distribución del Dataset.....	3
2.4. Falta de Evaluaciones Iterativas en Scrum .....	4
2.5. Deficiencias en el Análisis de Errores y Métricas Complementarias .....	4
2.6. Oportunidades de Mejora en la Optimización del Modelo .....	4
2.7. Falta de Expansión del Dataset y Data Augmentation .....	4
3. Conclusión .....	5

## 1. Introducción

Este informe presenta los hallazgos más relevantes identificados durante el desarrollo del proyecto de clasificación de noticias utilizando técnicas de Procesamiento de Lenguaje Natural (PLN) y redes neuronales. Se analizan aspectos relacionados con la implementación de modelos, el manejo de datos, la validación del desempeño y la integración de metodologías ágiles en el ciclo de vida del proyecto.

## 2. Hallazgos Claves

### 2.1. Desempeño Comparativo de Modelos

La aplicación de diferentes modelos permitió comparar el desempeño de estos. Se utilizaron LSTM y RNN, donde el modelo RNN no obtuvo buenos resultados, mientras que el LSTM logró un 86% de precisión en validación.

Se identificaron problemas de sobreajuste en el modelo LSTM, lo que sugiere la necesidad de ajustes en regularización.

### 2.2. Problemas de Generalización y Sobreajuste

El modelo LSTM presentó una discrepancia significativa entre entrenamiento y validación, lo que indica que aprendió patrones del conjunto de entrenamiento sin extrapolarlos correctamente a datos nuevos.

Se identificó la falta de estrategias de regularización adecuadas, como Dropout y Batch Normalization, lo que podría haber mitigado el sobreajuste.

No se exploraron modelos alternativos como GRU o Transformers ligeros, que podrían ofrecer mejor capacidad de generalización en conjuntos de datos con menor volumen.

### 2.3. Limitaciones en la Calidad y Distribución del Dataset

El dataset contenía 14,396 noticias, pero se trabajó con una muestra de 2,000 noticias, lo que pudo afectar la robustez del modelo.

No se aplicaron técnicas de balanceo de clases, lo que pudo causar sesgos en la clasificación.

La representación numérica de palabras mediante embeddings de 528 dimensiones pudo ser excesiva, contribuyendo al sobreajuste.

#### 2.4. Falta de Evaluaciones Iterativas en Scrum

La integración de Scrum con CRISP-DM permitió organizar las fases del proyecto, pero la falta de evaluaciones intermedias en Sprint 2 retrasó la identificación de problemas en la arquitectura del modelo.

La validación del modelo se realizó principalmente en Sprint 3, dejando poco margen para realizar ajustes significativos.

El Spike Técnico se concentró en la etapa final, en lugar de distribuirse a lo largo de los sprints para ajustes progresivos.

#### 2.5. Deficiencias en el Análisis de Errores y Métricas Complementarias

Se utilizó precisión como métrica principal, pero no se analizaron métricas más completas como F1-score, Recall y AUC-ROC, que hubieran permitido una evaluación más precisa del desempeño del modelo.

No se implementaron técnicas de interpretabilidad como SHAP o LIME, que habrían permitido entender mejor cómo el modelo toma decisiones y detectar posibles sesgos en la clasificación.

#### 2.6. Oportunidades de Mejora en la Optimización del Modelo

No se exploraron estrategias de ajuste de hiperparámetros con GridSearch o Bayesian Optimization, lo que podría haber mejorado la eficiencia del modelo.

No se aplicaron Schedulers de Tasa de Aprendizaje, lo que pudo afectar la convergencia del modelo durante el entrenamiento.

#### 2.7. Falta de Expansión del Dataset y Data Augmentation

La cantidad de datos utilizada pudo haber sido insuficiente para entrenar correctamente un modelo basado en LSTM.

No se aplicaron técnicas de Data Augmentation en texto, como la generación de sinónimos o reformulación de frases, que podrían haber mejorado la variabilidad del dataset y la robustez del modelo.

### 3. Conclusión

El proyecto logró implementar un modelo basado en LSTM para la clasificación de noticias, logrando una precisión del 86% en validación, lo que indica una mejora con respecto al modelo RNN, pero aún sin alcanzar de manera consistente el 85% de precisión mínima establecida en los objetivos SMART.

El uso de Scrum y CRISP-DM permitió estructurar el flujo de trabajo de manera eficiente, facilitando la priorización de actividades y la iteración en el desarrollo. Sin embargo, se identificaron áreas de mejora clave en la validación temprana del modelo, ajuste de hiperparámetros y optimización del preprocesamiento de datos.

El uso de Scrum y CRISP-DM permitió estructurar el flujo de trabajo, pero la falta de validaciones intermedias y ajustes iterativos impactó la detección temprana de errores.

Para futuros desarrollos, se recomienda:

- Optimizar la regularización del modelo con Dropout y Batch Normalization.
- Ajustar la distribución de datos y embeddings para reducir sesgos.
- Realizar validaciones iterativas dentro de cada sprint para detectar problemas de generalización antes del despliegue.
- Explorar arquitecturas más avanzadas como Transformers ligeros y modelos híbridos de PLN.
- Ampliar el dataset y aplicar técnicas de data augmentation para mejorar la representatividad y robustez del modelo.