

**Bitácora de ejecución**

**Análisis de Sentimientos en Ambientes con Comentarios aplicando Procesamiento de**

**Lenguaje Natural NLP**

Daniel Cortés Pérez

Director(a):

PhD. Ingeniería y Néstor Darío Duque Méndez

Universidad Nacional de Colombia Facultad de Administración, Departamento de Informática y Computación

Manizales, Colombia

2024

**Historia de revisiones**

|  |  |  |  |
| --- | --- | --- | --- |
| **Versión** | **Autor(es)** | **Descripción** | **Fecha** |
| 1.0 | Daniel Cortés Pérez | Creación del documento/ Revisión bibliografía | 05 de febrero de 2024 |
| 2.0 | Daniel Cortés Pérez | Ejecución de pruebas preliminares con las técnicas seleccionadas | 27 de febrero de 2024 |
| 3.0 | Daniel Cortés Pérez | Ejecución de pruebas aplicados a las bases de datos recopiladas | 18 de marzo de 2024 |
| 4.0 | Daniel Cortés Pérez | Evaluación del Rendimiento del Algoritmo | 20 de abril de 2024 |
| 5.0 | Daniel Cortés Pérez | Aplicación del Algoritmo en un Contexto Específico (Parte final) | 13 de mayo de 2024 |

**Índice**

Contenido

[Introducción 4](#_Toc178783331)

[Bitácora pruebas de ejecución del algoritmo desarrollado 5](#_Toc178783332)

[1. Revisión Bibliográfica 5](#_Toc178783333)

[ Recolección de Información 5](#_Toc178783334)

[ Análisis de la Literatura y Selección de Técnicas 5](#_Toc178783335)

[1. Selección y Preparación de Herramientas de Análisis de Datos 6](#_Toc178783336)

[ Instalación y Configuración de Entorno 6](#_Toc178783337)

[ Preprocesamiento de Datos y Preparación del Conjunto de Datos 6](#_Toc178783338)

[2. Desarrollo del Algoritmo de Análisis de Sentimientos 7](#_Toc178783339)

[ Diseño del Algoritmo 7](#_Toc178783340)

[ Implementación del Algoritmo 7](#_Toc178783341)

[3. Evaluación del Rendimiento del Algoritmo 8](#_Toc178783342)

[ Evaluación del Modelo en Datos de Prueba 8](#_Toc178783343)

[4. Aplicación del Algoritmo en un Contexto Específico 8](#_Toc178783344)

[ Análisis de Comentarios en un Contexto de Reseñas de Productos 8](#_Toc178783345)

[5. Conclusiones y Recomendaciones 9](#_Toc178783346)

[Conclusiones 9](#_Toc178783347)

[Recomendaciones 9](#_Toc178783348)

[Bibliografía 10](#_Toc178783349)

# Introducción

El análisis de sentimientos es una subdisciplina del procesamiento de lenguaje natural (NLP, por sus siglas en inglés) que permite identificar y clasificar las emociones presentes en textos escritos, tales como comentarios en línea, reseñas de productos y publicaciones en redes sociales. Este enfoque es fundamental en la era digital actual, donde las empresas, gobiernos y organizaciones dependen cada vez más del análisis de grandes volúmenes de datos para entender mejor las percepciones, opiniones y emociones de sus usuarios o ciudadanos. El objetivo principal de este proyecto es desarrollar un algoritmo que automatice el proceso de análisis de sentimientos, proporcionando una herramienta eficaz para analizar grandes cantidades de comentarios en diversos dominios, como comercio electrónico, servicios al cliente y redes sociales.

Este trabajo surge de la necesidad de mejorar la precisión y eficiencia en la identificación de sentimientos positivos, negativos y neutros en comentarios digitales. La propuesta se enfoca en aplicar técnicas avanzadas de procesamiento de lenguaje natural y aprendizaje automático para diseñar un modelo que no solo clasifique con precisión los sentimientos, sino que también identifique patrones emocionales que podrían ser útiles para la toma de decisiones estratégicas en diversos sectores. A través de una revisión exhaustiva de la literatura existente y la implementación de técnicas de aprendizaje automático, este proyecto busca optimizar los algoritmos existentes y aplicarlos en un contexto real para demostrar su aplicabilidad práctica.

A lo largo de esta bitácora, se documentarán todas las fases del desarrollo del algoritmo, desde la selección de técnicas y herramientas, hasta la implementación, evaluación y refinamiento del modelo. Además, se presentarán los resultados obtenidos de la aplicación del algoritmo en diferentes conjuntos de datos, destacando tanto los logros como los desafíos encontrados. Con este enfoque, se espera contribuir al campo del análisis de sentimientos, proporcionando un algoritmo robusto que pueda ser utilizado en múltiples contextos para mejorar la comprensión de las emociones y opiniones expresadas en línea.

# Bitácora pruebas de ejecución del algoritmo desarrollado

**Fecha**: 05/02/2024 **Ejecución No**: 1 **Detalles de la Prueba:**

### Revisión Bibliográfica

## Recolección de Información

**Fecha:** 05/02/2024

* **Descripción:**

Se inició la revisión de literatura sobre análisis de sentimientos en ambientes con comentarios, utilizando bases de datos científicas como IEEE, mendeley y Google Scholar, entre otras. La investigación se centró en estudios clave sobre técnicas de procesamiento de lenguaje natural (NLP) aplicadas al análisis de sentimientos, con el objetivo de identificar los enfoques más relevantes y recientes en esta área.

* **Métodos:** La búsqueda se realizó utilizando palabras clave como "análisis de sentimientos", "NLP", "procesamiento de lenguaje natural", "machine learning" y "algoritmos de clasificación de sentimientos". Se seleccionaron los estudios más citados y relevantes de los últimos 5 años para asegurar la pertinencia y actualidad de la revisión.
  + **Resultados:**
    - Se identificaron varias fuentes relevantes, entre las cuales destacan estudios como los de Liu (2012), que proporciona una base sólida sobre análisis de sentimientos y minería de opiniones, y Pang & Lee (2008), que presentan una clasificación de sentimientos utilizando técnicas de aprendizaje automático.
    - Esta fase permitió definir las tendencias actuales en la investigación, identificando los principales desafíos del campo, como la subjetividad del lenguaje, la ambigüedad emocional y la necesidad de conjuntos de datos bien etiquetados.

## Análisis de la Literatura y Selección de Técnicas

**Fecha:** 12/02/2024

* **Descripción:**
  + Tras la recolección de la información, se procedió a un análisis exhaustivo de las técnicas propuestas en la literatura. Este análisis permitió identificar las mejores herramientas para el desarrollo del algoritmo de análisis de sentimientos. Se priorizaron técnicas como la vectorización de palabras (TF-IDF), la lematización, y los algoritmos de clasificación como Naive Bayes y Regresión Logística, que han demostrado alta efectividad en la clasificación de textos cortos.
  + **Desafíos:** Uno de los principales desafíos fue la evaluación de los diferentes enfoques en función de la complejidad de implementación y los recursos computacionales necesarios. Algunos estudios sugerían el uso de modelos más complejos como los basados en redes neuronales, pero se optó por un enfoque más eficiente en términos de recursos.
  + **Resultados:**
    - Se seleccionaron las siguientes técnicas para el desarrollo del algoritmo: TF-IDF para la vectorización del texto, NLTK para el preprocesamiento (eliminación de stopwords, lematización y tokenización), y Scikit-learn para la implementación de modelos de aprendizaje automático como Regresión Logística y Naive Bayes.

### Selección y Preparación de Herramientas de Análisis de Datos

**Ejecución No**: 2

## Instalación y Configuración de Entorno

**Fecha:** 27/02/24

* **Descripción:**
  + Se configuró el entorno de desarrollo en Python, instalando bibliotecas clave para el procesamiento de lenguaje natural (NLTK), análisis de datos (pandas, numpy), y modelado de aprendizaje automático (Scikit-Learn). Esta configuración fue esencial para garantizar que las herramientas necesarias estuvieran disponibles y actualizadas.
  + **Herramientas utilizadas:** Python 3.8, NLTK, pandas, numpy, Scikit-Learn.
  + **Resultados:**
    - El entorno fue configurado correctamente, y las bibliotecas fueron instaladas sin problemas. Se realizaron pruebas preliminares para asegurar que las herramientas estuvieran listas para la fase de desarrollo del algoritmo.

## Preprocesamiento de Datos y Preparación del Conjunto de Datos

**Fecha:** 02/03/24

* **Descripción:**
  + Se procedió a la limpieza de un conjunto de datos de comentarios en línea. Los datos contenían ruido significativo, como enlaces, menciones, y caracteres especiales, los cuales fueron eliminados utilizando funciones personalizadas en Python. Se utilizaron técnicas de tokenización y lematización para normalizar el texto y mejorar la precisión del análisis.
  + **Desafíos:** El preprocesamiento de datos fue un proceso largo debido a la gran cantidad de datos y la variabilidad en la estructura del texto. Se requirieron varias iteraciones para garantizar que los datos estuvieran listos para el modelado.
  + **Resultados:**
    - Se preparó un conjunto de datos limpio y estructurado, listo para ser utilizado en el desarrollo del algoritmo. Este conjunto incluye 45,000 reseñas de productos y 30,000 comentarios de redes sociales, con etiquetas de sentimiento positivo, negativo y neutro.

### Desarrollo del Algoritmo de Análisis de Sentimientos

**Ejecución No**: 3

## Diseño del Algoritmo

**Fecha:** 18/03/24

* **Descripción:**
  + Basado en la revisión bibliográfica y las técnicas seleccionadas, se procedió al diseño inicial del algoritmo de análisis de sentimientos. Se diseñó un pipeline que incluye las siguientes etapas: recolección de datos, preprocesamiento, vectorización mediante TF-IDF, y clasificación utilizando Regresión Logística y Naive Bayes.
  + **Desafíos:** Asegurar que el pipeline fuera eficiente y modular, de manera que pudiera ser adaptado y escalado a conjuntos de datos más grandes en el futuro. También se tuvo en cuenta la posibilidad de incorporar modelos más avanzados, como redes neuronales, en futuras iteraciones.
  + **Resultados:**
    - Se definió una arquitectura clara para el algoritmo, con módulos específicos para la limpieza de datos, la vectorización de texto, y la implementación de modelos de clasificación.

## Implementación del Algoritmo

**Fecha:** 22/03/24

* **Descripción:**
  + Se implementó el algoritmo en Python, utilizando las bibliotecas previamente configuradas. El modelo fue entrenado con un subconjunto del conjunto de datos, utilizando una proporción 80/20 para el entrenamiento y validación. Se probaron ambos modelos (Regresión Logística y Naive Bayes) para evaluar su desempeño inicial.
  + **Resultados:**
    - La Regresión Logística presentó un F1 Score promedio del 0.85 en la clasificación de sentimientos, mientras que Naive Bayes alcanzó un 0.82. Estos resultados indican que ambos modelos son eficaces, pero la Regresión Logística ofrece un rendimiento ligeramente superior.

### Evaluación del Rendimiento del Algoritmo

**Ejecución No**: 4

## Evaluación del Modelo en Datos de Prueba

**Fecha:** 20/04/24

* **Descripción:**
  + Se evaluó el rendimiento del algoritmo utilizando un conjunto de prueba que no había sido utilizado en el entrenamiento. Se utilizaron métricas de precisión, recall, y F1 Score para medir la efectividad de los modelos. Además, se creó una matriz de confusión para analizar los errores de clasificación más comunes.
  + **Resultados:**
    - El modelo de Regresión Logística mantuvo una precisión del 85%, con un F1 Score de 0.84 en los datos de prueba. Los errores más comunes se produjeron en la clasificación de comentarios neutros, lo que indica una posible mejora futura en la diferenciación de sentimientos neutros frente a positivos y negativos.

### Aplicación del Algoritmo en un Contexto Específico

**Ejecución No**: 5

## Análisis de Comentarios en un Contexto de Reseñas de Productos

**Fecha:** 13/05/24

* **Descripción:**
  + Se aplicó el algoritmo a un conjunto de datos de reseñas de productos, específicamente comentarios de clientes sobre dispositivos electrónicos. El objetivo era identificar patrones en las reseñas, como la satisfacción general del cliente con el producto y las principales quejas.
  + **Resultados:**
    - El algoritmo clasificó correctamente el 87% de las reseñas como positivas o negativas. Los comentarios negativos se centraban en la durabilidad de los productos, mientras que los positivos destacaban la funcionalidad y el diseño. Estos hallazgos podrían ser útiles para las empresas en la mejora de sus productos.

### Conclusiones y Recomendaciones

## Conclusiones

* El uso de Regresión Logística y técnicas de preprocesamiento adecuadas proporcionó un rendimiento sólido en la clasificación de sentimientos.
* El preprocesamiento de datos fue clave para mejorar la precisión del algoritmo, especialmente en la eliminación de ruido y la normalización del texto.
* Existen oportunidades para mejorar la clasificación de sentimientos neutros, y se podría explorar el uso de modelos más complejos como redes neuronales.

## Recomendaciones

* Se sugiere aplicar el algoritmo a otros dominios, como análisis de redes sociales o retroalimentación de clientes en diferentes industrias.
* Considerar la incorporación de técnicas avanzadas de aprendizaje profundo, como LSTM, para mejorar el rendimiento en la clasificación de emociones complejas.

# Bibliografía

Bengio, Y., Courville, A., & Vincent, P. (2013). Representation Learning: A Review and New Perspectives. 1798-1828.

Bird, S. K. (2009). Natural Language Processing with Python. *O'Reilly Media.*

Bo Pang, L. L. (202). Thumbs up? Sentiment Classification using Machine Learning Techniques. *Conference on Empirical Methods in Natural Language Processing*, 59-67.

González Guerra, C. (2020). *Paridad de herramientas para el análisis de sentimientos sobre un CORPUS en español generado desde twitter.* Obtenido de Repositorio Institucional Academico: http://repositorio.unab.cl/xmlui/handle/ria/15339

H. Cunningham, D. M. (2002). GATE: A framework and graphical development environment for robust NLP tools and applications. *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics*.

Hunter, J. (2010). Matplotlib: A 2D Graphics Environment. *Computing in Science & Engineering.*

Hutto, C. &. (2014). *VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text.* USA: Vol. 8 No. 1 (2014): Eighth International AAAI Conference on Weblogs and Social Media.

Lee, B. P. (08 de 07 de 2008). *Opinion Mining and Sentiment Analysis.* USA: Foundations and Trends® in Information Retrieval: Vol. 2.

Lei Zhang, S. W. (2018). Deep learning for sentiment analysis: A survey.

Liu, B. (2012). *Sentiment Analysis and Opinion Mining.* USA: Springer Cham.

Liu, B. (2012). *Sentiment Analysis and Opinion Mining. Synthesis Lectures on Human Language Technologies, 5.* USA.

McKinney, W. (2010). Data Structures for Statistical Computing in Python. *Proceedings of the 9th Python in Science Conference.*

Md Shad Akhtar, A. K. (2016). A Hybrid Deep Learning Architecture for Sentiment Analysis. *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, 482-493.

Medhat, W. H. (2014). Sentiment analysis algorithms and applications: A survey. . *Ain Shams Engineering Journal, 5(4),* , 1093-1113.

Mikolov, T. S. (2013). Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems, 26.*

Oliphant, T. (2006). A guide to NumPy. . *Trelgol Publishing.*

Pedregosa, F. e. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research.*

Saif M. Mohammad, S. K. (Atlanta, USA). NRC-Canada: Building the State-of-the-Art in Sentiment Analysis of Tweets. *Proceedings of the seventh international workshop on Semantic Evaluation Exercises (SemEval-2013)*.

Ungar., M. A.-M. (2017, Vancouver, Canada). EmoNet: Fine-Grained Emotion Detection with Gated Recurrent Neural Networks. *In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 718–728.

Vidal Castro, C. L. (2017). *Comparación de rendimiento de técnicas de aprendizaje automático para análisis de afecto sobre textos en español.* Obtenido de SIBUBB: http://repobib.ubiobio.cl/jspui/handle/123456789/1772

Waskom., M. (2021). Seaborn: Statistical Data Visualization. *Journal of Open-Source Software.*