

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

**Lenguaje Natural NLP**

**Daniel Cortés Pérez**

Universidad Nacional de Colombia

Facultad de Administración, Departamento de Informática y Computación

Manizales, Colombia

2024

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

**Lenguaje Natural NLP**

**Daniel Cortés Pérez**

Trabajo de grado presentado como requisito parcial para optar al título de:

**Administrador(a) de Sistemas Informáticos**

Director(a):

PhD. Ingeniería Néstor Darío Duque Méndez

Modalidad del trabajo de grado:

Trabajo investigativo/Trabajo Final

Universidad Nacional de Colombia

Facultad de Administración, Departamento de Informática y Computación

Manizales, Colombia

2024

*A mis padres, a mi familia.*

*"Lo más importante que he aprendido es que la familia no es algo importante. Es todo. Mis logros no valen nada sin el amor y el apoyo incondicional de mi familia".*

*Barack Obama*

**Agradecimientos**

A mis padres, por ser mi mayor ejemplo de esfuerzo, perseverancia y apoyo incondicional. Gracias por sus consejos por enseñarme que no hay metas imposibles.

A mi familia, por su cariño y respaldo constante, quienes siempre han creído en mí y me han dado la fuerza para seguir adelante en cada paso de este camino.

A mi pareja, por su paciencia, comprensión y motivación en los momentos más difíciles. Tu compañía ha sido fundamental para alcanzar este logro.

A mis compañeros de carrera, por compartir conmigo este viaje académico lleno de retos, aprendizajes y grandes experiencias.

A mis maestros, por su dedicación y pasión por la enseñanza. Gracias por transmitir sus conocimientos y guiarme con sabiduría a lo largo de este proceso.

Este logro es gracias a todos ustedes.

Resumen

**Análisis de Sentimientos en Ambientes con Comentarios aplicando Procesamiento de Lenguaje Natural NLP.**

Este trabajo de grado se centra en el desarrollo de un algoritmo de análisis de sentimientos aplicado a comentarios en línea mediante técnicas de procesamiento de lenguaje natural (NLP). El objetivo principal es identificar de manera precisa los sentimientos positivos y negativos en los comentarios, optimizando el modelo para asegurar su eficacia. Utilizando pandas, los datos fueron cargados y preparados adecuadamente, mientras que numpy permitió su manipulación y procesamiento. Para la clasificación de sentimientos, se implementaron algoritmos de sklearn, los cuales fueron evaluados y optimizados utilizando métricas como ROC AUC y F1 Score.

La visualización de datos se realizó mediante seaborn y matplotlib, proporcionando gráficos y visualizaciones claras que facilitan la interpretación de los resultados. Este proceso no solo asegura la precisión del análisis de sentimientos, sino que también permite presentar los hallazgos de manera convincente.

El análisis de sentimientos es una herramienta crucial en áreas como marketing, ventas y desarrollo de productos, ya que proporciona información valiosa sobre las opiniones y sentimientos de la audiencia. La implementación de este proyecto demuestra la viabilidad y utilidad práctica de estas técnicas, subrayando la importancia de un análisis preciso para la toma de decisiones estratégicas.

**Palabras clave:** análisis de sentimientos, procesamiento de lenguaje natural, pandas, sklearn.

Abstract

**Sentiment Analysis in Comment Environments using NLP Natural Language Processing.**

This degree work focuses on the development of a sentiment analysis algorithm applied to online comments using Natural Language Processing (NLP) techniques. The main objective is to accurately identify positive and negative sentiments in comments, optimizing the model to ensure its effectiveness. Using pandas, the data were loaded and prepared appropriately, while numpy enabled their manipulation and processing. For sentiment classification, sklearn algorithms were implemented, which were evaluated and optimized using metrics such as ROC AUC and F1 Score.

Data visualization was performed using seaborn and matplotlib, providing clear graphs and visualizations that facilitate the interpretation of the results. This process not only ensures the accuracy of the sentiment analysis, but also allows the findings to be presented in a compelling manner.

Sentiment analysis is a crucial tool in areas such as marketing, sales and product development, as it provides valuable information about audience opinions and sentiments. The implementation of this project demonstrates the feasibility and practical utility of these techniques, highlighting the importance of accurate analysis for strategic decision making.

**Keywords:** sentiment analysis, natural language processing, pandas, sklearn.

Pág.

Contenido

[1. Presentación del Trabajo de Grado 3](#_Toc178664154)

[1.1 Planteamiento Del Problema o Situación Abordada 3](#_Toc178664155)

[1.2 Objetivo General 4](#_Toc178664156)

[1.3 Objetivos Específicos 4](#_Toc178664157)

[1.4 Metodología 5](#_Toc178664158)

[1.4.1 Recolección y Preparación de Datos: 5](#_Toc178664159)

[1.4.2 Manipulación de Datos: 5](#_Toc178664160)

[1.4.3 Desarrollo del Modelo: 5](#_Toc178664161)

[1.4.4 Evaluación del Modelo: 6](#_Toc178664162)

[1.4.5 Visualización de Resultados: 6](#_Toc178664163)

[1.5 Tablas NLP 6](#_Toc178664164)

[1.6 Figuras NLP 11](#_Toc178664173)

[2. Revisión de literatura 17](#_Toc178664182)

[ Análisis de Sentimientos 17](#_Toc178664183)

[ Técnicas y Algoritmos 17](#_Toc178664184)

[ Herramientas y Bibliotecas 17](#_Toc178664185)

[ Procesamiento de Lenguaje Natural (NLP) 17](#_Toc178664186)

[ Modelos de Representación de Texto 18](#_Toc178664187)

[ Preprocesamiento de Datos 18](#_Toc178664188)

[3. Desarrollo 19](#_Toc178664189)

[3.1 Recopilación de los Datos 19](#_Toc178664190)

[3.1.1 Reseñas de Películas: 19](#_Toc178664191)

[3.1.2 Tweets de Twitter: 19](#_Toc178664192)

[3.2 Preprocesamiento de Datos 20](#_Toc178664193)

[3.3 Desarrollo del Algoritmo 20](#_Toc178664194)

[3.4 Visualización de Resultados 21](#_Toc178664195)

[3.5 Optimización y Validación 21](#_Toc178664196)

[4. Conclusiones y recomendaciones 23](#_Toc178664197)

[4.1 Conclusiones 23](#_Toc178664198)

[ Importancia del Preprocesamiento de Datos: 23](#_Toc178664199)

[ Comparación de Algoritmos de Clasificación: 23](#_Toc178664200)

[ Relevancia de la Visualización de Datos: 23](#_Toc178664201)

[ Uso de Métricas de Evaluación: 23](#_Toc178664202)

[ Aplicabilidad en Distintos Contextos: 23](#_Toc178664203)

[ Desafíos en la Recolección de Datos: 24](#_Toc178664204)

[ Impacto del Tamaño del Conjunto de Datos: 24](#_Toc178664205)

[4.2 Recomendaciones 24](#_Toc178664206)

[ Ampliación del Conjunto de Datos: 24](#_Toc178664207)

[ Optimización y Evaluación Continua del Modelo: 24](#_Toc178664208)

[ Mejora en el Preprocesamiento de Datos: 25](#_Toc178664209)

[5. Bibliografía 26](#_Toc178664210)

Lista de figuras

Pág.

[1.5 Tablas NLP 6](#_Toc178664164)

[1.5.1 Figura 1: Base de datos # 1 Caso películas, histograma de datos. 6](#_Toc178664165)

[1.5.2 Figura 2: Base de datos # 1 Caso películas, función de matriz confusión. 7](#_Toc178664166)

[1.5.3 Figura 3: Base de datos # 1 Caso películas, Matriz de confusión subconjunto de entrenamiento. 7](#_Toc178664167)

[1.5.4 Figura 4: Base de datos # 1 Caso películas, Matriz de confusión subconjunto de prueba. 8](#_Toc178664168)

[1.5.5 Figura 5: Base de datos # 1 Caso películas, Histograma de coeficientes. 9](#_Toc178664169)

[1.5.6 Figura 6: Base de datos # 2 Caso Twitter, histograma de datos. 9](#_Toc178664170)

[1.5.7 Figura 7: Base de datos # 2 Caso Twitter, función de matriz confusión. 10](#_Toc178664171)

[1.5.8 Figura 8: Base de datos # 2 Caso Twitter, matriz de confusión. 10](#_Toc178664172)

Lista de tablas

Pág.

[1.6 Figuras NLP 11](#_Toc178664173)

[1.6.1 Tabla 1: Base de datos # 1 Caso películas, ilustración de la base de datos. 11](#_Toc178664174)

[1.6.2 Tabla 2: Base de datos # 1 Caso películas, totalidad de la base de datos. 11](#_Toc178664175)

[1.6.3 Tabla 3: Base de datos # 1 Caso películas, base de datos depurada. 12](#_Toc178664176)

[1.6.4 Tabla 5: Base de datos # 1 Caso películas, base de datos del subconjunto de entrenamiento. 13](#_Toc178664177)

[1.6.5 Tabla 6: Base de datos # 1 Caso películas, base de datos del subconjunto de prueba. 14](#_Toc178664178)

[1.6.6 Tabla 7: Base de datos # 2 Caso Twitter, totalidad de la base de datos. 15](#_Toc178664179)

[1.6.7 Tabla 8: Base de datos # 2 Caso Twitter, plenitud de la base de datos. 15](#_Toc178664180)

[1.6.8 Tabla 9: Base de datos # 2 Caso Twitter, base de datos del subconjunto de entrenamiento…...….………………………………………………………………………16](#_Toc178664181)

Introducción

En la era digital actual, la cantidad de datos generados por usuarios en plataformas en línea es inmensa. Comentarios, opiniones y reseñas se acumulan diariamente, proporcionando una mina de oro de información sobre percepciones y sentimientos. El análisis de sentimientos es una subdisciplina del procesamiento del lenguaje natural (NLP), emerge como una herramienta esencial para interpretar estas emociones y opiniones expresadas en textos. El análisis de sentimientos permite clasificar los comentarios como positivos, negativos o neutros, brindando a las empresas y organizaciones una comprensión profunda de la opinión pública. Esta capacidad es invaluable en diversas áreas como marketing, ventas y desarrollo de productos, donde comprender la voz del cliente puede definir estrategias y decisiones. **La importancia de este campo se refleja en la amplia gama de estudios y avances reportados en la literatura académica, como los trabajos pioneros de Pang & Lee (2008) y Liu (2012), que ofrecen una base teórica sólida y una visión general de las técnicas y aplicaciones del análisis de sentimientos. Pang & Lee (2008) proporcionan una revisión exhaustiva que destaca el análisis de sentimientos y la opinión como áreas significativas de interés en la recuperación de información y el NLP. Este trabajo subraya la importancia de distinguir entre los hechos y las opiniones en los textos, así como de clasificar las opiniones en categorías como positivas, negativas o neutras. Por otro lado, Liu (2012) profundiza en el análisis de sentimientos y la minería de opiniones desde una perspectiva de procesamiento de lenguaje natural, abordando tanto los aspectos teóricos como las aplicaciones prácticas de esta disciplina. Este trabajo de grado se enfoca en desarrollar un algoritmo eficiente para el análisis de sentimientos, aprovechando bibliotecas avanzadas de Python como pandas, numpy y sklearn. Utilizando técnicas de preprocesamiento de datos y modelos de clasificación, se busca optimizar la precisión y eficiencia del algoritmo. La visualización de resultados mediante herramientas como seaborn y matplotlib asegura una interpretación clara y convincente de los datos.**

# Presentación del Trabajo de Grado

**Este trabajo de grado aborda el desarrollo y la implementación de un algoritmo avanzado para el análisis de sentimientos, aplicado a comentarios en línea mediante técnicas de procesamiento de lenguaje natural (NLP). La capacidad de interpretar emociones y opiniones expresadas en textos digitales es fundamental en la actualidad, especialmente en un entorno donde la comunicación y la retroalimentación de los usuarios se producen de manera masiva y constante.**

**Para llevar a cabo este proyecto, he utilizado herramientas y bibliotecas avanzadas de Python, tales como pandas para la carga y preparación de datos, numpy para su manipulación, y sklearn para la clasificación de sentimientos. El proceso incluye desde la pre-procesación y limpieza de datos hasta la implementación de modelos de clasificación, optimizados y evaluados mediante métricas como ROC AUC y F1 Score para asegurar la máxima precisión y eficiencia. Además, he incorporado técnicas de visualización de datos utilizando seaborn y matplotlib, lo que permite una presentación clara y convincente de los resultados obtenidos. Estas visualizaciones no solo facilitan la comprensión de los patrones y tendencias en los datos, sino que también ayudan a comunicar los hallazgos de manera efectiva a diversas audiencias.**

**Este proyecto no solo subraya la importancia del análisis de sentimientos en el mundo digital actual, sino que también ofrece un enfoque práctico y efectivo para su implementación y optimización continua. Con este trabajo, espero contribuir al campo del procesamiento de lenguaje natural y proporcionar una herramienta valiosa para la interpretación de datos textuales, ayudando a las organizaciones a tomar decisiones más informadas y estratégicas.**

## **Planteamiento Del Problema o Situación Abordada**

**En la actualidad, las empresas y organizaciones se enfrentan a la necesidad creciente de comprender y analizar las opiniones y sentimientos de sus usuarios y clientes expresados a través de comentarios en línea. Las plataformas digitales generan una enorme cantidad de datos textuales que, si se analizan adecuadamente, pueden proporcionar valiosos insights para mejorar productos, servicios y estrategias de marketing. Sin embargo, la tarea de analizar manualmente estos datos es impracticable debido a su volumen y complejidad.**

**Este trabajo aborda estos desafíos mediante el desarrollo de un algoritmo robusto para el análisis de sentimientos en comentarios en línea, utilizando herramientas avanzadas de Python como pandas, numpy y sklearn. La metodología incluye la preparación y manipulación de datos, la clasificación de sentimientos y la visualización de resultados para facilitar la interpretación de datos. Además, se implementan métricas de evaluación como ROC AUC y F1 Score para optimizar y asegurar la precisión del modelo.**

**La solución propuesta no solo facilita la comprensión de las opiniones de la audiencia, sino que también permite a las organizaciones tomar decisiones informadas basadas en datos reales, mejorando así sus estrategias y respuestas a las necesidades del mercado.**

## Objetivo General

***Desarrollar un algoritmo basado en NLP para el análisis automático de sentimientos en comentarios en línea en diversos contextos y dominios.*** Esto surge de la necesidad de comprender y analizar de manera efectiva las vastas cantidades de datos generados en plataformas digitales, permitiendo la identificación de tendencias y la extracción de insights valiosos. La motivación detrás de este objetivo es mejorar la precisión y eficiencia en la interpretación de las emociones y opiniones expresadas en el texto, contribuyendo significativamente al campo del procesamiento del lenguaje natural y ofreciendo herramientas útiles para diversas aplicaciones prácticas.

## Objetivos Específicos

* Revisar la literatura sobre análisis de sentimientos en ambientes con comentarios.
* Identificar y seleccionar las técnicas, algoritmos y herramientas adecuadas para el desarrollo de la solución.
* Diseñar y desarrollar el algoritmo de análisis de sentimientos.
* Evaluar el desempeño del algoritmo en un conjunto de datos de comentarios en línea.
* Aplicar el algoritmo a un conjunto de datos de comentarios en línea en un contexto específico.

Estos objetivos se harán para profundizar en el entendimiento y manejo de emociones y opiniones en textos digitales, lo cual es esencial en la era de la información. Identificar técnicas adecuadas y desarrollar un algoritmo innovador permitirá superar limitaciones actuales en el análisis de sentimientos, mejorando la precisión y aplicabilidad en diversos contextos. La evaluación del algoritmo asegura su efectividad y la aplicación en contextos específicos demuestra su utilidad práctica, contribuyendo significativamente al avance del procesamiento del lenguaje natural y análisis de sentimientos.

## Metodología

La metodología de este trabajo de grado se basa en un enfoque sistemático para desarrollar y evaluar un algoritmo de análisis de sentimientos utilizando técnicas avanzadas de procesamiento de lenguaje natural (NLP). A continuación, se detallan los pasos seguidos:

### Recolección y Preparación de Datos:

* **Cargar Datos:** Se utilizaron bibliotecas de Python como pandas para cargar y explorar los datos de comentarios en línea. Esta fase incluyó la recolección de datos desde diversas fuentes para asegurar una muestra representativa y amplia.
* **Preprocesamiento de Datos:** Los datos fueron limpiados y preprocesados para eliminar ruido, manejar valores nulos y normalizar texto. Se emplearon técnicas como la eliminación de stopwords, la lematización y la tokenización utilizando bibliotecas como nltk.

### Manipulación de Datos:

* **Análisis Exploratorio:** Se realizó un análisis exploratorio de los datos con numpy y pandas para entender mejor la distribución de los comentarios y sus características principales.
* **Transformación de Datos:** Se transformaron los datos textuales en vectores numéricos mediante la técnica TF-IDF (Term Frequency-Inverse Document Frequency), utilizando la herramienta sklearn.

### Desarrollo del Modelo:

* **Selección de Modelos:** Se seleccionaron y compararon varios algoritmos de clasificación, como Regresión Logística, Naive Bayes, implementados con sklearn.
* **Entrenamiento y Validación:** Los modelos fueron entrenados y validados utilizando técnicas de validación cruzada para asegurar su rendimiento y evitar sobreajuste.

### Evaluación del Modelo:

* **Métricas de Evaluación:** Se utilizaron métricas como ROC AUC, F1 Score, Precisión y Recall para evaluar la eficacia y precisión de los modelos. Estas métricas permitieron comparar el rendimiento de los diferentes algoritmos y seleccionar el más adecuado.
* **Optimización:** Se implementaron técnicas de optimización como la búsqueda de hiperparámetros y el ajuste fino de modelos para mejorar su rendimiento.

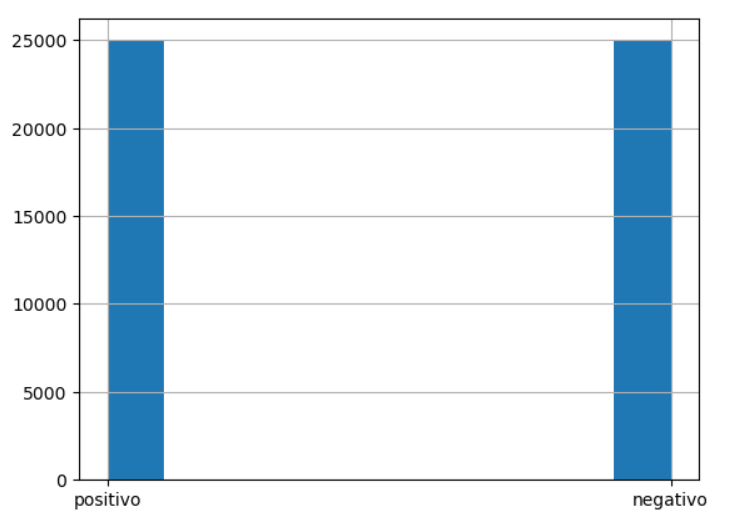
### Visualización de Resultados:

* **Gráficos y Visualizaciones:** Se utilizaron seaborn y matplotlib para crear visualizaciones claras y detalladas de los resultados, lo que facilitó la interpretación y presentación de los hallazgos.

## Tablas NLP

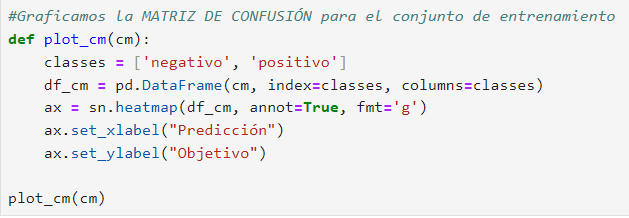
A lo largo del desarrollo del trabajo se recopilaron gran variedad de tablas y figuras que ayudan a comprender y dar mayor claridad al momento de trabajar con el proyecto, aclarando que se trabajaron con 2 bases de datos distintas. Las figuras y tablas recopiladas son:

### Figura 1: Base de datos # 1 Caso películas, histograma de datos.



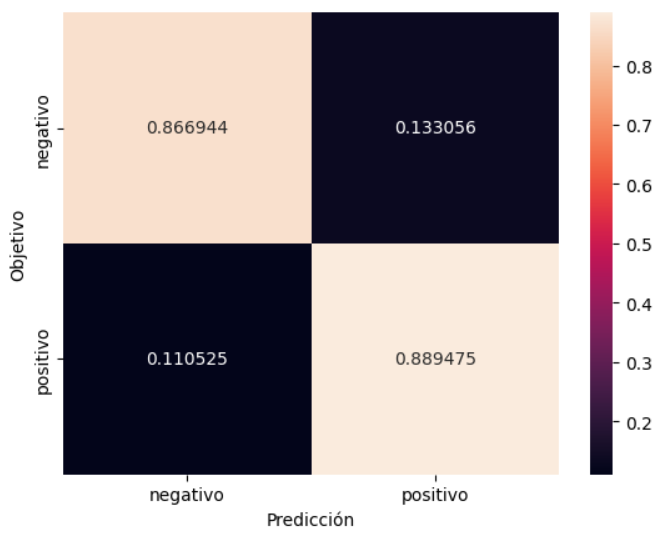
Este histograma ayudar a conocer el comportamiento de los sentimientos ya que distribuye la totalidad de los comentarios en su respectivo sentimiento.

### Figura 2: Base de datos # 1 Caso películas, función de matriz confusión.



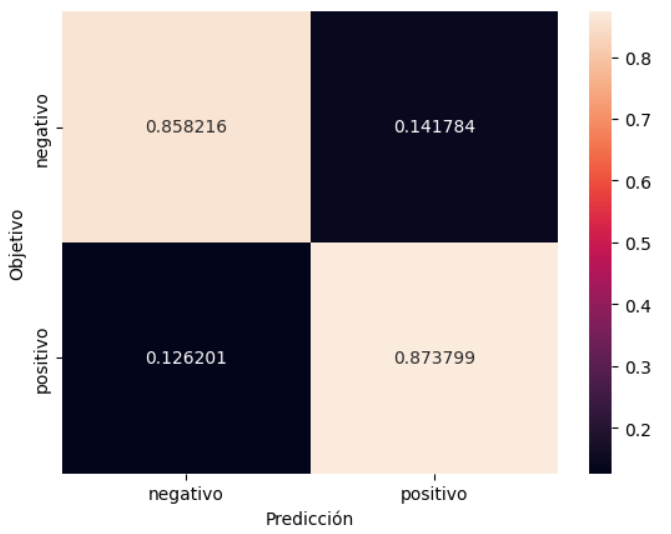
Esta función crea la matriz de confusión la cual permite leer de forma más fácil el rendimiento del algoritmo. Esta matriz se hace tanto para el subconjunto de entrenamiento como el subconjunto de prueba.

### Figura 3: Base de datos # 1 Caso películas, Matriz de confusión subconjunto de entrenamiento.



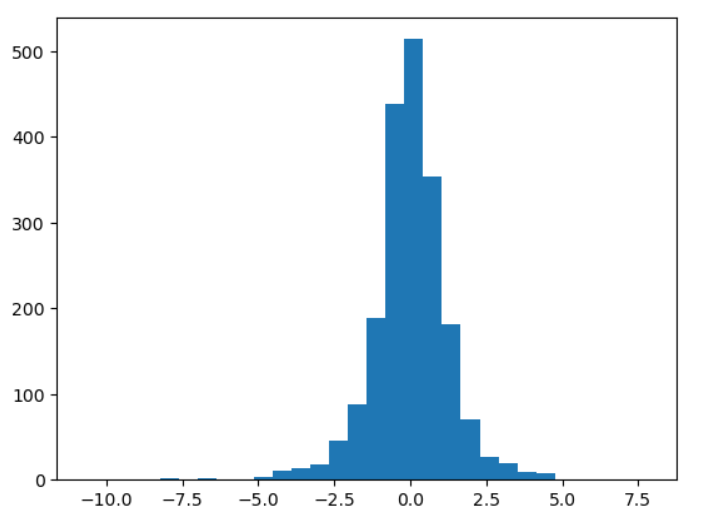
Esta figura permite ver gráficamente como fue el comportamiento del algoritmo para el subconjunto de entrenamiento, en este caso acertó el 86,6% para los comentarios negativos, quiere decir que se equivocó un 13,3%. Para los comentarios positivos acertó un 88,9% y se equivocó un 11% de las veces. Estos resultados dan como conclusión un rendimiento superior al 86% de efectividad.

### Figura 4: Base de datos # 1 Caso películas, Matriz de confusión subconjunto de prueba.



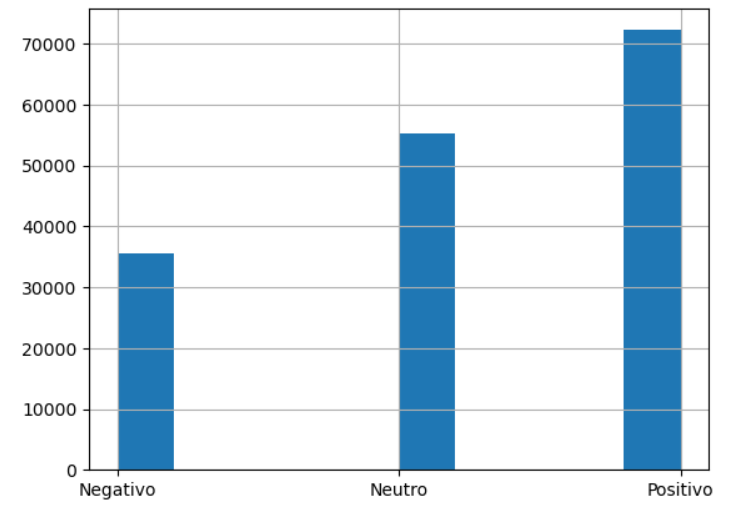
Esta figura permite ver gráficamente como fue el comportamiento del algoritmo para el subconjunto de prueba, en este caso acertó el 85,8% para los comentarios negativos, quiere decir que se equivocó un 14,1%. Para los comentarios positivos acertó un 87,3% y se equivocó un 12,6% de las veces. Estos resultados dan como conclusión un rendimiento superior al 85% de efectividad. Otra conclusión que dejan la Figura 3 y la Figura 4, es que, al trabajar con diferentes volúmenes de datos, la diferencia de efectividad del algoritmo es equivalente a un 1% en este caso.

### Figura 5: Base de datos # 1 Caso películas, Histograma de coeficientes.



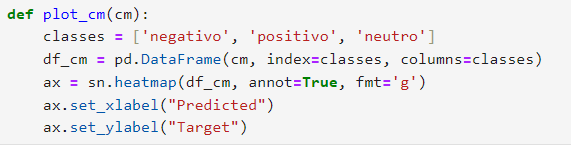
Este histograma permite ver los coeficientes de las palabras más populares utilizadas en los comentarios analizados. Esto ayudo dando conocer cuáles son las palabras más positivas y negativas, lo cual posteriormente ayudaba facilitando la clasificación del comentario.

### Figura 6: Base de datos # 2 Caso Twitter, histograma de datos.



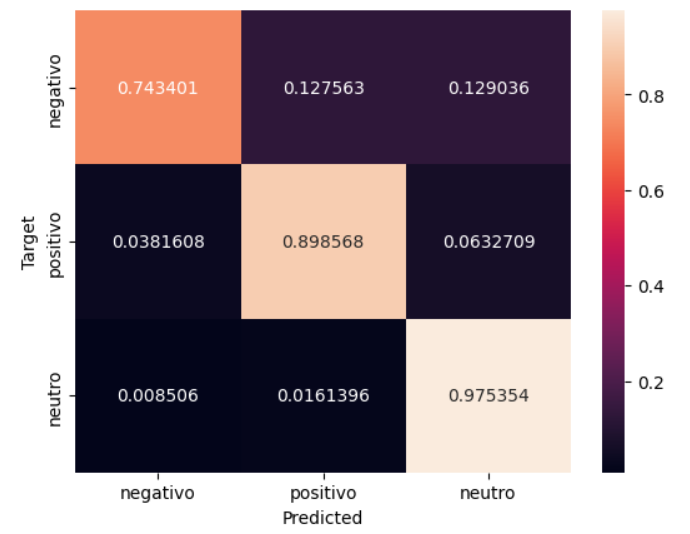
Este histograma ayudar a conocer el comportamiento de los sentimientos ya que distribuye la totalidad de los comentarios en su respectivo sentimiento. En este caso con un sentimiento nuevo el cual es “Neutro”.

### Figura 7: Base de datos # 2 Caso Twitter, función de matriz confusión.



Esta función crea la matriz de confusión la cual permite leer de forma más fácil el rendimiento del algoritmo. Esta matriz se hace tanto para el subconjunto de entrenamiento como el subconjunto de prueba. En este caso se tiene en cuenta el sentimiento a clasificar el cual es “Neutro”.

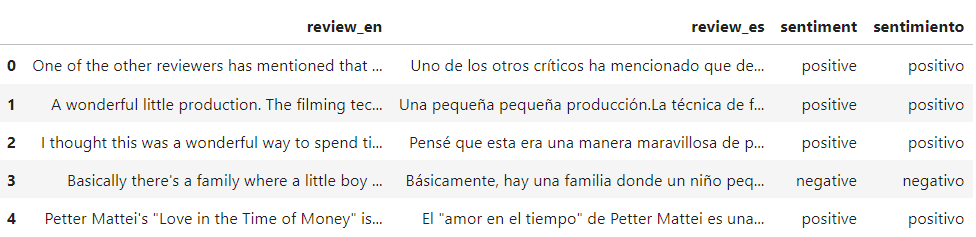
### Figura 8: Base de datos # 2 Caso Twitter, matriz de confusión.



Esta figura permite ver gráficamente como fue el comportamiento del algoritmo para el subconjunto de entrenamiento, en este caso acertó el 74,3% para los comentarios negativos, se equivocó un 12,7% diciendo que era positivo y un 12,9% diciendo que el comentario era neutro. También acertó un 89,8% para los comentarios positivos y en este momento se equivocó diciendo que el 3% de los comentarios eran negativos y el 6% eran neutros. Por último, el algoritmo acertó el 97,5% para los comentarios neutros, mientras que se equivocó un 0,8% para los comentarios negativos y un 1,6% para los comentarios positivos. Estos resultados dan como conclusión un rendimiento superior al 74% de efectividad.

## Figuras NLP

### **Tabla 1:** Base de datos # 1 Caso películas, ilustración de la base de datos.



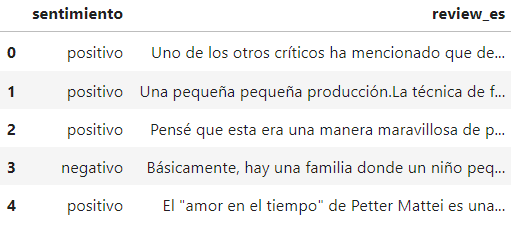
Esta tabla ilustra brevemente como está conformada la base de datos, en este caso tenemos 4 columnas; reseña en inglés, reseña en español, sentimiento en inglés y sentimiento en español.

### **Tabla 2:** Base de datos # 1 Caso películas, totalidad de la base de datos.



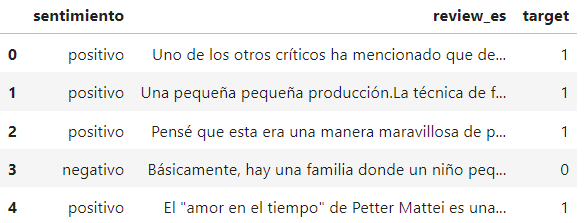
**Esta tabla permite ver la totalidad de dimensiones de la base de datos con la que se va a trabajar, en este caso 50.000 filas por 4 columnas.**

### **Tabla 3:** Base de datos # 1 Caso películas, base de datos depurada.



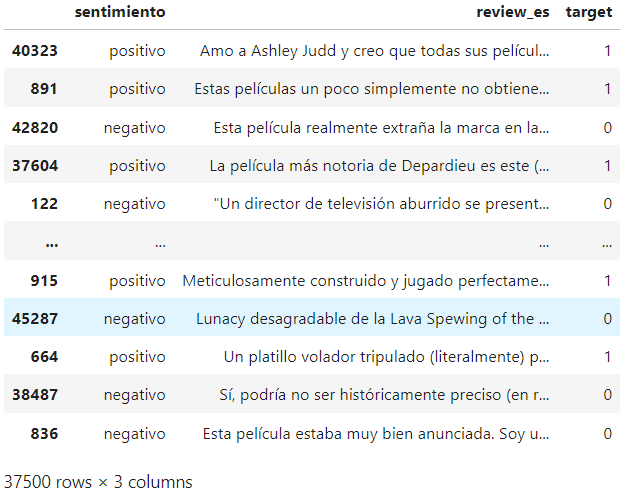
Esta tabla contiene los datos con los que se va a trabajar después de hacer una limpieza en la base datos, quedando así con solamente 2 columnas; sentimiento en español y reseña en español.

****Tabla 4:** Base de datos # 1 Caso películas, Plenitud base de datos.**



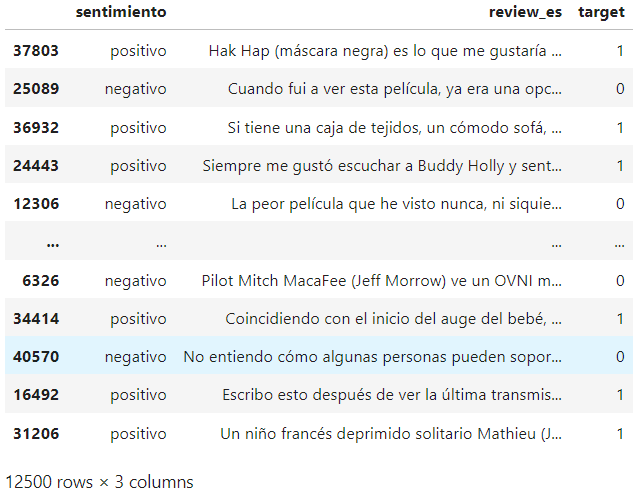
Esta tabla contiene la base datos definitiva con la cual se va a trabajar, después de agregar una nueva columna llamada “target”, la cual tiene como función asignar el valor 1 si el sentimiento es positivo y 0 si es negativo.

### **Tabla 5:** Base de datos # 1 Caso películas, base de datos del subconjunto de entrenamiento.



Esta tabla contiene los datos con los cuales va a trabajar el subconjunto de entrenamiento para el desarrollo de la matriz de confusión, en este caso está conformada por 37.500 filas por 3 columnas.

### **Tabla 6:** Base de datos # 1 Caso películas, base de datos del subconjunto de prueba.

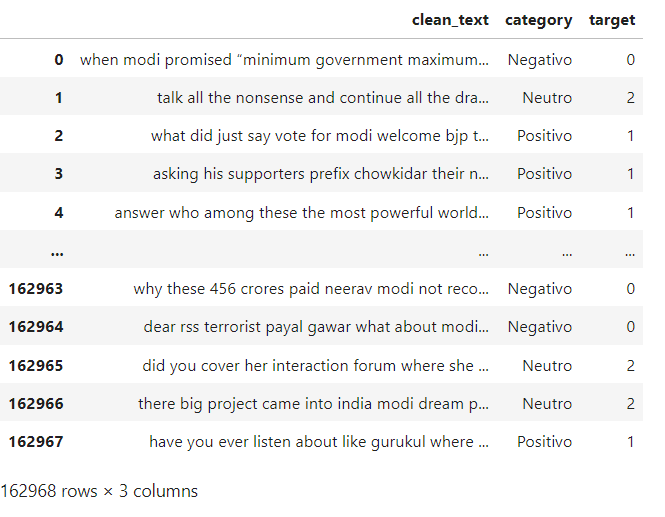
Esta tabla contiene los datos con los cuales va a trabajar el subconjunto de prueba para el desarrollo de la matriz de confusión, en este caso está conformada por 12.500 filas por 3 columnas.

### **Tabla 7:** Base de datos # 2 Caso Twitter, totalidad de la base de datos.



**Esta tabla permite ver la totalidad de dimensiones de la base de datos con la que se va a trabajar, en este caso 162.968 filas por 2 columnas.**

### **Tabla 8:** Base de datos # 2 Caso Twitter, plenitud de la base de datos.



Esta tabla contiene la base datos definitiva con la cual se va a trabajar, después de agregar una nueva columna llamada “target”, la cual tiene como función asignar el valor 0 si el sentimiento es negativo, 1 si es positivo y 2 si es neutro.

### **Tabla 9:** Base de datos # 2 Caso Twitter, base de datos del subconjunto de entrenamiento.



Esta tabla contiene los datos con los cuales va a trabajar el subconjunto de entrenamiento para el desarrollo de la matriz de confusión, en este caso está conformada por 122.226 filas por 3 columnas.

# Revisión de literatura

En el desarrollo de mi trabajo de grado, la revisión de la literatura y la bibliografía desempeñaron un papel fundamental en la construcción del marco teórico y metodológico. A continuación, se detallan las principales fuentes y contribuciones que guiaron el proyecto:

* Análisis de Sentimientos

El análisis de sentimientos es una subdisciplina del procesamiento de lenguaje natural (NLP) que se enfoca en la identificación y extracción de opiniones subjetivas en textos. Esta área ha sido ampliamente estudiada debido a su relevancia en aplicaciones como la minería de opiniones, el análisis de redes sociales, y la retroalimentación de clientes.

* Técnicas y Algoritmos

Liu (2012) en su obra "Sentiment Analysis and Opinion Mining" proporciona una visión completa sobre los métodos y aplicaciones del análisis de sentimientos. Liu destaca la importancia de entender las opiniones de los usuarios para mejorar productos y servicios. Este trabajo fue fundamental para comprender las bases teóricas y prácticas del análisis de sentimientos.

Medhat et al. (2014) en su revisión titulada "Sentiment analysis algorithms and applications: A survey" exploran los diferentes algoritmos y enfoques utilizados en el análisis de sentimientos, desde métodos basados en léxico hasta técnicas de aprendizaje automático. Esta revisión fue crucial para la selección de las técnicas a utilizar en mi proyecto.

* Herramientas y Bibliotecas

Hutto y Gilbert (2014) desarrollaron VADER (Valence Aware Dictionary and sEntiment Reasoner), un modelo basado en reglas diseñado específicamente para el análisis de sentimientos en redes sociales. Este modelo se caracteriza por su precisión en textos cortos y ruidosos, como los tweets, lo cual fue particularmente relevante para mi proyecto.

El uso de bibliotecas como NLTK, Scikit-Learn, y Pandas fue esencial en el desarrollo del análisis de sentimientos. Estas herramientas proporcionaron las funcionalidades necesarias para el preprocesamiento de datos, la tokenización, la vectorización y la implementación de modelos de clasificación.

* Procesamiento de Lenguaje Natural (NLP)

El campo del NLP ha experimentado un rápido crecimiento y evolución, impulsado por la disponibilidad de grandes volúmenes de datos textuales y el avance de las técnicas de aprendizaje automático. En este contexto, mi proyecto se apoyó en varios estudios y recursos clave.

* Modelos de Representación de Texto

Mikolov et al. (2013) introdujeron Word2Vec, un modelo de representación vectorial de palabras que ha revolucionado la forma en que los textos se procesan y analizan. Este modelo permite representar palabras en espacios vectoriales, capturando sus similitudes semánticas. La implementación de Word2Vec fue crucial para mejorar la precisión de los modelos de clasificación en mi proyecto.

* Preprocesamiento de Datos

Bird et al. (2009) en "Natural Language Processing with Python" proporcionan una guía exhaustiva sobre técnicas de preprocesamiento de datos textuales, incluyendo tokenización, lematización y eliminación de stopwords. Estas técnicas fueron implementadas en mi proyecto para preparar los datos antes de su análisis.

Cada paso de mi trabajo incorporo estos aprendizajes, desde la revisión de literatura, donde establecí las brechas existentes, hasta la metodología, donde apliqué las técnicas identificadas para crear un algoritmo funcional, demostrando la aplicabilidad y eficacia del algoritmo desarrollado en entornos reales.

# Desarrollo

#### La recopilación de datos se realizó desde dos fuentes principales:

## Recopilación de los Datos

Para la recopilación de datos en mi trabajo de grado, se utilizaron dos bases de datos principales que proporcionaron una diversidad de opiniones y sentimientos: reseñas de películas de una página web y tweets de Twitter. A continuación, detallo el proceso de obtención de cada una:

### Reseñas de Películas:

* + Fuente: La base de datos de reseñas de películas fue extraída de un sitio web conocido por su amplia colección de opiniones de usuarios sobre diversas películas.
  + Proceso de Extracción: Utilicé técnicas de web scraping para recolectar estas reseñas. Mediante el uso de la biblioteca BeautifulSoup de Python, se accedió a las páginas web, se navegó por el contenido HTML y se extrajeron las reseñas relevantes junto con las calificaciones asociadas.
  + Formato de los Datos: Las reseñas obtenidas incluían texto de opiniones, los cuales se les asignaron “tarjetas” para también darles un valor numérico, lo que permitió realizar un análisis cuantitativo y cualitativo de los sentimientos.

### Tweets de Twitter:

* + Fuente: Twitter, una plataforma ampliamente utilizada, proporcionó una rica fuente de datos textuales en forma de tweets.
  + Proceso de Extracción: Utilicé la API de Twitter para acceder y recopilar tweets. Mediante el uso de tweepy, una biblioteca de Python para trabajar con la API de Twitter, se configuraron consultas específicas para recolectar tweets que contenían ciertos hashtags o palabras clave relacionadas con temas de interés.
  + **Consideraciones de Preprocesamiento**: Los tweets recopilados incluían datos adicionales como la fecha de publicación, el número de retweets y likes, y metadatos del usuario. Se realizó un preprocesamiento adicional para eliminar enlaces, menciones, hashtags y otros elementos no textuales que no aportaban al análisis de sentimientos.

Razón para esta Metodología: La combinación de reseñas de películas y tweets proporcionó una base de datos diversa y equilibrada. Las reseñas de películas ofrecieron un contexto estructurado y detallado de las opiniones de los usuarios, mientras que los tweets proporcionaron opiniones más espontáneas y variadas en tiempo real. Esta dualidad permitió desarrollar un algoritmo de análisis de sentimientos que pudiera generalizar bien en diferentes contextos y tipos de texto.

## **Preprocesamiento de Datos**

**El preprocesamiento es una etapa crucial que incluye:**

* Eliminación de Ruido: **Remover caracteres especiales, enlaces y otros elementos no textuales.**
* Manejo de Valores Nulos: **Tratamiento de datos faltantes o incompletos.**
* Normalización del Texto: **Convertir todo el texto a minúsculas y normalizar las palabras.**

## **Desarrollo del Algoritmo**

**El desarrollo del algoritmo involucró varias etapas:**

* Tokenización y Lematización: **Utilizando bibliotecas como NLTK para dividir el texto en tokens y reducir las palabras a su forma base.**
* Vectorización: **Transformar el texto en vectores numéricos utilizando TF-IDF.**
* Entrenamiento del Modelo: **Implementar modelos de clasificación con la biblioteca Scikit-Learn. Se probaron diferentes algoritmos como la Regresión Logística y Naive Bayes para determinar cuál ofrecía mejores resultados.**
* Evaluación del Modelo: **Utilizar métricas como la precisión, Recall, F1 Score y ROC AUC para evaluar el desempeño del modelo. La evaluación se realizó mediante validación cruzada para asegurar la robustez del modelo.**

## **Visualización de Resultados**

**La visualización de datos es fundamental para interpretar los resultados. Se utilizaron bibliotecas como Seaborn y Matplotlib para crear gráficos que representen la distribución de sentimientos y los resultados de clasificación.**

## **Optimización y Validación**

**Se optimizó el modelo ajustando hiperparámetros y utilizando técnicas de validación cruzada. Esto aseguró que el modelo fuera robusto y pudiera generalizar bien a datos no vistos.**

# Conclusiones y recomendaciones

## Conclusiones

Al culminar mi trabajo de grado se han derivado varias conclusiones fundamentales que subrayan tanto los logros alcanzados como las áreas de mejora y futuras investigaciones. A continuación, se presentan las conclusiones más relevantes:

* Importancia del Preprocesamiento de Datos: **El preprocesamiento adecuado de los datos, incluyendo la limpieza, tokenización y eliminación de ruido, es crucial para mejorar la precisión de los modelos de análisis de sentimientos. Sin un preprocesamiento riguroso, los resultados pueden ser significativamente menos precisos.**
* Comparación de Algoritmos de Clasificación: **Al comparar diferentes algoritmos de clasificación, como la Regresión Logística y Naive Bayes, se encontró que la Regresión Logística ofreció una mejor precisión general en la identificación de sentimientos positivos, negativos y neutros.**
* Relevancia de la Visualización de Datos: **La visualización de datos con herramientas como Seaborn y Matplotlib resultó esencial para identificar patrones y tendencias en los datos. Estas visualizaciones no solo facilitaron la comprensión de los resultados, sino que también mejoraron la comunicación de los hallazgos.**
* Uso de Métricas de Evaluación: **Las métricas de evaluación como el F1 Score y el ROC AUC fueron fundamentales para medir la eficacia y precisión de los modelos de clasificación. Estas métricas proporcionaron una visión clara de las fortalezas y debilidades de cada modelo, permitiendo ajustes y optimizaciones adicionales.**
* Aplicabilidad en Distintos Contextos: **El análisis de sentimientos demostró ser una herramienta poderosa y versátil aplicable en diversos contextos, desde la retroalimentación de reseñas hasta el monitoreo de redes sociales. La capacidad de entender y analizar opiniones en tiempo real ofrece una ventaja competitiva significativa.**
* Desafíos en la Recolección de Datos: **La recopilación de datos de fuentes como reseñas de películas y tweets presentó desafíos en términos de acceso y limpieza. La implementación de técnicas de scraping y el uso de APIs específicas fueron necesarios para obtener datos relevantes y de calidad.**
* Impacto del Tamaño del Conjunto de Datos: **El tamaño y la diversidad del conjunto de datos impactan directamente en la precisión del modelo. Conjuntos de datos más grandes y variados tienden a proporcionar modelos más robustos y generalizables.**

## Recomendaciones

### Ampliación del Conjunto de Datos:

Para futuros trabajos, recomiendo ampliar el conjunto de datos utilizados en el análisis de sentimientos. Esto podría incluir la incorporación de reseñas de otras plataformas de medios sociales, blogs, foros y sitios de reseñas especializadas en diferentes dominios. La diversidad en las fuentes de datos puede mejorar la generalización del modelo, permitiendo una mejor interpretación de los sentimientos en distintos contextos y asegurando una mayor representatividad y precisión del análisis.

* + - Optimización y Evaluación Continua del Modelo:

Recomiendo establecer un proceso continuo de optimización y evaluación del modelo para asegurar su efectividad a lo largo del tiempo. Esto incluye la implementación de un sistema de monitoreo para rastrear el rendimiento del modelo en tiempo real, la actualización periódica de los datos de entrenamiento y la reevaluación de los parámetros del modelo. Además, utilizar métricas como el F1-Score, ROC AUC y el análisis de matriz de confusión puede ayudar a identificar áreas de mejora y asegurar que el modelo mantenga su precisión y consistencia en la identificación de sentimientos positivos, negativos y neutros.

### Mejora en el Preprocesamiento de Datos:

Una recomendación importante es mejorar el proceso de preprocesamiento de los datos. Al recomendar ampliar los conjuntos de datos también es importante implementar técnicas avanzadas de limpieza de datos, como la eliminación de ruido y la corrección de errores ortográficos, ya que esto puede aumentar significativamente la calidad de los datos de entrada y, por lo tanto, la precisión del modelo de análisis de sentimientos. Además, considerar el uso de técnicas de normalización de texto y análisis de lenguaje específico de dominio puede ofrecer un análisis más preciso y contextualizado, lo que es crucial para interpretar correctamente los sentimientos expresados en diferentes contextos.

# 

# 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.*