



**REPORTE 1 - PROYECTO FINAL
CATEGORIZACIÓN Y ANÁLISIS DE SENTIMIENTOS
DE ARTÍCULOS DE NOTICIAS**

GRUPO 16

Andrés Felipe Gualdrón Gutiérrez
Juan Manzano Barona
Jersson Hernán Morales Hernández
Carina Lizebeth Ordoñez Araque

1 INTRODUCCIÓN

Este proyecto propone un enfoque innovador que integra técnicas de aprendizaje automático supervisado y no supervisado para la categorización y re-categorización de noticias, complementado por el uso del análisis de sentimientos mediante VADER (Valence Aware Dictionary and sEntiment Reasoning). Esta metodología se presenta como una herramienta robusta y adaptable para navegar el complejo panorama informativo contemporáneo.

El sistema resultante ofrecerá a los usuarios una visión integral y matizada del entorno de noticias en constante evolución, facilitando una comprensión más profunda de las diversas categorías de artículos. Al utilizar un amplio conjunto de datos que abarca 210,000 artículos y sus respectivas categorías, junto con el análisis de sentimientos proporcionado por VADER, este proyecto busca ofrecer insights más precisos y detallados.

Con el fin de mitigar el impacto de la polarización en la opinión pública, esta propuesta permitirá a las partes interesadas tomar decisiones más informadas y promover una difusión de noticias más equilibrada y consciente. La combinación de métodos de aprendizaje automático y análisis de sentimientos no solo optimiza el filtrado de información relevante, sino que también contribuye a un consumo informativo más equilibrado y reflexivo, esencial en un contexto mediático tan dinámico y desafiante.

2 PROBLEMA Y SU CONTEXTO

El acceso a la información por parte de la población ha tenido un gran cambio durante las últimas décadas, en las que se han ido desarrollando mecanismos que facilitan el acceso a los datos, noticias e información general de una forma mucho más directa y rápida. Las nuevas tecnologías, la creación y desarrollo de dispositivos móviles y el gran avance en la conectividad han propiciado un entorno de creación y difusión de todo tipo de contenidos de una forma inmediata y global, lo que supone un reto importante en cuanto a la filtración de la información a la que se accede.

Una de las consecuencias de este modelo, que se ha reconocido como una problemática social en muchos países, está relacionada con la polarización de la opinión pública. La proliferación de creadores y difusores de contenidos ha contribuido a alimentar esta polarización mucho más allá de la política o la religión, llegando a prácticamente cualquier aspecto social, cultural o deportivo.

En este contexto, cobra una especial relevancia el filtrado de la información a la que accedemos y, para ello, la validación y el refinamiento de la categorización de los artículos de noticias y opinión se convierten en un aspecto fundamental.

3 PREGUNTA DE NEGOCIO Y ALCANCE DEL PROYECTO

Este proyecto tiene como objetivo el desarrollo de un sistema automatizado que permita la validación, el refinamiento y una posible re-categorización de artículos de noticias, utilizando como fuente una base de datos del portal HuffPost que incluye, aproximadamente, 210.000 artículos.

Este sistema se llevará a cabo mediante la realización de un análisis de sentimientos o minería de opiniones que permite identificar el tono emocional presente en un texto. Para ello, se utilizará un enfoque híbrido de aprendizaje supervisado y no supervisado para realizar la categorización de la información, además del algoritmo VADER (Valence Aware Dictionary for Sentiment Reasoning) para el análisis de sentimientos.

3.1 Metodología

- Categorización y Re-categorización de Noticias
 - Análisis No Supervisado:

- Implementar técnicas de modelado de temas como Latent Dirichlet Allocation (LDA) o Non-Negative Matrix Factorization (NMF).
- Descubrir patrones temáticos latentes en el conjunto de datos.
- Identificar posibles subcategorías y temas emergentes no capturados por las categorías existentes.
- Análisis Supervisado:
 - Utilizar las etiquetas de categoría existentes para entrenar un modelo de clasificación supervisada (por ejemplo, SVM, Random Forest, o un modelo basado en BERT).
 - Evaluar la precisión del modelo en comparación con las categorías existentes.
- Integración y Refinamiento:
 - Comparar los resultados del análisis no supervisado con las categorías existentes y las predicciones del modelo supervisado.
 - Identificar discrepancias y posibles áreas de mejora en la categorización.
 - Re-entrenar el modelo supervisado con la estructura de categorías refinada.
- Análisis de Sentimientos
 - Implementar VADER (Valence Aware Dictionary and sEntiment Reasoner) para el análisis de sentimientos.
 - Aplicar VADER a cada artículo para obtener puntuaciones de sentimiento (positivo, negativo, neutral).
 - Agregar los resultados de sentimiento por categoría y a lo largo del tiempo.
- Desarrollo de un Panel de Control
 - Crear un panel interactivo para visualizar:
 - Distribución y evolución de categorías y subcategorías identificadas
 - Tendencias de sentimiento para diferentes categorías de noticias utilizando los resultados de VADER
 - Artículos clave para temas específicos
 - Análisis de frecuencia de palabras por categoría y subcategoría
 - Indicadores de polarización basados en la distribución de sentimientos de VADER

3.2 Aplicaciones Potenciales

- Asistir a los medios de comunicación en la comprensión de temas tendencia y el sentimiento público, en general, con una categorización más precisa.
- Ayudar a los legisladores a medir el sentimiento público sobre diversos temas y a desarrollar políticas para mitigar la polarización, respaldados por análisis de sentimientos adaptados al contexto noticioso.
- Ayudar a las empresas a monitorear el sentimiento de las noticias relacionadas con su industria a través de fuentes de noticias, con una comprensión de las categorías y subcategorías relevantes.

4 CONJUNTO DE DATOS

Los datos corresponden a artículos de noticias del portal HuffPost en Estados Unidos que fueron recopilados entre 2012 y 2022. El dataset cuenta con cerca de 210 mil titulares de noticias (HEADLINE) junto con:

- CATEGORY: Categoría en la que el artículo fue publicado
- SHORT_DESCRIPTION: Resumen del artículo
- AUTHORS: Lista de los autores quienes contribuyeron en el artículo
- DATE: Fecha de publicación del artículo
- LINK: Hipervínculo del artículo original de la noticia

Los datos están contenidos en un archivo tipo *.JSON, de 87.3MB, disponible para acceso público permanentemente en la plataforma Kaggle con las siguientes citas:

- <https://www.kaggle.com/datasets/rmisra/news-category-dataset>
- Misra, Rishabh. "News Category Dataset." arXiv preprint arXiv: 2209.11429 (2022).
- Misra, Rishabh and Jigyasa Grover. "Sculpting Data for ML: The first act of Machine Learning." ISBN 9798585463570 (2021).

5 EXPLORACION DE DATOS

5.1 Análisis de Valores Nulos y Duplicados

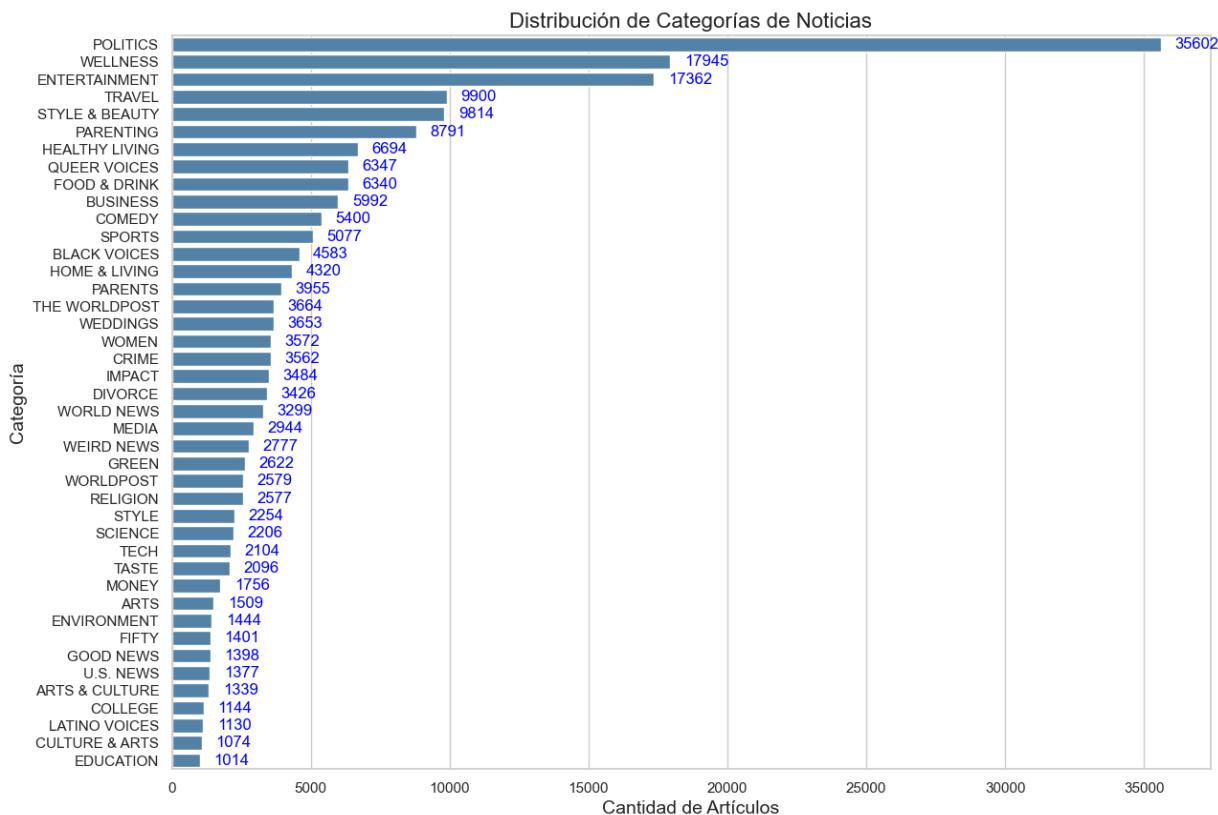
- Valores Nulos: No se encontraron valores nulos en ninguna de las columnas.
- Duplicados: Se identificaron 13 registros duplicados en la base de datos.

5.2 Categorías de las Noticias

Se obtuvieron 42 categorías entre las que se encuentran principalmente: política (POLITICS), bienestar (WELLNESS), entretenimiento (ENTERTAINMENT), viajes (TRAVEL), y estilo y belleza (STYLE & BEAUTY).

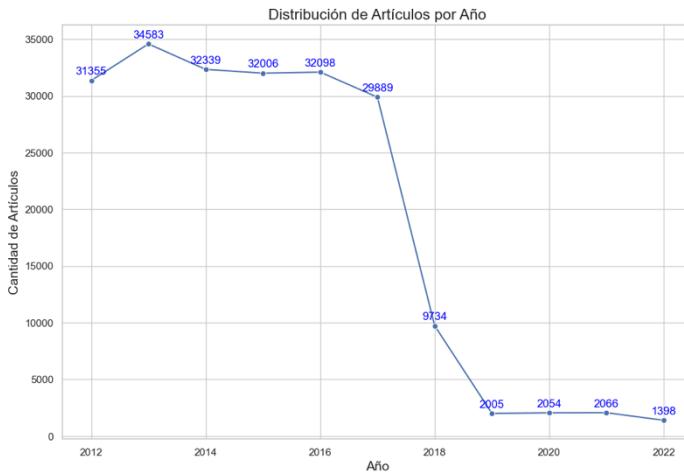
5.3 Distribución de Categorías de Noticias

Se realizó un análisis de la distribución de categorías de noticias para identificar la cantidad de artículos por categorías. POLITICS es la categoría con más artículos (35,602), seguida por WELLNESS (17,945) y ENTERTAINMENT (17,362). Entre las categorías menos frecuentes se encuentran EDUCATION, CULTURE & ARTS, y COLLEGE, cada una con poco más de 1,000 artículos.



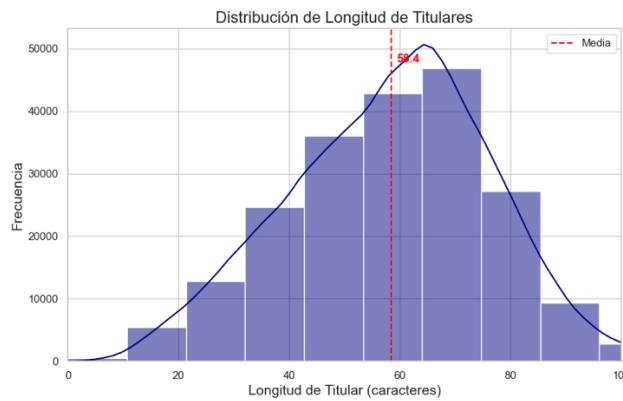
5.4 Distribución de Artículos por Año

Se llevó a cabo un análisis de la Distribución de Artículos por Año. Se observa un volumen constante de publicaciones entre 2012 y 2018, con más de 30,000 artículos por año. A partir de 2019, hay una notable disminución en la cantidad de publicaciones, alcanzando menos de 3,000 artículos por año desde el año 2020.



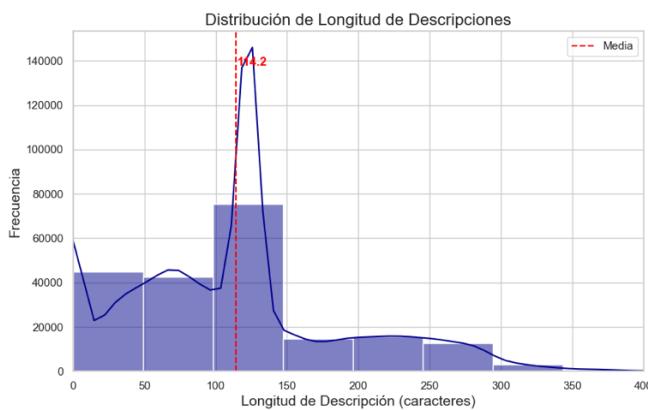
5.5 Distribución de la Longitud de Titulares

El análisis de la Distribución de la Longitud de Titulares muestra que la longitud promedio de los titulares es de 58.4 caracteres, con la mayoría de ellos en un rango de 40 a 70 caracteres, lo cual indica un estilo conciso y consistente.



5.6 Distribución de la Longitud de Descripciones

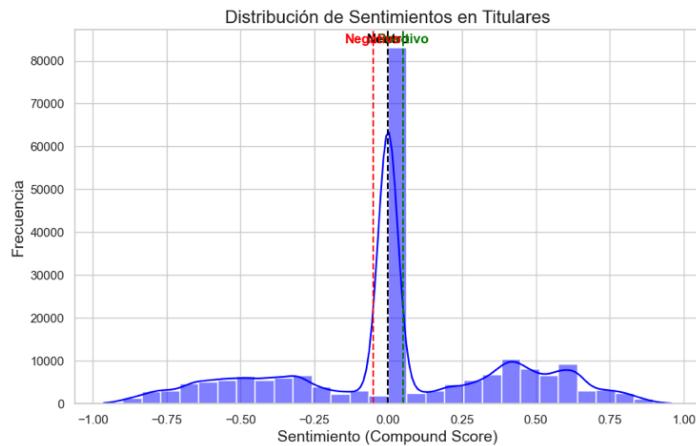
En el análisis de la distribución de la longitud de descripciones se identificó que la longitud promedio de las descripciones es de 114.2 caracteres. La distribución muestra una alta concentración alrededor de los 100-150 caracteres, aunque también existe un volumen considerable de descripciones más breves.



5.7 Análisis de Sentimientos en Titulares

Se realizó un Análisis de Sentimientos preliminar a los titulares (HEADLINES), revelando que la mayoría de ellos presentan un tono cercano al neutro. La mayoría de los titulares muestran una puntuación neutra en el "COMPOUND SCORE".

- Distribución: Existe una tendencia hacia sentimientos neutros, con pocos titulares extremadamente positivos o negativos.
- Sentimiento por Categoría: Las categorías como CRIME y WORLD NEWS presentan un sentimiento promedio negativo, mientras que GOOD NEWS y HOME & LIVING tienden hacia sentimientos más positivos.



6 REPOSITORIOS

6.1 GIT

El repositorio Git creado para almacenar el contenido referente al proyecto final: **CATEGORIZACIÓN Y ANÁLISIS DE SENTIMIENTOS DE ARTÍCULOS DE NOTICIAS**, y para realizar el registro del versionamiento tanto del código desarrollado para el despliegue de la herramienta como para los datos usados en el entrenamiento y prueba del modelo, se puede acceder en el siguiente vínculo:

https://github.com/jhermoher/miad2024_dsa_G16.git

El repositorio fue creado desde el portal Git en la cuenta de uno de los integrantes del equipo (refiérase al Anexo -1). El repositorio para el proyecto se creó desde una máquina local con la siguiente estructura:

```
proyecto/
    ├── .dvc/          # Versionamiento de los datos
    ├── data/          # Datos y documentación relacionada
    ├── dashboard/     # Código del panel de control
    ├── docs/          # Documentación adicional
    ├── notebooks/     # Jupyter notebooks con análisis
    ├── src/           # Código fuente del proyecto
    ├── submittals/    # Reportes y entregables
    ├── tests/          # Pruebas unitarias y de integración
    └── requirements.txt # Dependencias del proyecto
```

Los soportes de la estructuración del repositorio se muestran a continuación:

Creación de las carpetas y primer -commit- para estructurar el repositorio.

```
Last login: Sat Oct 26 16:49:39 on console
[jerssonn@Jerssons-MacBook-Pro miad2024_dsa_G16 % mkdir data
[jerssonn@Jerssons-MacBook-Pro miad2024_dsa_G16 % mkdir notebooks
[jerssonn@Jerssons-MacBook-Pro miad2024_dsa_G16 % mkdir src
[jerssonn@Jerssons-MacBook-Pro miad2024_dsa_G16 % mkdir dashboard
[jerssonn@Jerssons-MacBook-Pro miad2024_dsa_G16 % mkdir docs
[jerssonn@Jerssons-MacBook-Pro miad2024_dsa_G16 % mkdir test
[jerssonn@Jerssons-MacBook-Pro miad2024_dsa_G16 % mkdir submittals
[jerssonn@Jerssons-MacBook-Pro miad2024_dsa_G16 % echo "# My Project" > README.md
[touch .gitignore
[jerssonn@Jerssons-MacBook-Pro miad2024_dsa_G16 % git add .
[jerssonn@Jerssons-MacBook-Pro miad2024_dsa_G16 % git commit -m "Initial project setup"
[main 7615fdb] Initial project setup
  2 files changed, 1 insertion(+), 42 deletions(-)
  create mode 100644 .gitignore
```

Estructuración del repositorio.

```
[jerssom@Jerssons-MacBook-Pro miad2024_dsa_G16 % git push -f origin main
Enumerating objects: 8, done.
Counting objects: 100% (8/8), done.
Delta compression using up to 8 threads
Compressing objects: 100% (4/4), done.
Writing objects: 100% (6/6), 3.27 KiB | 3.27 MiB/s, done.
Total 6 (delta 0), reused 0 (delta 0), pack-reused 0
To https://github.com/jhermoher/miad2024_dsa_G16.git
 + 7615fdb...dc86448 main -> main (forced update)
[jerssom@Jerssons-MacBook-Pro miad2024_dsa_G16 % git status
On branch main
Your branch is up to date with 'origin/main'.

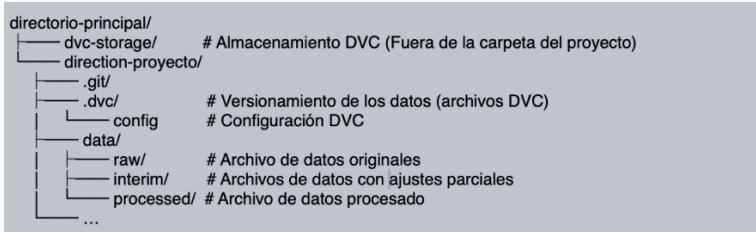
nothing to commit, working tree clean
[jerssom@Jerssons-MacBook-Pro miad2024_dsa_G16 % for dir in */; do touch "$dir/.gitkeep"; done
[jerssom@Jerssons-MacBook-Pro miad2024_dsa_G16 % git add .
[jerssom@Jerssons-MacBook-Pro miad2024_dsa_G16 % git commit -m "Add project structure with folders"
[main 9559765] Add project structure with folders
7 files changed, 0 insertions(+), 0 deletions(-)
create mode 100644 dashboard/.gitkeep
create mode 100644 data/.gitkeep
create mode 100644 docs/.gitkeep
create mode 100644 notebooks/.gitkeep
create mode 100644 src/.gitkeep
create mode 100644 submittals/.gitkeep
create mode 100644 test/.gitkeep
[jerssom@Jerssons-MacBook-Pro miad2024_dsa_G16 % git push origin main
Enumerating objects: 4, done.
Counting objects: 100% (4/4), done.
Delta compression using up to 8 threads
Compressing objects: 100% (2/2), done.
Writing objects: 100% (3/3), 431 bytes | 431.00 KiB/s, done.
Total 3 (delta 0), reused 0 (delta 0), pack-reused 0
To https://github.com/jhermoher/miad2024_dsa_G16.git
 dc86448..9559765 main -> main
jerssom@Jerssons-MacBook-Pro miad2024_dsa_G16 % ]
```

El archivo README.md, en donde se describe el proyecto final y la estructura del repositorio fue modificado de manera progresiva, directamente desde el portal.

6.2 DVC

Para el registro y almacenamiento de las versiones de los datos se creó una carpeta en un maquina local, adicionalmente, la plataforma se integró al repositorio Git creado para el proyecto, y finalmente, se creó y se configuró un servicio AWS-S3 para su acceso por los integrantes del grupo (refiérase al Anexo -2).

La estructura definida para el almacenamiento de los datos fue la siguiente:



Los soportes del proceso descrito arriba se muestran a continuación:

Creación del directorio de almacenamiento DVC y adición del archivo de datos original.

```
[(venv) jerssom@Jerssons-MacBook-Pro miad2024_dsa_G16 % cd ..
[(venv) jerssom@Jerssons-MacBook-Pro Github % mkdir dvc-storage
[(venv) jerssom@Jerssons-MacBook-Pro Github % ls
MIAD_NLP_2024          MIAD_NLP_2024_Ini      Proyectos_ml_nlp      Unsupervised_ML_2024      dvc-storage           miad2024_dsa_G16
[(venv) jerssom@Jerssons-MacBook-Pro Github % cd miad2024_dsa_G16
[(venv) jerssom@Jerssons-MacBook-Pro miad2024_dsa_G16 % dvc remote add -d mystorage ../dvc-storage
Setting 'mystorage' as a default remote.
[(venv) jerssom@Jerssons-MacBook-Pro miad2024_dsa_G16 % dvc remote list
mystorage      /Users/jerssom/Documents/GitHub/dvc-storage
[(venv) jerssom@Jerssons-MacBook-Pro miad2024_dsa_G16 % ls
README.md      dashboard      data      docs      notebooks      src      submittals      test      venv
[(venv) jerssom@Jerssons-MacBook-Pro miad2024_dsa_G16 % cd data
[(venv) jerssom@Jerssons-MacBook-Pro data % ls
interim      processed      raw
[(venv) jerssom@Jerssons-MacBook-Pro data % cd raw
[(venv) jerssom@Jerssons-MacBook-Pro raw % dvc add News_Category_Dataset_v3.json
100% Adding...|]

To track the changes with git, run:
    git add .gitignore News_Category_Dataset_v3.json.dvc

To enable auto staging, run:
    dvc config core.autostage true
[(venv) jerssom@Jerssons-MacBook-Pro raw % dvc status
Data and pipelines are up to date.
```

-Commit- del repositorio DVC en el repositorio Git del proyecto

```
(venv) jerssonn@Jerssons-MacBook-Pro raw % git add data/raw/News_Category_Dataset_v3.json.dvc
warning: could not open directory 'data/raw/data/raw': No such file or directory
fatal: pathspec 'data/raw/News_Category_Dataset_v3.json.dvc' did not match any files
(venv) jerssonn@Jerssons-MacBook-Pro raw % cd ..
(venv) jerssonn@Jerssons-MacBook-Pro data % cd ..
(venv) jerssonn@Jerssons-MacBook-Pro miad2024_dsa_G16 % git add data/raw/News_Category_Dataset_v3.json.dvc
(venv) jerssonn@Jerssons-MacBook-Pro miad2024_dsa_G16 % git add .gitignore
(venv) jerssonn@Jerssons-MacBook-Pro miad2024_dsa_G16 % git commit -m "Add dataset tracking with DVC"
[main c67e6d6] Add dataset tracking with DVC
 1 file changed, 5 insertions(+)
 create mode 100644 data/raw/News_Category_Dataset_v3.json.dvc
(venv) jerssonn@Jerssons-MacBook-Pro miad2024_dsa_G16 % dvc push
Collecting
Pushing
1 file pushed
(venv) jerssonn@Jerssons-MacBook-Pro miad2024_dsa_G16 % git push origin main
Enumerating objects: 12, done.
Counting objects: 100% (12/12), done.
Delta compression using up to 8 threads
Compressing objects: 100% (9/9), done.
Writing objects: 100% (10/10), 1.03 KiB | 1.03 MiB/s, done.
Total 10 (delta 0), reused 0 (delta 0), pack-reused 0
remote: Resolving deltas: 100% (2/2), completed with 1 local object.
To https://github.com/jhermoher/miad2024_dsa_G16.git
 9559765..c67e6d6 main -> main
(venv) jerssonn@Jerssons-MacBook-Pro miad2024_dsa_G16 %
```

Adición de dos (2) versiones de datos y -commit- en el repositorio Git

```
(venv) jerssonn@Jerssons-MacBook-Pro miad2024_dsa_G16 % cd data
(venv) jerssonn@Jerssons-MacBook-Pro data % cd interim
(venv) jerssonn@Jerssons-MacBook-Pro interim % dvc add NoDuplicates_News_Category_Dataset.csv
100% Adding...|
```

To track the changes with git, run:

```
git add NoDuplicates_News_Category_Dataset.csv.dvc .gitignore
```

To enable auto staging, run:

```
dvc config core.autostage true
(venv) jerssonn@Jerssons-MacBook-Pro interim % dvc add Dropped_News_Category_Dataset.csv
100% Adding...|
```

To track the changes with git, run:

```
git add .gitignore Dropped_News_Category_Dataset.csv.dvc
```

To enable auto staging, run:

```
dvc config core.autostage true
(venv) jerssonn@Jerssons-MacBook-Pro interim % dvc push
Collecting
Pushing
2 files pushed
(venv) jerssonn@Jerssons-MacBook-Pro interim % cd ..
(venv) jerssonn@Jerssons-MacBook-Pro data % cd ..
(venv) jerssonn@Jerssons-MacBook-Pro miad2024_dsa_G16 % git add data/interim/NoDuplicates_News_Category_Dataset.csv.dvc
(venv) jerssonn@Jerssons-MacBook-Pro miad2024_dsa_G16 % git add data/interim/Dropped_News_Category_Dataset.csv.dvc
(venv) jerssonn@Jerssons-MacBook-Pro miad2024_dsa_G16 % git add .gitignore
(venv) jerssonn@Jerssons-MacBook-Pro miad2024_dsa_G16 % git commit -m "Add interim dataset tracking with DVC"
[main 0b290cc] Add interim dataset tracking with DVC
 2 files changed, 19 insertions(+)
 create mode 100644 data/interim/Dropped_News_Category_Dataset.csv.dvc
 create mode 100644 data/interim/NoDuplicates_News_Category_Dataset.csv.dvc
```

```
(venv) jerssonn@Jerssons-MacBook-Pro miad2024_dsa_G16 % git stash
Saved working directory and index state WIP on main: 0b290cc Add interim dataset tracking with DVC
(venv) jerssonn@Jerssons-MacBook-Pro miad2024_dsa_G16 % git pull origin main
remote: Enumerating objects: 5, done.
remote: Counting objects: 100% (5/5), done.
remote: Compressing objects: 100% (3/3), done.
remote: Total 3 (delta 0), reused 0 (delta 0), pack-reused 0 (from 0)
Unpacking objects: 100% (3/3), 999 bytes | 166.00 KiB/s, done.
From https://github.com/jhermoher/miad2024_dsa_G16
 * branch      main    -> FETCH_HEAD
   c67e6d6..be781d9 main    -> origin/main
Successfully rebased and updated refs/heads/main.
(venv) jerssonn@Jerssons-MacBook-Pro miad2024_dsa_G16 % git stash pop
On branch main
Your branch is ahead of 'origin/main' by 1 commit.
  (use "git push" to publish your local commits)

Changes not staged for commit:
  (use "git add <file>..." to update what will be committed)
  (use "git restore <file>..." to discard changes in working directory)
    modified:  .dvc/config

Untracked files:
  (use "git add <file>..." to include in what will be committed)
    data/interim/.gitignore
    data/raw/.gitignore
    venv/

no changes added to commit (use "git add" and/or "git commit -a")
Dropped refs/stash@{0} (7d7a9e4eb45fae761ea9b9fd62166960ffca97a)
(venv) jerssonn@Jerssons-MacBook-Pro miad2024_dsa_G16 % git status
On branch main
Your branch is ahead of 'origin/main' by 1 commit.
  (use "git push" to publish your local commits)

Changes not staged for commit:
  (use "git add <file>..." to update what will be committed)
  (use "git restore <file>..." to discard changes in working directory)
    modified:  .dvc/config

Untracked files:
  (use "git add <file>..." to include in what will be committed)
    data/interim/.gitignore
    data/raw/.gitignore
    venv/

no changes added to commit (use "git add" and/or "git commit -a")
(venv) jerssonn@Jerssons-MacBook-Pro miad2024_dsa_G16 % git push origin main
Enumerating objects: 8, done.
Counting objects: 100% (8/8), done.
Delta compression using up to 8 threads
Compressing objects: 100% (6/6), done.
Writing objects: 100% (6/6), 763 bytes | 763.00 KiB/s, done.
Total 6 (delta 1), reused 0 (delta 0), pack-reused 0
remote: Resolving deltas: 100% (1/1), completed with 1 local object.
To https://github.com/jhermoher/miad2024_dsa_G16.git
  be781d9..00c4dd6 main -> main
(venv) jerssonn@Jerssons-MacBook-Pro miad2024_dsa_G16 %
```

Configuración de AWS-S3: dsa-finalproject-dvc-store como repositorio remoto

```
(venv) jerssonm@Jerssons-MacBook-Pro miad2024_dsa_G16 % aws iam list-users
{
    "Users": []
}

(venv) jerssonm@Jerssons-MacBook-Pro miad2024_dsa_G16 % dvc remote add -d aws_remote s3://dsa-finalproject-dvc-store
Setting 'aws_remote' as a default remote.
```

Adición de los archivos al servicio S3 y actualización del repositorio Git

```
(venv) jerssonm@Jerssons-MacBook-Pro miad2024_dsa_G16 % dvc push
Collecting
Pushing
3 files pushed
(venv) jerssonm@Jerssons-MacBook-Pro miad2024_dsa_G16 % cd .dvc
(venv) jerssonm@Jerssons-MacBook-Pro .dvc % cat config
[core]
    remote = aws_remote
    ['remote "mystorage"']
    url = ../../dvc-storage
    ['remote "aws_remote"']
    url = s3://dsa-finalproject-dvc-store
(venv) jerssonm@Jerssons-MacBook-Pro .dvc % cd ..
zsh: command not found: cd..
(venv) jerssonm@Jerssons-MacBook-Pro .dvc % cd ..
(venv) jerssonm@Jerssons-MacBook-Pro miad2024_dsa_G16 % git commit .dvc/config -m "Adición S3 como remote"
[main f6463ec] Adición S3 como remote
 1 file changed, 6 insertions(+)
(venv) jerssonm@Jerssons-MacBook-Pro miad2024_dsa_G16 % git push origin main
Enumerating objects: 7, done.
Counting objects: 100% (7/7), done.
Delta compression using up to 8 threads
Compressing objects: 100% (4/4), done.
Writing objects: 100% (4/4), 470 bytes | 470.00 KiB/s, done.
Total 4 (delta 1), reused 0 (delta 0), pack-reused 0
remote: Resolving deltas: 100% (1/1), completed with 1 local object.
To https://github.com/jhermoher/miad2024_dsa_G16.git
 00c4d60..f6463ec main -> main
(venv) jerssonm@Jerssons-MacBook-Pro miad2024_dsa_G16 %
```

7 REPORTE TRABAJO EN EQUIPO

El primer reporte relacionado con el proyecto final del curso busca documentar el avance alcanzado en las primeras 3 semanas.

Semana 1: Durante la primera semana, formamos nuestro equipo de trabajo y realizamos una fase exploratoria inicial. Como equipo identificamos y analizamos los posibles problemas que podíamos abordar en el proyecto, asegurándonos de que para cada problema existiera una clara pregunta de negocio y un conjunto de datos disponible. Refiérase al Anexo -3.

Semana 2: En la segunda semana, tomamos la decisión final sobre el problema específico a abordar, priorizando aquellos que tenían datos inmediatamente disponibles. Definimos concretamente la pregunta de negocio, el alcance de nuestro proyecto y los conjuntos de datos que utilizaríamos. Identificamos cómo resolver la pregunta de negocio mediante visualizaciones descriptivas y un modelo predictivo (Refiérase al Plan de manejo de Datos – DMP, entregado por el grupo). También desarrollamos una maqueta inicial del prototipo donde mostramos claramente los elementos y su relación con la pregunta de negocio. Refiérase al Anexo -4.

Semana 3: En la tercera semana nos centramos en aspectos técnicos y de implementación. Establecimos nuestros repositorios en Git para el código y DVC para los datos, realizamos una exploración detallada de los datos asegurando el versionado tanto del código como de los datos. Basándonos en los hallazgos de nuestro análisis de datos, realizamos ajustes sobre la maqueta del prototipo inicial.

A diferencia de las semanas anteriores, para avanzar significativamente en las diferentes actividades de esta tercera semana, las tareas fueron distribuidas a cada integrante de la siguiente manera:

- Andrés Gualdrón. Exploración de los datos disponibles.
- Juan Manzano. Generación del reporte correspondiente a la primera entrega.
- Jersson Morales. Creación de los repositorios Git y DVC.
- Lizebeth Ordoñez. Creación del mockup del tablero.

Las mismas se realizaron completamente con responsabilidad y calidad, como se muestran en este reporte.

8 MAQUETA PROTOTIPO

Este prototipo presenta una interfaz de visualización de datos diseñada para analizar y presentar información sobre el sentimiento de artículos de noticias recopilados del portal HuffPost. Su objetivo es ofrecer a los usuarios una comprensión integral del contenido mediático y su polarización. La estructura del mockup se divide en dos secciones principales: Análisis Descriptivo y Análisis Predictivo, cada una con componentes clave que facilitan el entendimiento y la interpretación de los datos.

8.1 Análisis Descriptivo

Análisis de Sentimientos y Categorización de Noticias

Total Artículos **21 324**

Distribución Categorías **42**

Proporción de Sentimientos

12.8% -1.22%	25.2% -5.45%	62% +6.67%
------------------------	------------------------	----------------------

Distribución de Contenido por Categoría

Frecuencia de Sentimientos por Categoría

Matriz de Confusión

	Predicción Positiva	Predicción Negativa
Real Positivo	80	10
Real Negativo	5	90

Comparativa de Sentimientos entre Categorías

Detección de Noticias Polarizantes

Categoría	Fecha de Publicación	Artículo	Sentimiento	Puntaje de Sentimiento	Palabras Claves
Politics	19 May, 2021 : 10:10 AM	Trump repunta en las encuestas de Florida	Positivo	69	Satisfacción, esperanza
Crimen	18 May, 2021 : 3:12 PM	Niña de 11 años fingió estar muerta para escapar luego de	Negativo	97	Desesperación, horror

1 2 3 4 5 ... 20

Este mockup presenta una interfaz intuitiva y funcional diseñada para el análisis y la visualización del sentimiento de artículos de noticias extraídos del portal HuffPost. Su objetivo principal es proporcionar a los usuarios una comprensión profunda del contenido mediático y su polarización. La estructura del mockup está organizada en varias secciones clave:

- Indicadores Generales
 - Total de Artículos: Indica el número total de artículos analizados, proporcionando una visión general de la magnitud del conjunto de datos.
 - Distribución de Categorías: Representa la cantidad de artículos en cada categoría, facilitando el análisis del contenido en diferentes temáticas.
 - Proporción de Sentimientos: Muestra el porcentaje de artículos clasificados como positivos, neutros y negativos, junto con la variación de estos sentimientos a lo largo del tiempo.
- Distribución de Contenido por Categoría
 - Gráfica de Barras: Ilustra la cantidad de artículos por categoría, permitiendo una comparación sencilla entre las diferentes secciones de contenido. Cada categoría se presenta en un color distintivo para mejorar la claridad visual.
- Frecuencia de Sentimiento por Categoría
 - Gráfica de Barras Agrupadas: Expone la distribución de sentimientos (positivo, neutro, negativo) dentro de cada categoría, lo que permite a los usuarios comparar rápidamente las emociones asociadas a los artículos en las diversas categorías.
- Comparativa de Sentimientos entre Categorías
 - Gráfico de Líneas: Representa las tendencias de sentimiento a lo largo del tiempo para cada categoría, permitiendo a los usuarios observar cómo varían las emociones en función de diferentes temas y períodos.
- Matriz de Confusión
 - Visualización del Rendimiento: Proporciona una representación gráfica del rendimiento del modelo de análisis de sentimientos, mostrando los verdaderos positivos, falsos positivos, verdaderos negativos y falsos negativos. Esta sección incluye una breve explicación para ayudar a los usuarios a interpretar los resultados.
- Detección de Noticias Polarizantes
 - Tabla Interactiva: Incluye información sobre artículos considerados polarizantes, abarcando su categoría, título, sentimiento, puntaje de sentimiento y palabras clave asociadas. Esta tabla permite a los usuarios ordenar y filtrar los datos según diversos criterios.

Además, se incluye una sección lateral de filtros que permite a los usuarios seleccionar criterios como fecha, categoría o sentimiento, optimizando así la experiencia de análisis y personalizando la visualización de datos según sus necesidades.

8.2 Análisis Predictivo

Se ofrece a los usuarios la oportunidad de ingresar un artículo con el propósito de predecir su sentimiento y realizar una posible reclasificación de categoría. El formulario cuenta con los siguientes campos:

- Título: (Campo de texto para que el usuario ingrese el título del artículo)
- Contenido: (Caja de texto donde el usuario debe ingresar el contenido del artículo)
- Botón: Predecir Sentimiento (para iniciar la predicción del sentimiento y la posible reclasificación de la categoría)
- Botón: Cancelar (para cerrar el formulario o limpiar los campos)

Clasificador de Sentimientos de Artículos

Filtro: Título:
Fecha: Seleccionar

Predicción Sentimientos: Contenido: Contenido del artículo

Predecir Sentimiento **Cancel**

Resultado de Clasificación (después de hacer clic en "Predecir Sentimiento"):

- Título del Artículo: (Título ingresado por el usuario)
- Contenido del Artículo: (Contenido ingresado por el usuario)
- Sentimiento Predicho:
 - Sentimiento: (Positivo, Neutro o Negativo)
 - Puntuación del Sentimiento: (valor entre -1 y 1)
- Palabras Clave: (Lista de palabras clave extraídas del contenido)
- Categoría Sugerida: (Categoría en la que el artículo se clasifica según el análisis)

Resultado de Clasificación

Título del Artículo: "El futuro de la tecnología: Innovaciones que cambiarán el mundo"
Contenido del Artículo: "La tecnología está avanzando a pasos agigantados. Las nuevas innovaciones están transformando la manera en que vivimos y trabajamos, ofreciendo soluciones que facilitan nuestra vida diaria."

Predicción

Sentimiento: Positiva
Puntuación del Sentimiento: 0.85 (Escala de -1 a 1)
Sentimiento:
Palabras Claves: Tecnología, innovaciones, transformando, soluciones.

Categoría: Innovación

ANEXO -1. SCREENSHOT. REPOSITORIO GIT DEL PROYECTO FINAL

The screenshot shows a GitHub repository page for 'miad2024_dsa_G16'. The repository is public and has 9 commits. The README file contains the following text:

CATEGORIZACIÓN Y ANÁLISIS DE SENTIMIENTOS DE ARTÍCULOS DE NOTICIAS

Key details from the repository page:

- Code**: The main branch is 'main'.
- Commits**: 9 commits by 'jhermoher'.
- Activity**: No releases or packages published.
- Contributors**: 2 contributors: 'jhermoher' and 'jerssonMH'.

ANEXO -2. Screenshot. REPOSITORIO REMOTO DVC EN AWS-S3 DEL PROYECTO FINAL

The screenshot shows the AWS S3 console interface. The left sidebar is titled "Amazon S3" and includes sections for Buckets, Access Grants, Access Points, Object Lambda Access Points, Multi-Region Access Points, Batch Operations, IAM Access Analyzer for S3, Block Public Access settings for this account, Storage Lens, Dashboards, Storage Lens groups, AWS Organizations settings, Feature spotlight, and a link to the AWS Marketplace for S3.

The main content area displays the contents of the "md5/" folder within the "files/" bucket. The title bar shows the path: Amazon S3 > Buckets > dsa-finalproject-dvc-store > files/ > md5/. The top right corner shows the region as N. Virginia and the user as vclabs/user3581208=Jersson_Hernan_Morales_Hernandez @ 5451-2... .

The "Objects" tab is selected, showing three items:

Name	Type	Last modified	Size	Storage class
29/	Folder	-	-	-
76/	Folder	-	-	-
bc/	Folder	-	-	-

Below the table, there are buttons for Copy S3 URI, Copy URL, Download, Open, Delete, Actions, Create folder, and Upload. A search bar at the top says "Find objects by prefix".

At the bottom of the page, there are links for CloudShell, Feedback, © 2024, Amazon Web Services, Inc. or its affiliates., Privacy, Terms, and Cookie preferences.

ANEXO -3. TEMAS CONSIDERADOS PARA EL PROYECTO.

1. Segmentación de Clientes y Predicción de Abandono

Este proyecto combinaría la segmentación de clientes (aprendizaje no supervisado) con la predicción de abandono (aprendizaje supervisado) para ayudar a las empresas a comprender su base de clientes y predecir qué clientes es probable que abandonen. Utilizaría algoritmos como K-means o clustering jerárquico para la segmentación, considerando factores como la frecuencia de compra, valor monetario y tiempo desde la última compra.

La parte supervisada emplearía Random Forest, Gradient Boosting o Regresión Logística, incorporando características demográficas, historial de compras e interacciones de soporte al cliente. El proyecto utilizaría el conjunto de datos de comercio electrónico de una tienda minorista del Reino Unido, disponible en Kaggle, y se presentaría en un tablero con la distribución de segmentos de clientes y puntuaciones de riesgo de abandono.

2. Categorización de Artículos de Noticias y Análisis de Sentimiento

Este proyecto combinaría el modelado de temas (aprendizaje no supervisado) con el análisis de sentimiento (aprendizaje supervisado) para extraer información de artículos de noticias. El componente no supervisado utilizaría algoritmos como LDA o NMF para el modelado de temas, incluyendo preprocesamiento de texto y creación de matrices documento-término (TF-IDF).

Para el análisis de sentimiento, se implementaría VADER o BERT ajustado, clasificando los artículos como positivos, negativos o neutrales. El proyecto utilizaría un conjunto de datos de categorías de noticias de Kaggle que contiene titulares y descripciones, presentando un tablero con distribución de temas a lo largo del tiempo y tendencias de sentimiento por categoría.

3. Sistema de Recomendación de Películas

Este proyecto crearía un sistema personalizado de recomendación de películas utilizando filtrado colaborativo (aprendizaje no supervisado) y predicción de calificaciones (aprendizaje supervisado). El componente no supervisado emplearía factorización matricial mediante Descomposición en Valores Singulares para encontrar usuarios y películas similares.

La parte supervisada utilizaría Random Forest Regression o Redes Neuronales, considerando características como edad del usuario, género de la película y año de lanzamiento. El proyecto utilizaría el conjunto de datos MovieLens, que contiene calificaciones de usuarios y metadatos de películas, presentando un tablero con recomendaciones personalizadas y predicciones de calificación para pares usuario-película.

ANEXO -4. PRIMER MOCKUP DEL TABLERO

