SAMSUNG

HATE: Discurso de odio

MOD

Big Data Course

Together for Tomorrow!
Enabling People
Education for Future Generations

MOD

MOD

UNIT 4. Conclusiones

4.1. Conclusiones

Hate en redes sociales

- Violencia hacia minorías
- Intolerancia y discriminación
- Problemas de salud mental

Prevención

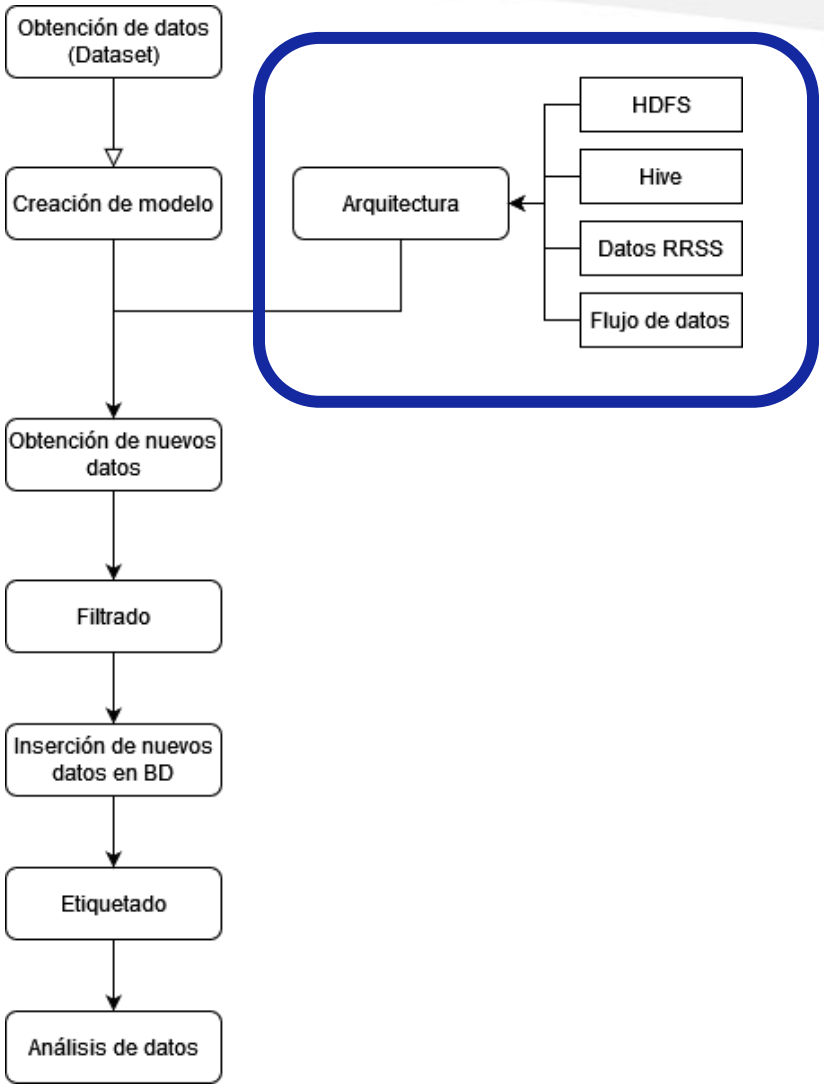
- Prevención antes de la reacción
- Identificación automática de hate
- Analítica de días según el hate
- Visualización en tiempo real de la cantidad de hate

Herramientas

- Sistema de ficheros distribuido (HDFS)
- Sistema de consultas sobre HDFS (Hive)
- Base de datos (MariaDB)
- Filtrado de datos (Apache Spark)
- Entrenamiento de Modelo de predicción de Hate (Pyspark y Python)
- Obtención de datos de Redes Sociales (Python)
- Flujo de datos (NiFi)
- Visualización de datos (PowerBI y jupyter)
- Lenguajes de programación (shell)

Herramientas

- Sistema de ficheros distribuido (HDFS)
- Sistema de consultas sobre HDFS (Hive)
- Base de datos (MariaDB)
- Filtrado de datos (Apache Spark)
- Entrenamiento de Modelo de predicción de Hate (Pyspark y Python)
- Obtención de datos de Redes Sociales (Python)
- Flujo de datos (NiFi)
- Visualización de datos (PowerBI y jupyter)
- Lenguajes de programación (shell)



Zerotier

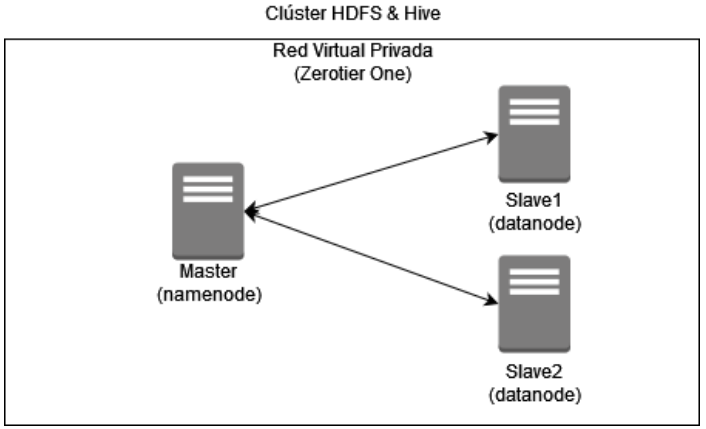
- Red virtual
- Acceso a los equipos del clúster HDFS.
- Asignación y configuración de IP´s

HDFS

Configuración
Namenode y Datanodes

Elección directorio
home HDFS

Comprobación
acceso correcto



HDFS

```
 GNU nano 2.3.1

Este fichero es una prueba del clúster HDFS
```

```
[hadoop@Slave2 ~]$ hdfs dfs -ls /user/hadoop/tmp
[hadoop@Slave2 ~]$ hdfs dfs -put cursoSamsung.txt /user/hadoop/tmp
[hadoop@Slave2 ~]$ nano cursoSamsung.txt
[hadoop@Slave2 ~]$
```

```
[hadoop@Master ~]$ hdfs dfs -ls /user/hadoop/tmp
[hadoop@Master ~]$ hdfs dfs -ls /user/hadoop/tmp
Found 1 items
-rw-r--r--   3 hadoop hadoop         45 2023-07-28 02:09 /user/hadoop/tmp/cursoSamsung.txt
[hadoop@Master ~]$
```

```
[hadoop@Slave1 ~]$ hdfs dfs -ls /user/hadoop/tmp
[hadoop@Slave1 ~]$ hdfs dfs -ls /user/hadoop/tmp
Found 1 items
-rw-r--r--   3 hadoop hadoop         45 2023-07-28 02:09 /user/hadoop/tmp/cursoSamsung.txt
[hadoop@Slave1 ~]$ hdfs dfs -get /user/hadoop/tmp/cursoSamsung.txt ~
[hadoop@Slave1 ~]$ ls
authors.avsc  codegen_authors.java  Desktop    Downloads  hadoopdata     hive-site.xml      hola1            Music     proyecto  Templates  Videos
authors.java  cursoSamsung.txt      Documents  hadoop     hadoop.tar.gz  hive-site.xml.org  init_iptables.sh  Pictures  Public    tmp
[hadoop@Slave1 ~]$ cat cursoSamsung.txt
Este fichero es una prueba del clúster HDFS
[hadoop@Slave1 ~]$
```

Hive

Configuración servidor hive en Master

Configuración hive en Slaves

Comprobación acceso correcto

Hive

```
hive> SELECT COUNT(comment_id) from prediccion.comments
    > WHERE source_id = 3
    > AND community_id = 'roastme';

17709
Time taken: 20.536 seconds, Fetched: 1 row(s)
```

```
hive> SELECT comment from prediccion.comments
    > WHERE source_id = 1
    > AND community_id = 'samsung'
    > LIMIT 5;


I LOVE YOU Samsung
I will make money,, and make Samsung my brand product tech ðŸ™ðŸ™
Samsung ðŸ‘ðŸ»ðŸ‘ðŸ»ðŸ‘ðŸ»ðŸ‘ðŸ»
Interpol is corrupted
now i know why samsung is awesome
```

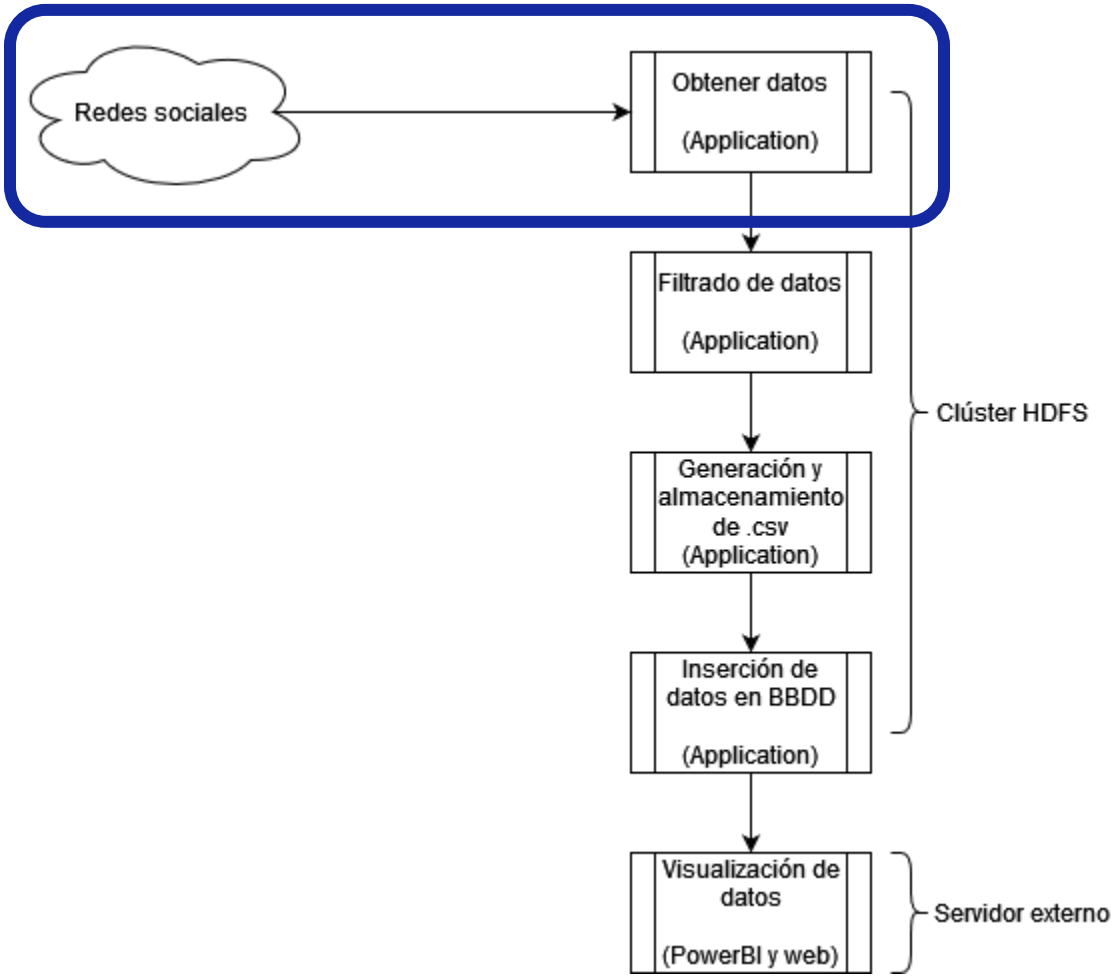
MariaDB

```
MariaDB [(none)]> SELECT COUNT(comment_id) from prediccion.comments
    -> WHERE source_id = 3
    -> AND community_id = 'roastme';
+-------------------+
| COUNT(comment_id) |
+-------------------+
|             17709 |
+-------------------+
1 row in set (0.07 sec)
```

```
MariaDB [(none)]> SELECT comment from prediccion.comments
    -> WHERE source_id = 1
    -> AND community_id = 'samsung'
    ->  LIMIT 5;
+-------------------------------------------------------------------------------+
| comment                                                                       |
+-------------------------------------------------------------------------------+
| I LOVE YOU Samsung                                                            |
| I will make money,, and make Samsung my brand product tech ðŸ™ðŸ™           |
| Samsung ðŸ‘ðŸ»ðŸ‘ðŸ»ðŸ‘ðŸ»ðŸ‘ðŸ»                                                |
| Interpol is corrupted                                                         |
| now i know why samsung is awesome                                             |
+-------------------------------------------------------------------------------+
```

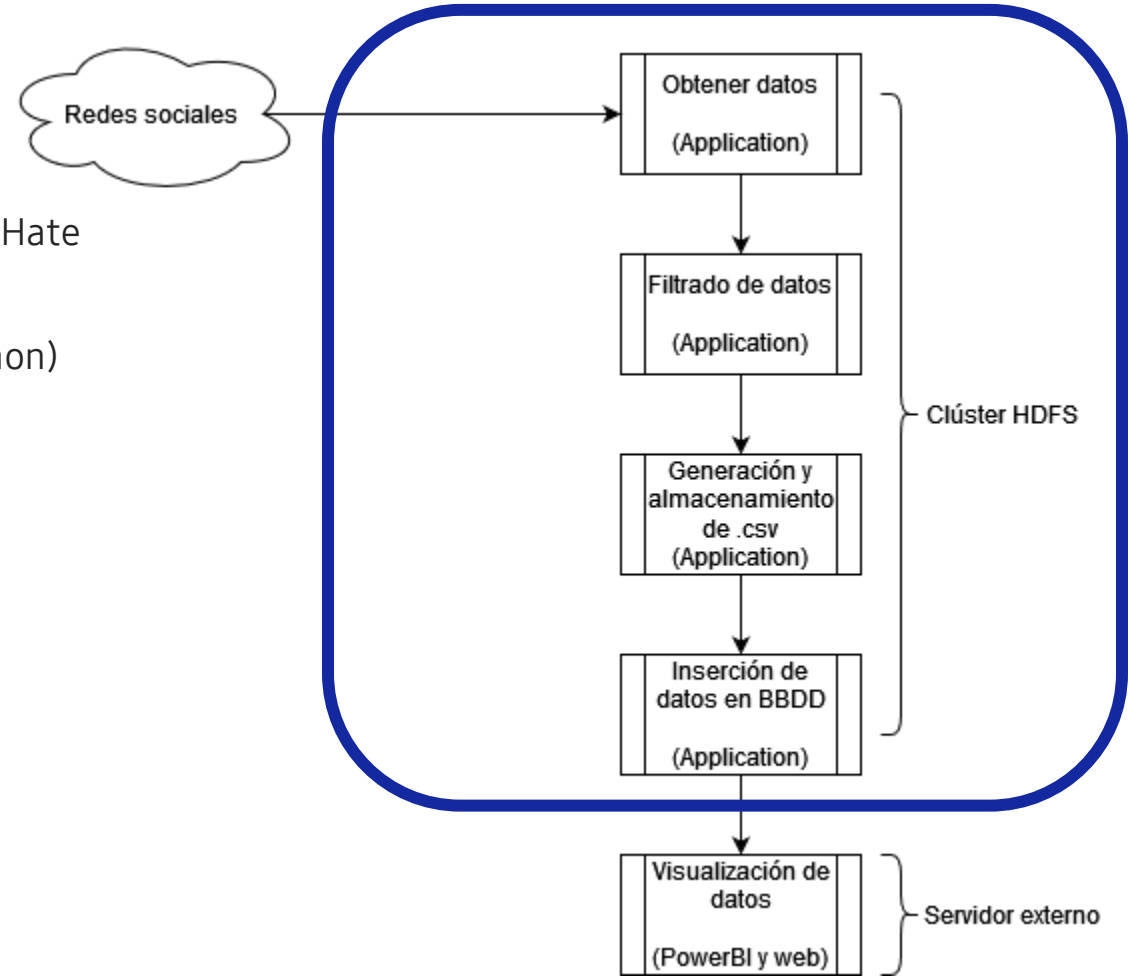
Diagrama

- Sistema de ficheros distribuido (HDFS)
- Sistema de consultas sobre HDFS (Hive)
- Base de datos (MariaDB)



Diagrama

- Filtrado de datos (Apache Spark)
- Entrenamiento de Modelo de predicción de Hate (Pyspark y Python)
- Obtención de datos de Redes Sociales (Python)
- Flujo de datos (NiFi)
- Visualización de datos (PowerBI y jupyter)
- Lenguajes de programación (shell)



Formato de los datos

comentario	hate	fecha	Información adicional
Texto del comentario	0	26-07-2023	id, usuario, fuente ...

Datos

Estructurados

No Estructurados

En Streaming

Datos

Estructurados

kaggle

Más de 160 mil datos etiquetados

Entrenamiento del modelo


```
from pyspark.ml.classification import GBTClassifier

gbt = GBTClassifier(maxIter=10)
gbtModel = gbt.fit(train)
gbtPreds = gbtModel.transform(validate)
gbtPreds.show(5)
```

```
Accuracy of GBT is = 0.759571
```

Datos

Estructurados

```
_________________________________________________________________
 Layer (type)                 Output Shape              Param #
=================================================================
 embedding (Embedding)        (None, 300, 100)          5000000

 spatial_dropout1d (Spatial   (None, 300, 100)          0
 Dropout1D)

 lstm (LSTM)                  (None, 100)               80400

 dense (Dense)                (None, 1)                 101

=================================================================
```

Entrenamiento del modelo


```
Epoch 10/10
95/95 [==============================] - 161s 2s/step - loss: 0.3907 - accuracy: 0.8255
```

Datos

Estructurados

No Estructurados

En Streaming

kaggle

Datos

No Estructurados

Más de 125 mil datos

Datos

No Estructurados

Comentario
A beautiful world
Hello
U dumb ass

Modelo de IA

Hate
0
0
1

Datos

Estructurados

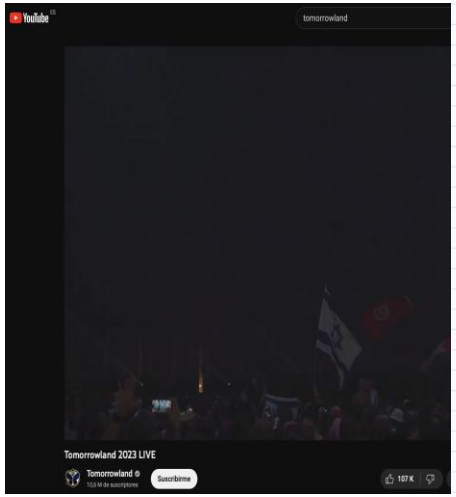
kaggle

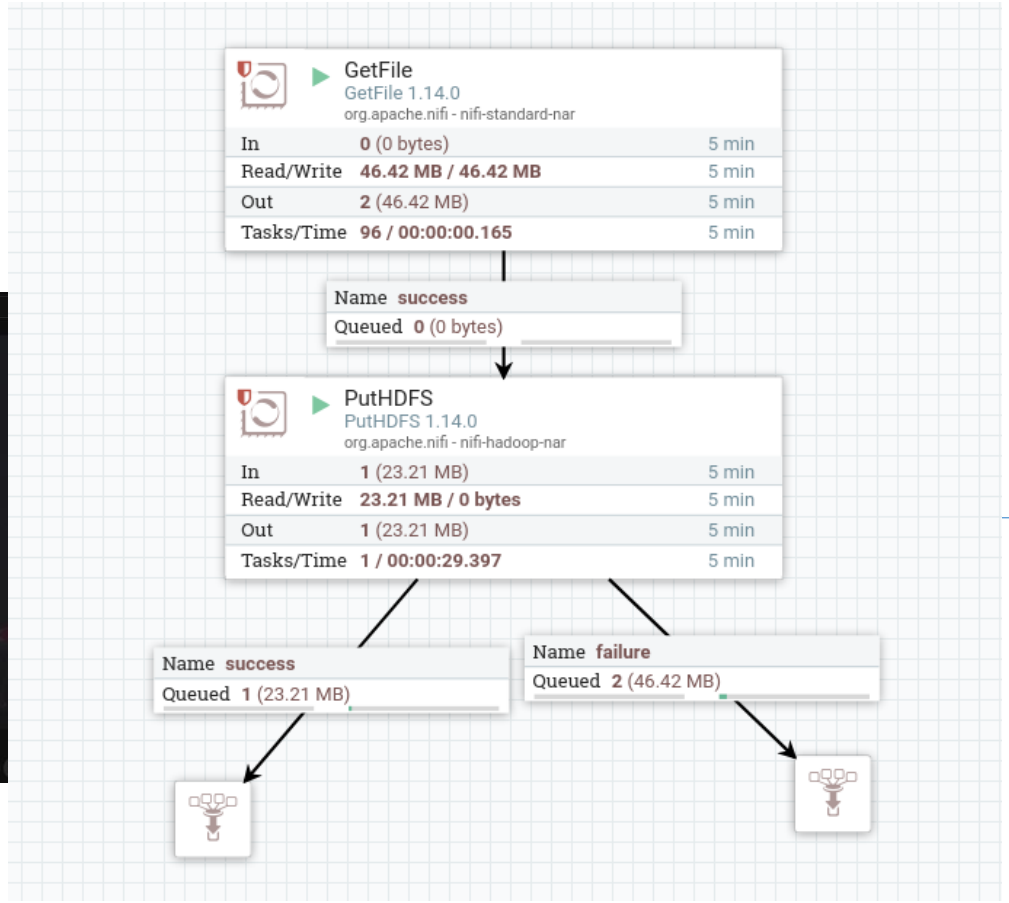
No Estructurados

En Streaming

Datos

YouTube LIVE





En Streaming

hadoop HDFS

MariaDB

Datos

Estructurados

kaggle

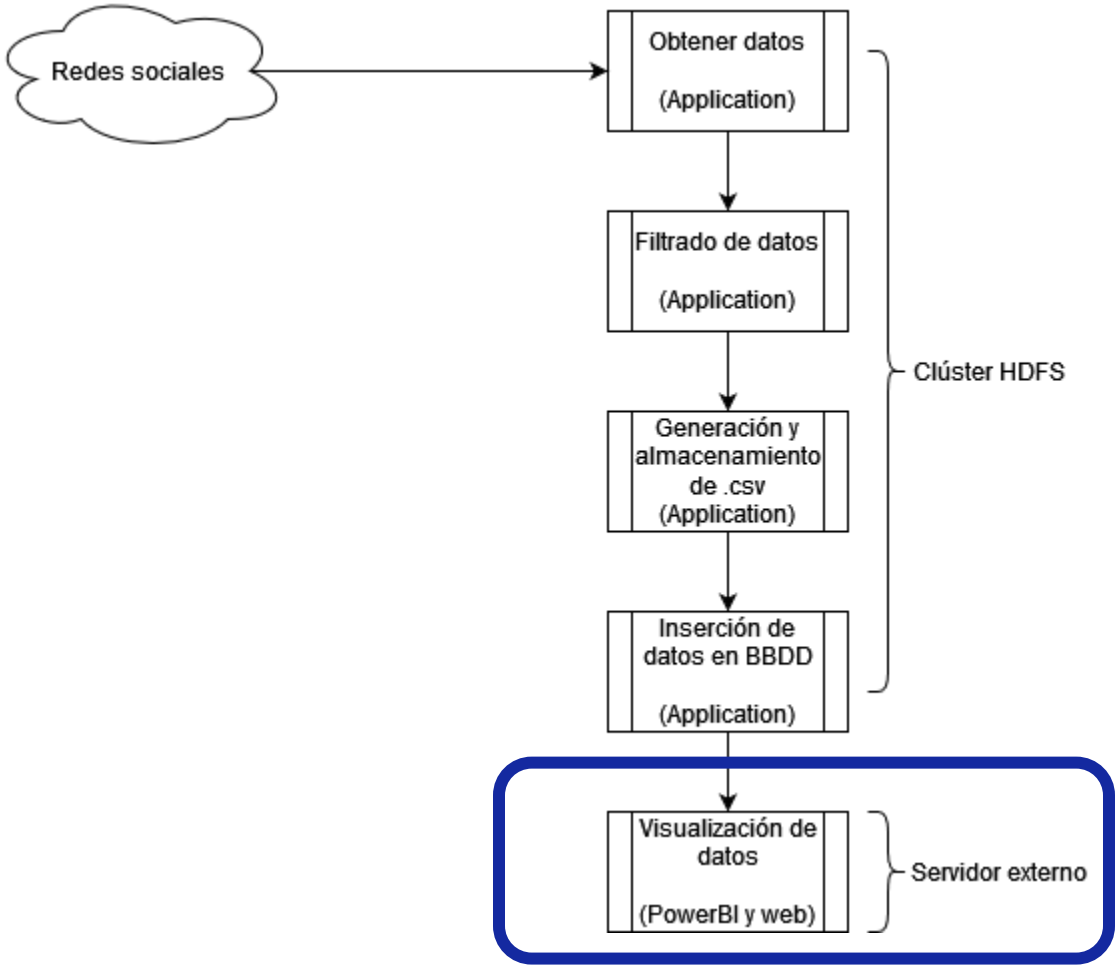
No Estructurados

En Streaming

YouTube LIVE

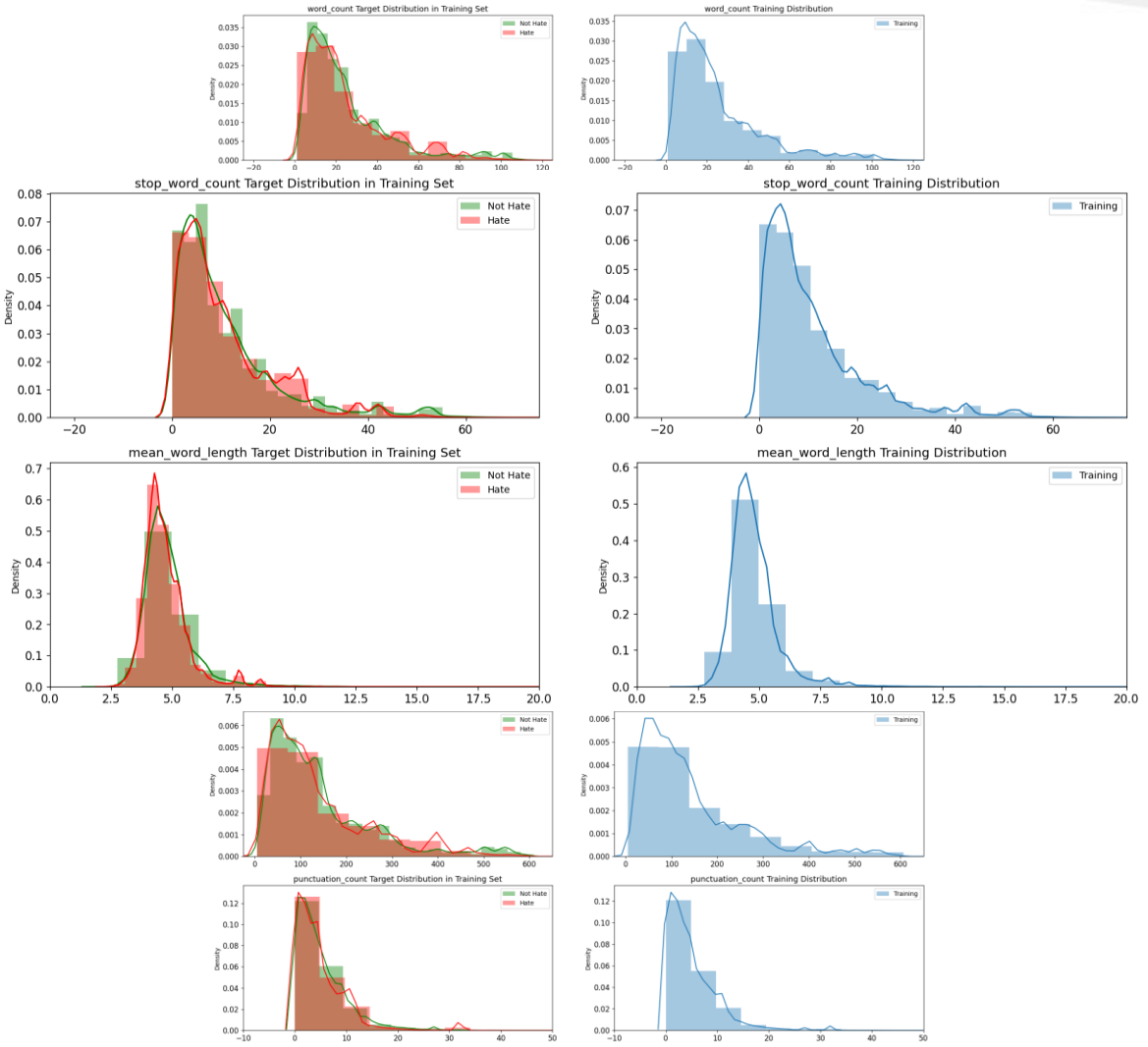
Diagrama

- Visualización de datos (PowerBI y jupyter)
- Lenguajes de programación (shell)



Analisis

- Datos Estructurados
- Previo al Entrenamiento del Modelo

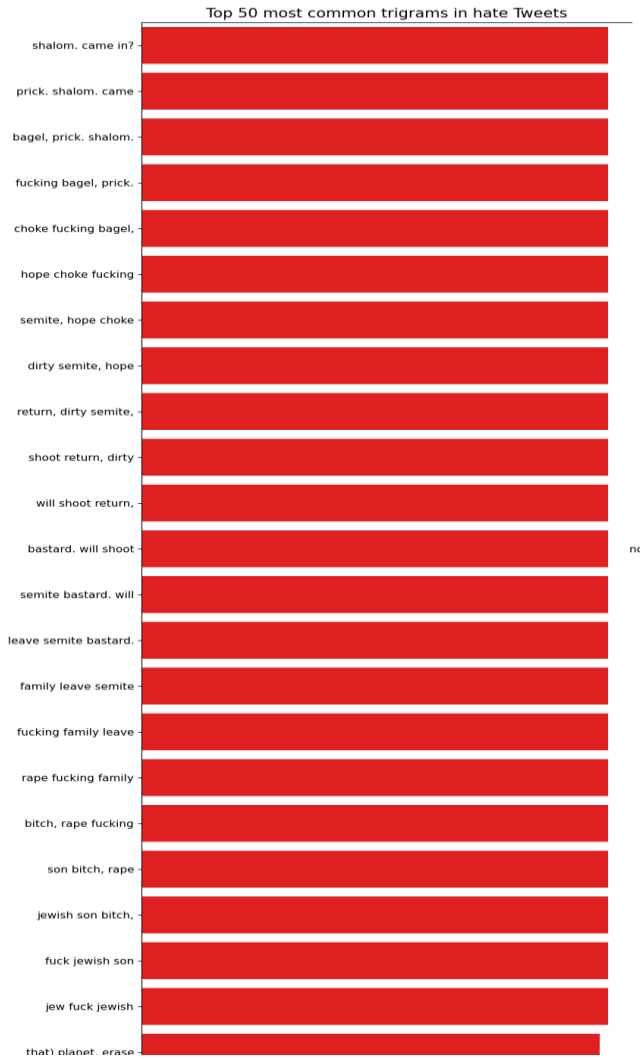


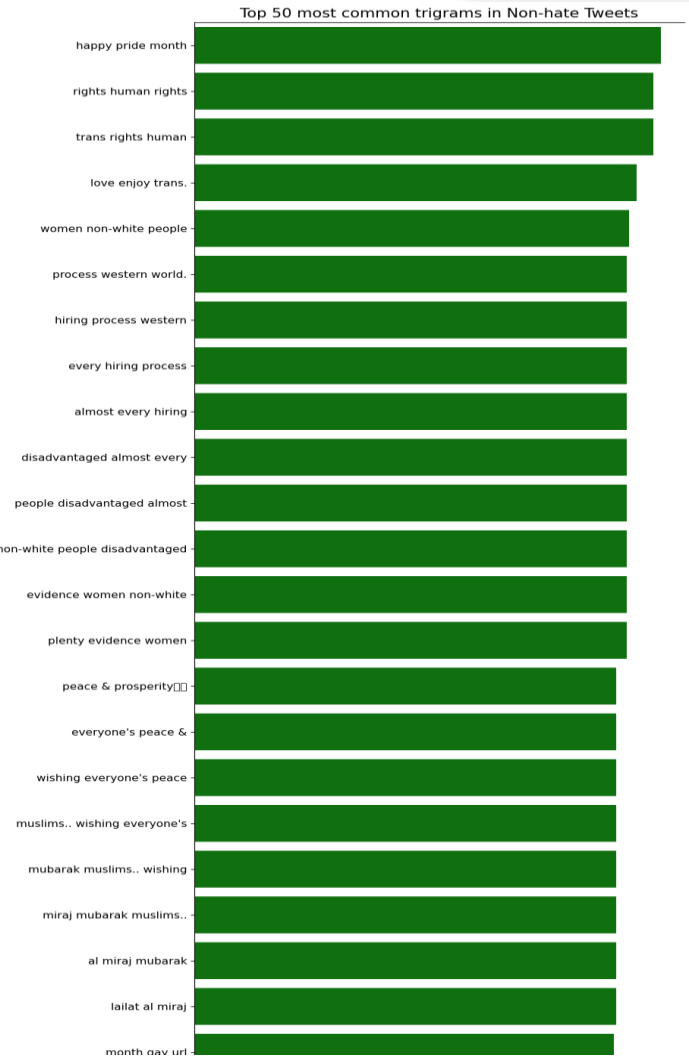
jupyter

Analisis

- Datos Estructurados
- Previo al Entrenamiento del Modelo
- N-Gramas

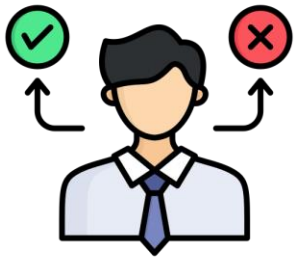
jupyter

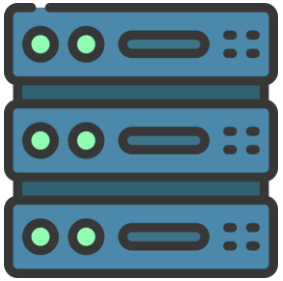




PowerBI

- Software de Visualización
- Conexión a Servidor
- Ayuda a la toma de decisiones





PowerBI

Visualización de analíticas

Conclusiones

- Diseño y Desarrollo de un prototipo de sistema de predicción de hate
- Prevención de hate
- Visualización de hate en tiempo real
- Análisis de hate en redes sociales
- Expectativas de una publicación
- Importancia de la salud mental
- Necesidad de colaboración con profesionales

SAMSUNG

Together for Tomorrow!

Enabling People

Education for Future Generations