جامعة عبد المالك السعدي
Université Abdelmalek Essaâdi

كلية العلوم و التقنيات بطنجة
Faculté des Sciences et Techniques de Tanger

IASD
INTELLIGENCE ARTIFICIELLE
ET SCIENCES DE DONNEES

Report

Real-Time Twitter Sentiment Analysis

Supervised by : Pr. Yasyn EL YUSUFI

Realized by

El ghazi Loubna

Zaoui Hanane

2023/2024

1-Project Overview

This project focuses on the development of a sentiment analysis application utilizing big data technologies. Our primary goal was to build a scalable and efficient system capable of processing and analyzing large volumes of Twitter data in real-time. The core components of our project include data ingestion, storage, processing, and front-end visualization.

2-Technologies Used

Apache Kafka is an open-source distributed streaming platform designed for handling real-time data feeds with high fault tolerance and high throughput. It is commonly used for data collection, storage, and analysis, as well as for building real-time data pipelines and streaming applications.

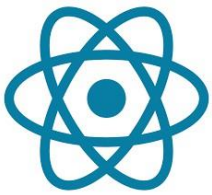
Apache Hadoop is an open-source framework that allows for the distributed storage and processing of large data sets across clusters of computers using simple programming models. It is designed to scale up from single servers to thousands of machines, each offering local computation and storage.

Apache Spark is an open-source, distributed computing system designed for big data processing, with built-in modules for streaming, SQL, machine learning, and graph processing. Spark provides an interface for programming entire clusters with implicit data parallelism and fault tolerance.

FastAPI is a modern, fast (high-performance), web framework for building APIs with Python 3.6+ based on standard Pythontype hints. It is designed to be easy to use and deploy, providing automatic interactive API documentation.

MongoDB is a NoSQL database that stores data in flexible, JSON-like documents. This allows for schema flexibility, which can be particularly useful when dealing with unstructured data like social media posts.

React is a JavaScript library for building user interfaces, particularly single-page applications where data can change dynamically.



Docker simplifies the development, deployment, and management of applications by encapsulating them in portable containers.

3-Architechture and Implementation

-Data Ingestion with Apache Kafka

Apache Kafka was used to ingest Twitter data streams. Kafka was configured to collect streaming data from Twitter and transfer it to Hadoop for storage and processing.

-Data Storage and Processing with Apache Hadoop

The collected Twitter data was stored in Hadoop's HDFS (Hadoop Distributed File System). We utilized Apache Spark with **PySpark** for processing the data. The data from the twitter_training.csv file was preprocessed using:

Tokenization: Splitting the text into individual words.
Stop Words Removal: Filtering out common stop words to reduce noise in the data.
N-Grams: Generating bigrams to capture context between adjacent words.
Feature Generation: Using CountVectorizer to convert text into numerical features suitable for machine learning models.

-Machine Learning Model Training

Supervised learning models were trained on the preprocessed data:

LOGISTIC REGRESSION
RANDOM FOREST
NAÏVE BAYES

The models were evaluated, and the best-performing model(In our case **LOGISTIC REGRESSION MODEL**) was saved for future predictions.

-Real-Time Sentiment Prediction

The saved model was used to predict the sentiment (Negative, Positive, Neutral, Irrelevant) of incoming tweets in real-time. The results were validated against the **twitter_validation.csv dataset.**

-Database Integration with MongoDB

Predicted sentiment data was stored in **MongoDB** to facilitate efficient retrieval and presentation.

-Web Application Development with FastAPI and React

The backend was developed using FastAPI to create a robust RESTful API for data processing and retrieval. The frontend was built using React with DaisyUI to provide an interactive user interface.

-Project Deployment on GitHub

The entire project, including the codebase and documentation, was uploaded to GitHub for version control and collaboration.

4- Data Description

Training Data: twitter_training.csv

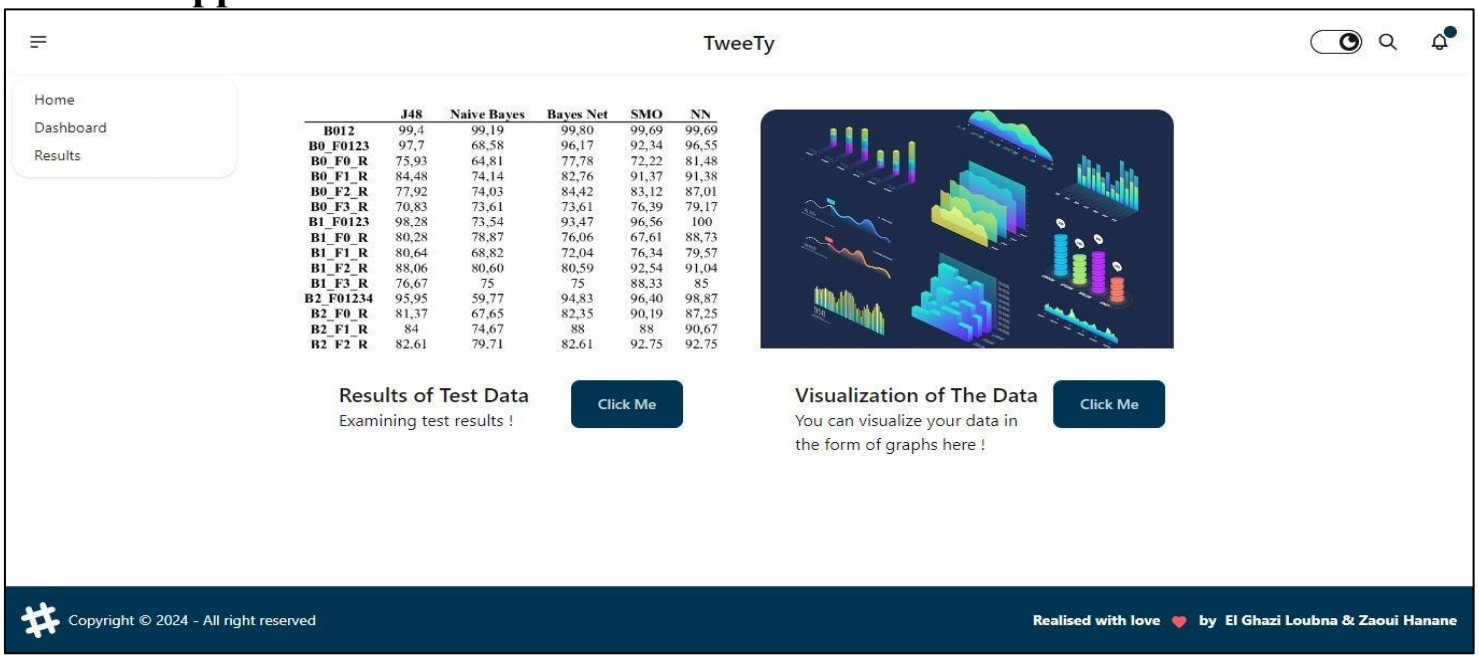
- **Tweet ID**: int
- **Entity**: string
- **Sentiment**: string (Label)
- **Tweet content**: string

Validation Data: twitter_validation.csv

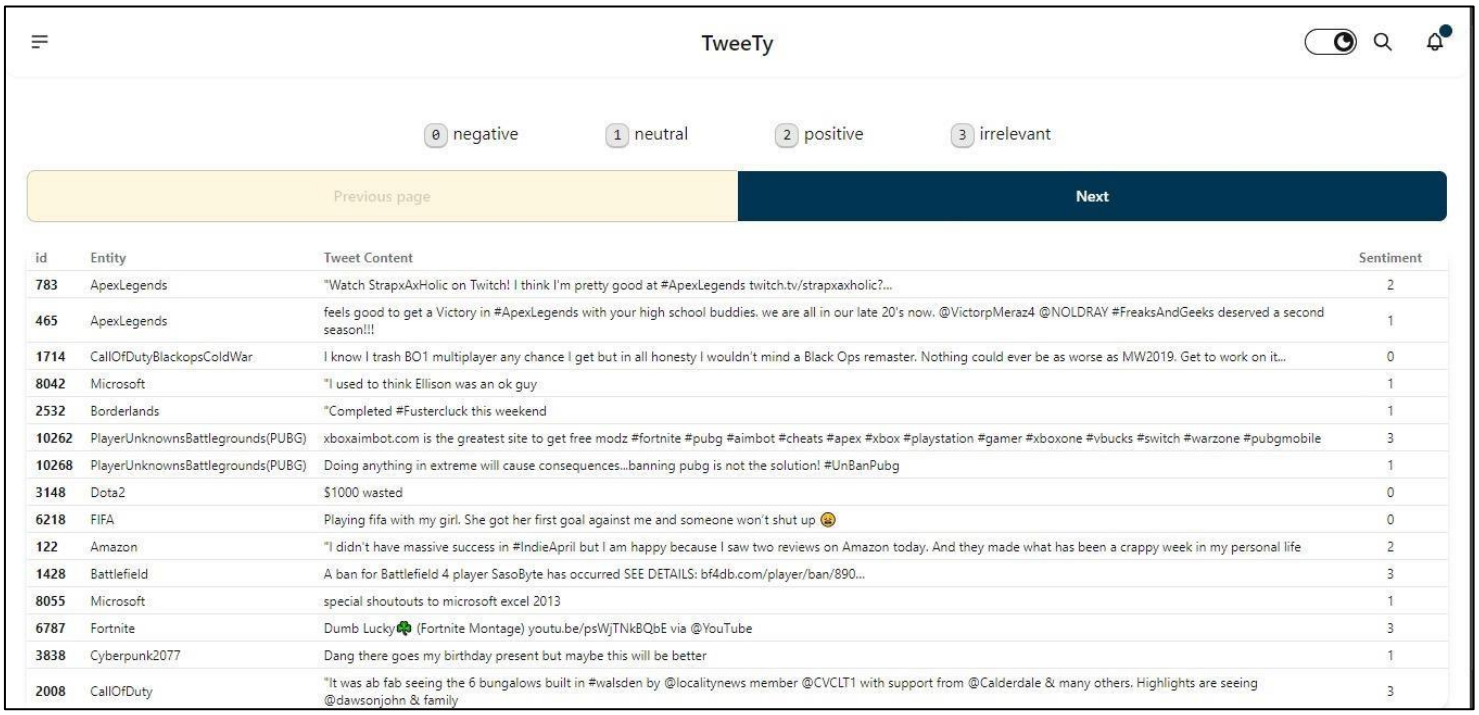
- Contains 998 tweets with the same structure as the training data.
- The "Sentiment column is the target for prediction.

Source of Data : https://www.kaggle.com/datasets/jp797498e/twitter-entity-sentiment-analysis

5-The application Web



	J48	Naive Bayes	Bayes Net	SMO	NN
B012	99,4	99,19	99,80	99,69	99,69
B0_F0123	97,7	68,58	96,17	92,34	96,55
B0_F0_R	75,93	64,81	77,78	72,22	81,48
B0_F1_R	84,48	74,14	82,76	91,37	91,38
B0_F2_R	77,92	74,03	84,42	83,12	87,01
B0_F3_R	70,83	73,61	73,61	76,39	79,17
B1_F0123	98,28	73,54	93,47	96,56	100
B1_F0_R	80,28	78,87	76,06	67,61	88,73
B1_F1_R	80,64	68,82	72,04	76,34	79,57
B1_F2_R	88,06	80,60	80,59	92,54	91,04
B1_F3_R	76,67	75	75	88,33	85
B2_F01234	95,95	59,77	94,83	96,40	98,87
B2_F0_R	81,37	67,65	82,35	90,19	87,25
B2_F1_R	84	74,67	88	88	90,67
B2_F2_R	82.61	79.71	82.61	92.75	92.75



id	Entity	Tweet Content	Sentiment
783	ApexLegends	"Watch StrapxAxHolic on Twitch! I think I'm pretty good at #ApexLegends twitch.tv/strapxaxholic?...	2
465	ApexLegends	feels good to get a Victory in #ApexLegends with your high school buddies. we are all in our late 20's now. @VictorpMeraz4 @NOLDRAY #FreaksAndGeeks deserved a second season!!!	1
1714	CallOfDutyBlackopsColdWar	I know I trash BO1 multiplayer any chance I get but in all honesty I wouldn't mind a Black Ops remaster. Nothing could ever be as worse as MW2019. Get to work on it...	0
8042	Microsoft	"I used to think Ellison was an ok guy	1
2532	Borderlands	"Completed #Fustercluck this weekend	1
10262	PlayerUnknownsBattlegrounds(PUBG)	xboxaimbot.com is the greatest site to get free modz #fortnite #pubg #aimbot #cheats #apex #xbox #playstation #gamer #xboxone #vbucks #switch #warzone #pubgmobile	3
10268	PlayerUnknownsBattlegrounds(PUBG)	Doing anything in extreme will cause consequences...banning pubg is not the solution! #UnBanPubg	1
3148	Dota2	$1000 wasted	0
6218	FIFA	Playing fifa with my girl. She got her first goal against me and someone won't shut up 😩	0
122	Amazon	"I didn't have massive success in #IndieApril but I am happy because I saw two reviews on Amazon today. And they made what has been a crappy week in my personal life	2
1428	Battlefield	A ban for Battlefield 4 player SasoByte has occurred SEE DETAILS: bf4db.com/player/ban/890...	3
8055	Microsoft	special shoutouts to microsoft excel 2013	1
6787	Fortnite	Dumb Lucky🍀 (Fortnite Montage) youtu.be/psWjTNkBQbE via @YouTube	3
3838	Cyberpunk2077	Dang there goes my birthday present but maybe this will be better	1
2008	CallOfDuty	"It was ab fab seeing the 6 bungalows built in #walsden by @localitynews member @CVCLT1 with support from @Calderdale & many others. Highlights are seeing @dawsonjohn & family	3

6- Conclusion

This project demonstrates the integration of various technologies to build a real-time sentiment analysis application. Using Apache Kafka for data ingestion, Apache Hadoop for storage and processing, and machine learning models for sentiment prediction, we have developed a robust and efficient system. The web interface built with React and DaisyUI provides an intuitive platform for users to interact with the application.