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 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

-Stream Processing: Uses Apache Kafka to handle real-time data streams from Twitter.

-Data Preprocessing : Employs Pyspark for necessary data preprocessing:

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.

-Dockerized Setup

We used Docker and Docker Compose to streamline the setup and deployment of our application, ensuring that it runs consistently across different environments.

Producer: The producer component is responsible for ingesting data (tweets) into the Kafka topic. It reads the Twitter data from a source (e.g., `twitter_training.csv`) and publishes each tweet as a message to a Kafka topic.

Consumer: The consumer component subscribes to the Kafka topic to consume the messages (tweets) in real-time. It processes the incoming tweets, applies preprocessing steps using Pyspark, and then uses the trained machine learning model to predict the sentiment of each tweet. The results are then logged into MongoDB and served to the frontend via the FastAPI backend.

-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

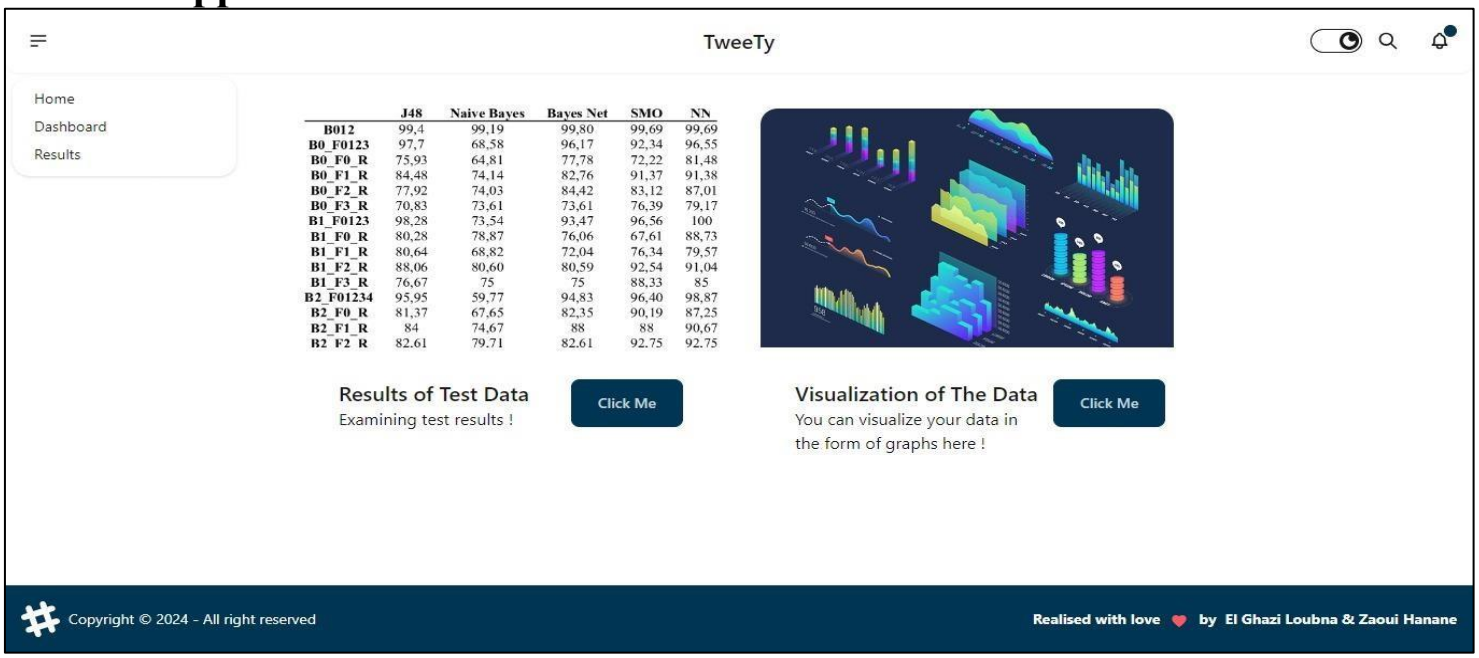
- **Tweet ID**: int
- **Entity**: string
- **Sentiment**: string (Label)
- **Tweet content**: string

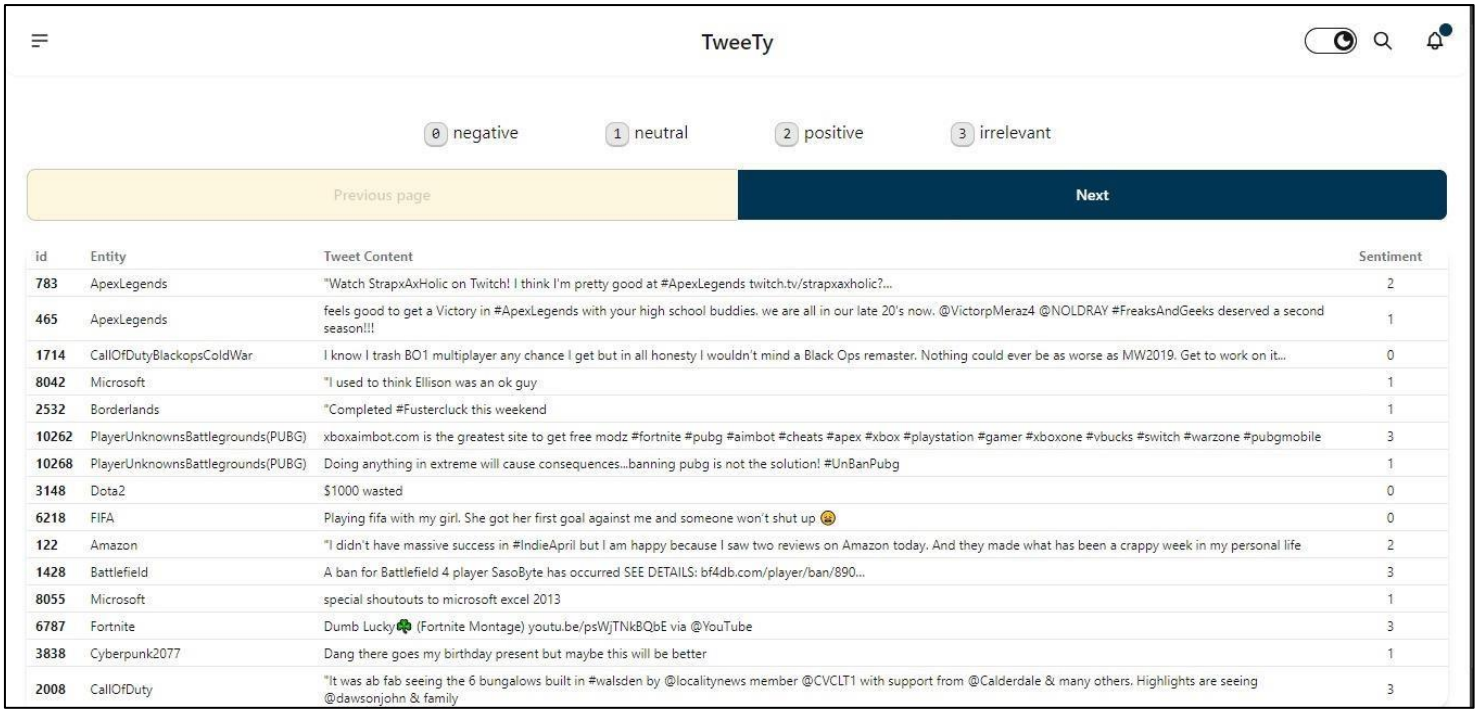
Validation Data: twitter_validation.csv

- 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





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 has been an incredible learning journey, allowing us to apply theoretical knowledge to a practical, real-world problem. The integration of various technologies to achieve real-time sentiment analysis has not only honed our technical skills but also showcased the potential of big data and machine learning in social media analytics.