

James-tech4 / JAMES-DSPT08-PHASE-4-PROJECT

<> Code

Issues

Pull requests

Actions

Projects

Wiki

Security

Insights

Setting

This repository contains files for my Phase 4 Final Project

☆ 0 stars

🍴 0 forks

👁 1 watching

🔗 Branches

📈 Activity

🏷 Tags

🌐 Public repository

🔑

🔗 1 Branch

🏷 0 Tags

🔑

🏷

🔍 Go to file

t


Go to file

+

Add file

Code









...


 James-tech4



Final changes

7ac7682 · 2 minutes ago

🕒

 images	Changes	38 minutes ago
 Notebook.pdf	Updated	2 hours ago
 README.md	Update README.md	35 minutes ago
 index.ipynb	Updated	2 hours ago
 presentation.pdf	Final changes	2 minutes ago
 presentation.pptx	Final changes	8 minutes ago
 twitter_training.csv	Updated	last week
 twitter_validation.csv	Updated	last week

 README

# BUILDING AN NLP MODEL FOR ENTITY-LEVEL SENTIMENT ANALYSIS ON TWITTER



- Student Name: JAMES WACHIRA MUTHEE
- Mode of Study: PART TIME
- Project: End of Phase Project
- Technical Mentor: Daniel Ekale

## Table of Contents

1. Project Overview
2. Problem Statement
3. Business Objectives
4. Target Audience
5. Data
6. Data Understanding
7. Exploratory Data Analysis
8. Data Preprocessing
9. Modelling & Evaluation
10. Validating the Model
11. Findings
12. Conclusion

### 1. Project Overview

In today's digital world, twitter plays a key role in shaping public opinion and brand perception. This project develops an entity-level sentiment analysis model to classify tweets as Positive, Negative, Neutral, or Irrelevant regarding a specific entity (e.g., brands, products, public figures). The goal is to help businesses extract actionable insights, track sentiment trends, and enhance decision-making.

### 2. Problem Statement

Traditional sentiment analysis focuses on entire texts rather than specific entities within them. This project bridges that gap by building a model that identifies sentiment at the entity level, providing more precise insights for businesses.

### 3. Business Objectives

The aim of this project is to build a model that will be able to accurately classify tweets at the entity level into Positive, Negative, Neutral, or Irrelevant sentiments. This will help businesses and organizations to:

- Improve brand and reputation management
- Enhance customer engagement strategies
- Enable data-driven decision-making

### 4. Target Audience

The target audience for this project are:

- Businesses & brands
- Marketing & PR teams
- Customer support teams
- Investors & analysts

### 5. Data

- Dataset sourced from [Kaggle](https://www.kaggle.com/datasets/James-tech4/twitter-sentiment-analysis).

### 6. Data Understanding

Both the training and validation datasets had column names that did not make sense. The column names were renamed as below:

1. ID-A unique identifier for each tweet.
2. Entity-The topic being discussed in the tweet.
3. Sentiment-The emotion assigned to the tweet
4. Tweet - The actual text of the tweet

- The training dataset had 74681 rows and 4 columns while the validation dataset had 999 rows and 4 columns
- The training dataset had 686 missing values in the tweet column while the validation dataset had no missing values
- Both datasets had one column with integer data type and 3 columns with categorical data type

## 7. Exploratory Data Analysis

Some visualizations were done on the training data to get some deeper insights on it's distribution.

## 8. Data Preprocessing

Preprocessing steps applied are:

- Stopword removal
- Lemmatization
- Tokenization
- Label encoding.

## 9. Modeling & Evaluation

Built two models:

- Logistic Regression Model (Baseline)



## Releases

No releases published

[Create a new release](#)

## Packages

No packages published

[Publish your first package](#)

## Languages

- Jupyter Notebook 100.0%