INTERACTIVE WEB APPLICATION FOR TWITTER SENTIMENT ANALYSIS AND HASHTAG TREND DETECTION USING PYSPARK AND FLASK

SUBMITTED TO: ISHAN GUPTA

ADVAIT HARISH (C0936461)
BETSY VARGHESE (C0937312)
ANICA REMIN FERNANDEZ (C0945331)
SAJIN DEV SAHADEVAN (C0933891)
ATHULYA JAYAN (C0936177)

AGENDA

- OBJECTIVE
- INTRODUCTION
- METHODOLOGY
- MACHINE LEARNING APPROACH
- EDA & VISUALIZATIONS
- RESULTS
- CONCLUSION
- FUTURE WORK

PROJECT OBJECTIVE

- GOAL: DEVELOP SCALABLE BIG DATA ANALYTICS + ML PIPELINE FOR TWEET SENTIMENT & CATEGORY CLASSIFICATION
- SCOPE: DATA INGESTION → CLEANING → EDA → FEATURE ENGINEERING → PREDICTIVE MODELING
- TOOLS: APACHE SPARK (PYSPARK), MLLIB, FLASK

INTRODUCTION

- TWITTER AS A LARGE-SCALE SOURCE OF REAL-TIME TEXT DATA
- IMPORTANCE OF BIG DATA PLATFORMS FOR SENTIMENT ANALYSIS
- APACHE SPARK’S ROLE: IN-MEMORY DISTRIBUTED COMPUTING
- PROJECT’S END-TO-END PIPELINE OVERVIEW

BIG DATA ARCHITECTURE OVERVIEW

PIPELINE

DATA SOURCE

↓

INGESTION

↓

PROCESSING

↓

ML

↓

VISUALIZATION

INTEGRATION OF PYSPARK & FLASK

METHODOLOGY OVERVIEW

PHASES:

1. DATA INGESTION
2. CLEANING & PREPROCESSING
3. FEATURE ENGINEERING
4. MODEL TRAINING & EVALUATION

DATA INGESTION & STORAGE

- DATASET: TWITTER.CSV (TWEETS, METADATA, TIMESTAMPS)
- LOADED INTO SPARK DATAFRAME WITH SCHEMA INFERENCE
- STORAGE STRATEGY FOR SCALABILITY

DATA CLEANING & PREPROCESSING

- HANDLING MISSING/NULL CRITICAL VALUES
- REMOVING URLS, MENTIONS, SPECIAL CHARACTERS FROM TWEETS
- TIMESTAMP CONVERSION TO SPARK TIMESTAMP TYPE

FEATURE ENGINEERING

- EXTRACTING WORD COUNTS, SENTIMENT SCORES
- ENCODING CATEGORICAL FIELDS USING STRINGINDEXER
- COMBINING FEATURES INTO A VECTOR WITH VECTORASSEMBLER

DATA SPLITTING

DATA SPLITTING

- TRAINING SET: 70%
- TESTING SET: 30%
- SPARK RANDOMSPLIT USAGE

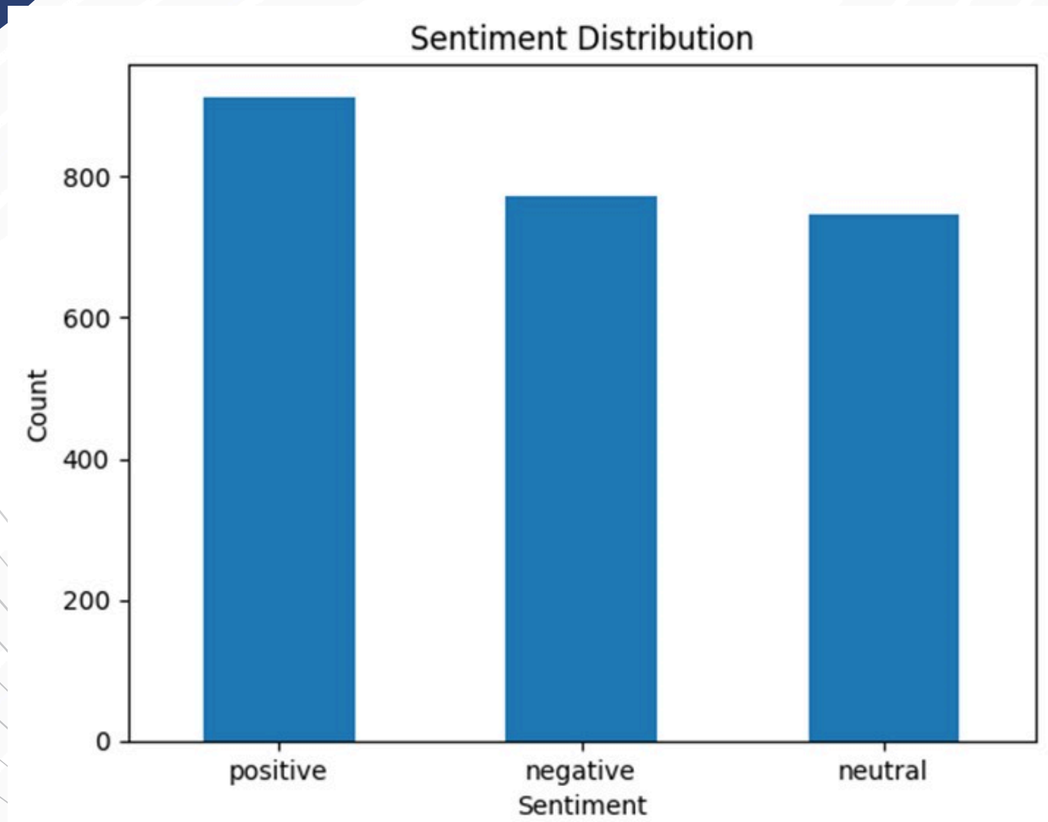
MACHINE LEARNING MODELS

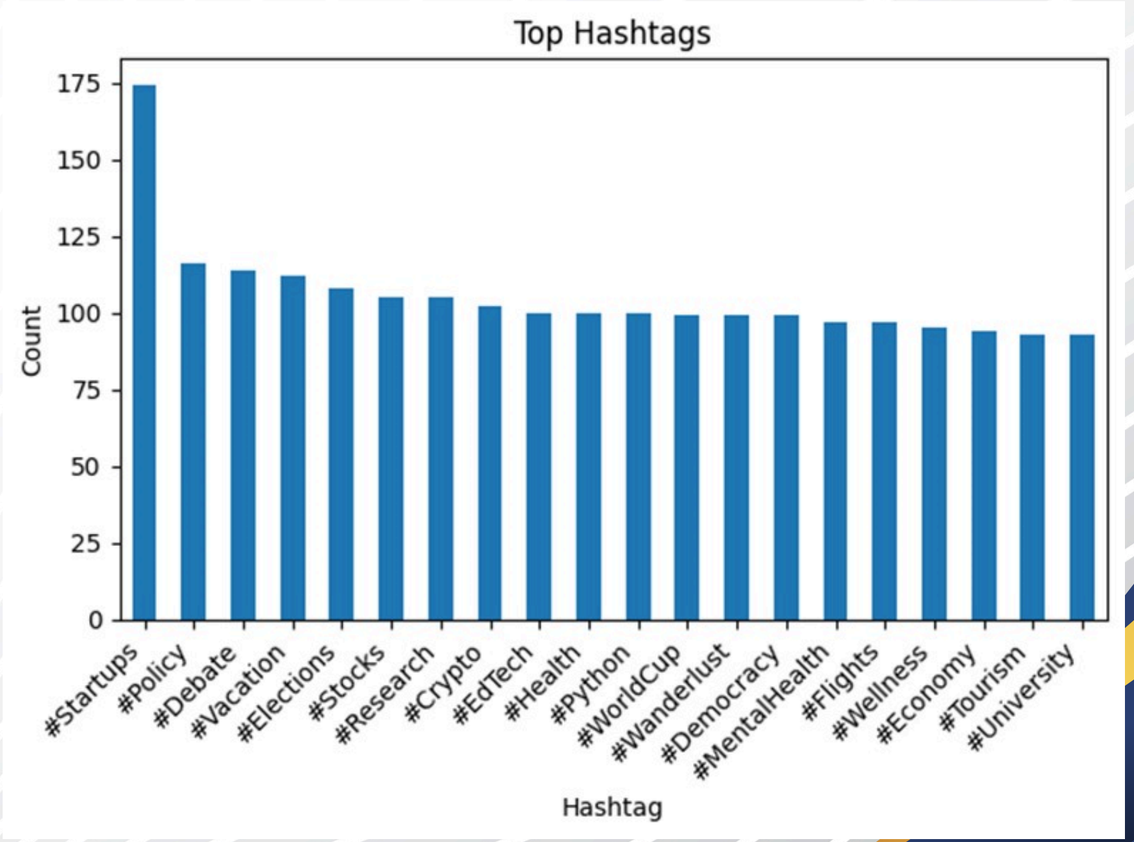
- MODELS USED: LOGISTIC REGRESSION, DECISION TREE CLASSIFIER
- EVALUATION METRICS: ACCURACY, PRECISION, RECALL, F1-SCORE, CONFUSION MATRIX

EXPLORATORY DATA ANALYSIS (EDA)

- COUNT & DISTRIBUTION ANALYSIS OF TWEET METADATA
- WORD FREQUENCY PATTERNS
- SENTIMENT DISTRIBUTION

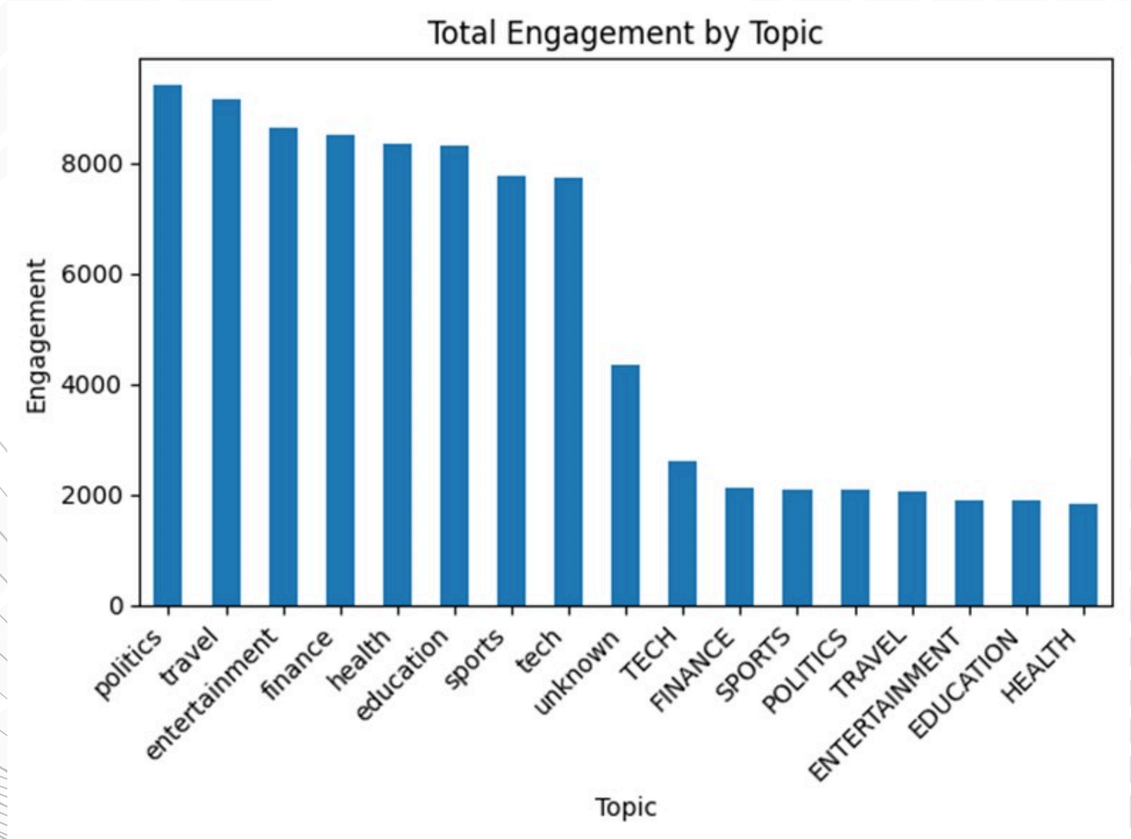
VISUALIZATIONS





SAMPLE CHARTS:

- WORD CLOUD / WORD FREQUENCY BAR CHART
- SENTIMENT DISTRIBUTION PIE CHART
- HASHTAG TRENDS OVER TIME



MODEL PERFORMANCE

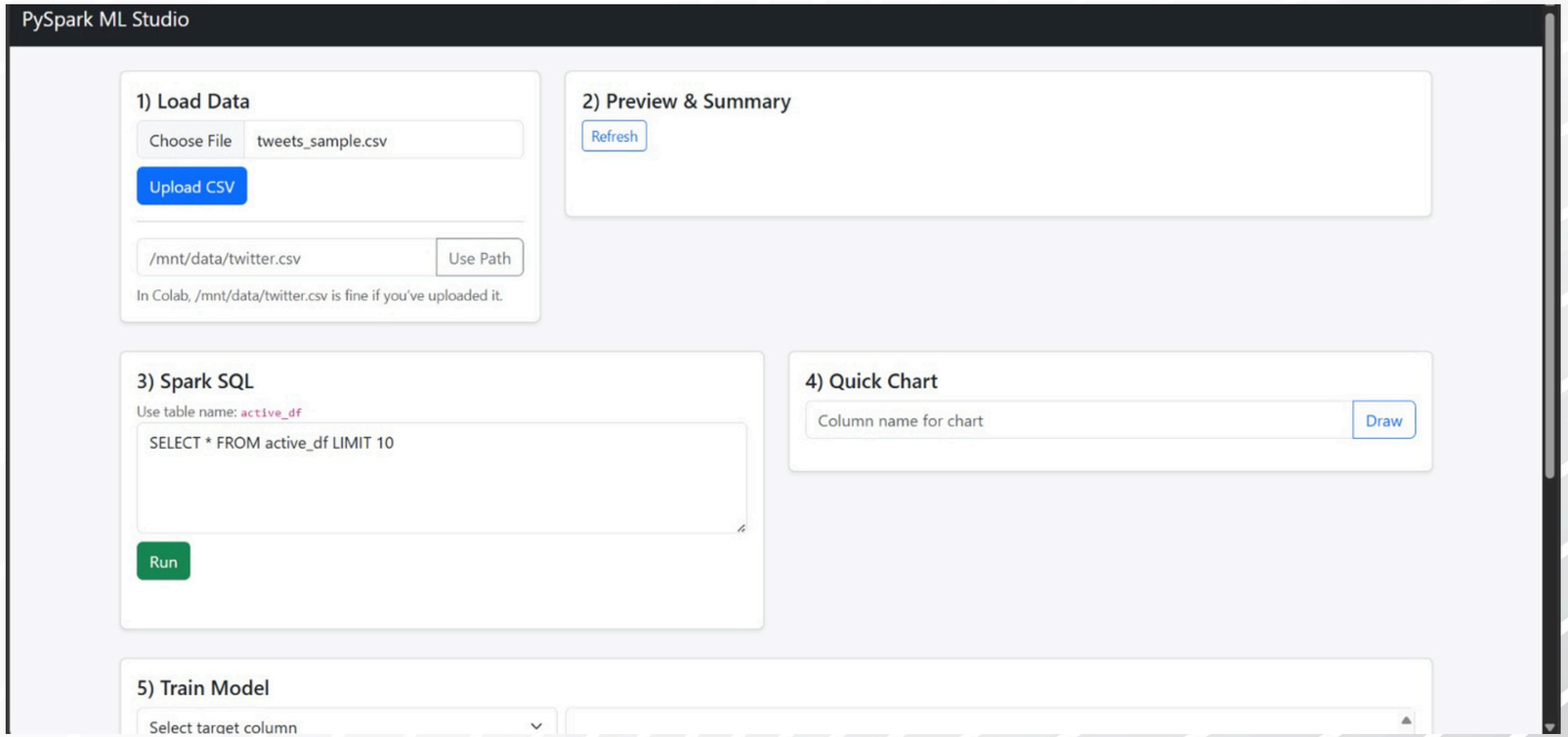
- LOGISTIC REGRESSION RESULTS SUMMARY
- DECISION TREE RESULTS SUMMARY
- COMPARATIVE ACCURACY TABLE

KEY FINDINGS

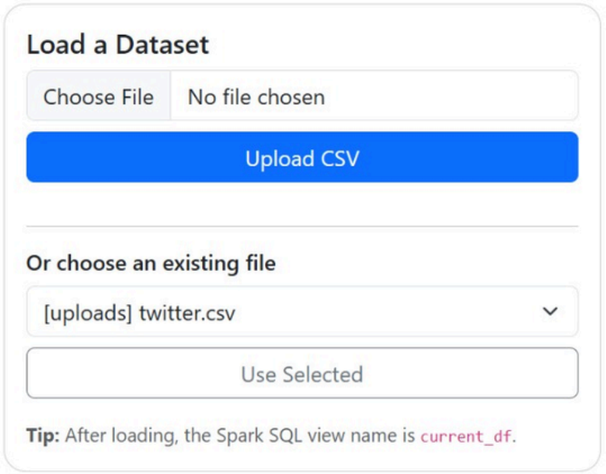
- DECISION TREE OUTPERFORMED LOGISTIC REGRESSION
- FEATURE ENGINEERING BOOSTED PERFORMANCE
- BALANCED PRECISION-RECALL ACROSS CLASSES

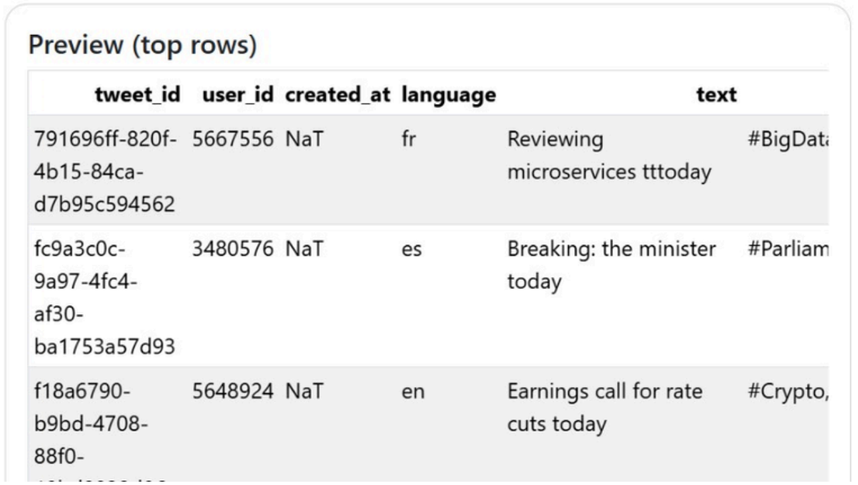
- Flask-based interactive dashboard
- Input tweet → Output sentiment prediction
- Visualization of hashtag trends

APPLICATION INTERFACE



MODEL OUTPUT





Query Builder

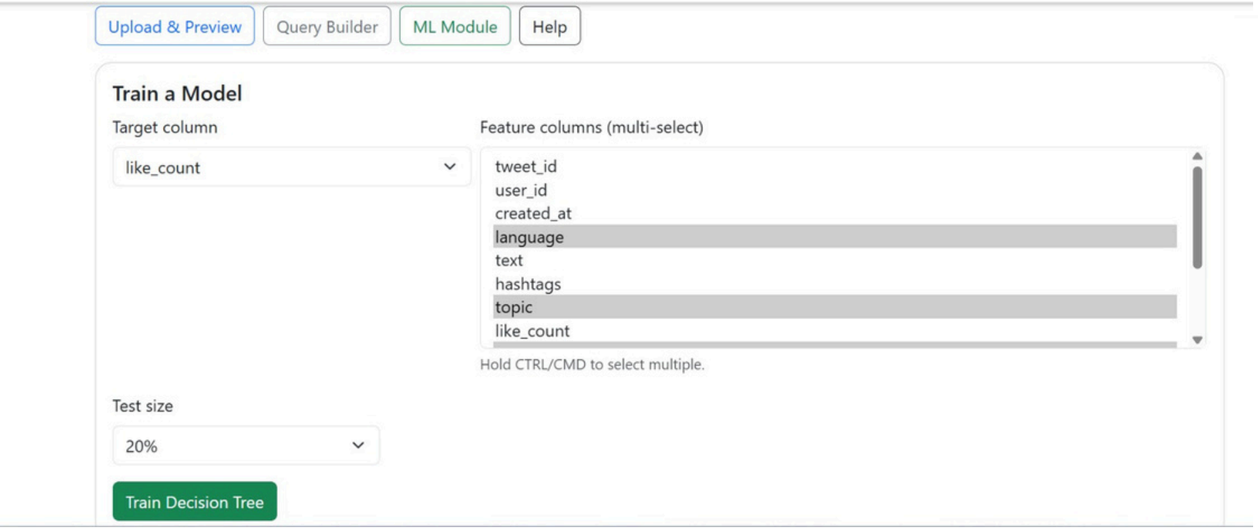
Query the current dataset via Spark SQL view `current_df`.

```
e.g. SELECT language, COUNT(*) AS cnt FROM current_df GROUP BY
language ORDER BY cnt DESC
```

Run Query

Results (17 rows; showing up to 1000)

topic	avg_score
TRAVEL	0.554879
HEALTH	0.541009
TECH	0.532915
travel	0.532389
SPORTS	0.527286
sports	0.526754
ENTERTAINMENT	0.524609
EDUCATION	0.522071
politics	0.522038
FINANCE	0.518724
None	0.517053
entertainment	0.515776
health	0.510022

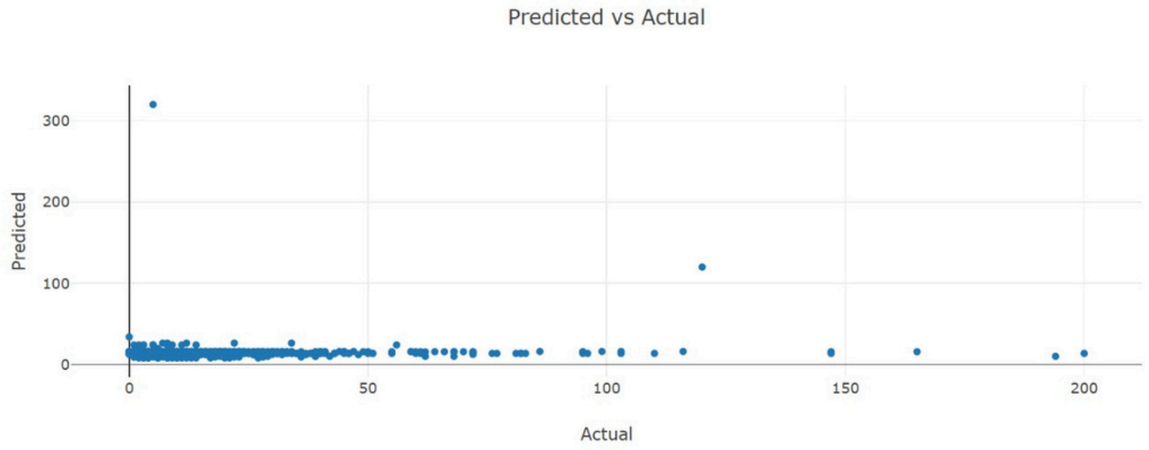


Model Report

Trained DecisionTreeRegressor with 4 feature(s). Test size: 20%.

- **RMSE:** 22.5655
- **MAE:** 12.2355
- **R^2:** -0.245

Chart (Evaluation)



Predict & Download

Predict on FULL dataset

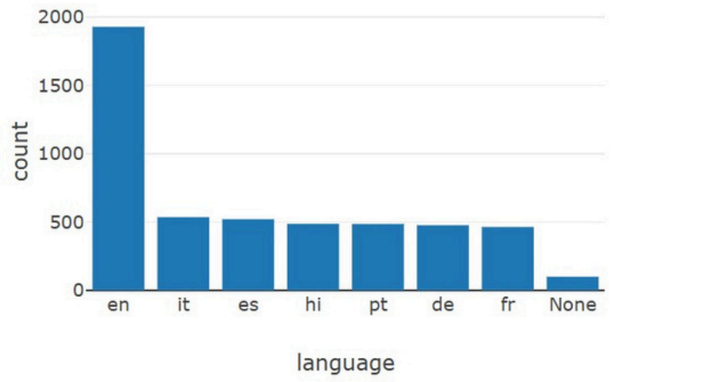
Download Predictions CSV

Uses the latest trained model to predict for every row in the current datase

Quick Chart

Column name (e.g., language, topic)

Draw



CHALLENGES FACED

- HANDLING LARGE-SCALE UNSTRUCTURED TEXT DATA
- ENSURING BALANCED CLASS DISTRIBUTION
- MODEL PERFORMANCE TUNING

FUTURE WORK

- HYPERPARAMETER TUNING
- ENSEMBLE METHODS
- REAL-TIME ANALYSIS WITH SPARK STREAMING
- MULTI-LABEL CLASSIFICATION & TOPIC MODELING
- ADVANCED VISUALIZATION DASHBOARD

CONCLUSION

- SUCCESSFUL DEMONSTRATION OF END-TO-END PYSPARK ML PIPELINE
- EFFECTIVE CLASSIFICATION OF TWEET SENTIMENTS & CATEGORIES
- SCALABILITY FOR LARGE DATASETS

THANK YOU