BIG DATA PROJECT:
AI TWEETS SENTIMENT ANALYSIS

Sedera RASOANAIVO

AGENDA

2

OBJECTIVES & DOMAIN KNOWLEDGE

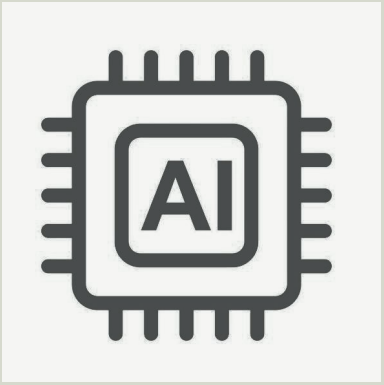
WORKFLOW

RAW TWEETS ANALYSIS

ML ANALYSIS

CONCLUSIONS & CHALLENGES

BACKGROUND & OBJECTIVE



OBJECTIVES:

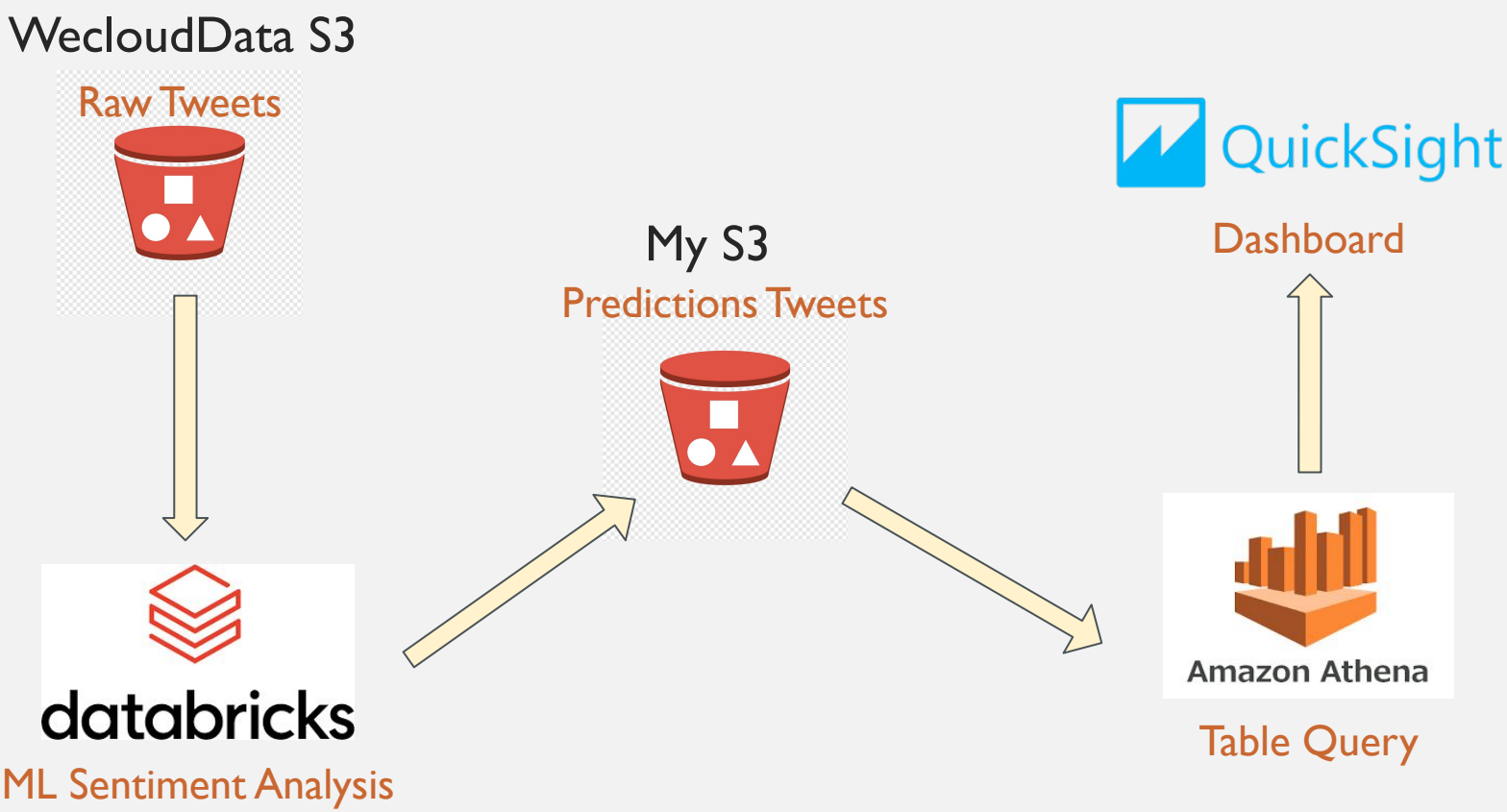
- Performing Sentiment Analysis related to the AI topic
- Creating Dashboard
- Withdrawing Insights.

BACKGROUNDS:

- Artificial Intelligence (AI): branch of computer science that makes machines think and act clever mimicking human
- AI has a range of applications with the potential to transform how we work and our daily lives
- It generates different reaction to people

WORKFLOW

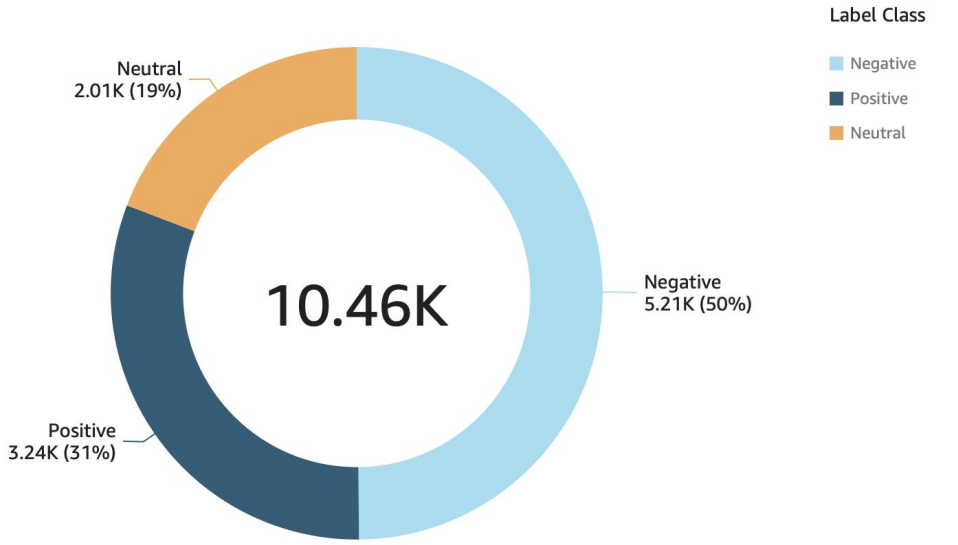
4


WecloudData S3
Raw Tweets
databricks
ML Sentiment Analysis
My S3
Predictions Tweets
QuickSight
Dashboard
Amazon Athena
Table Query

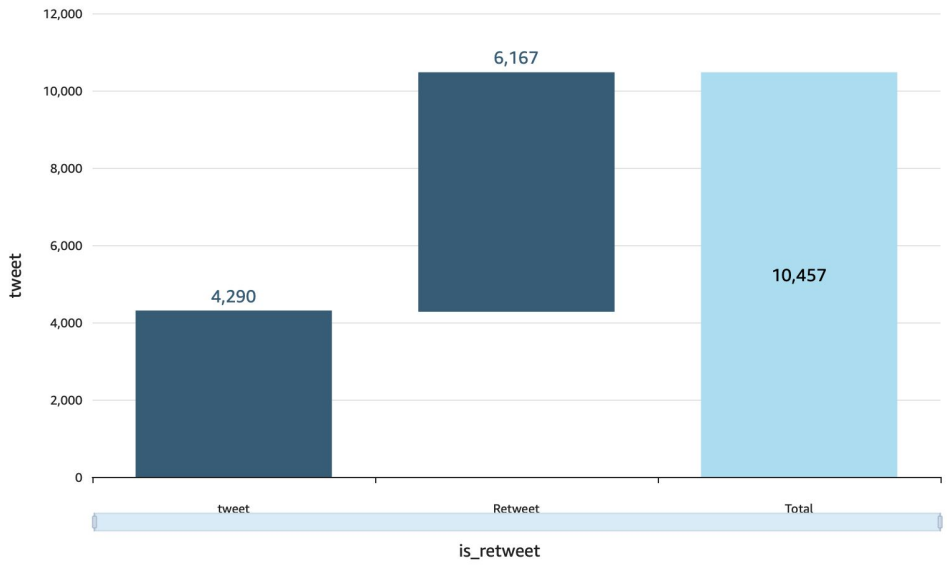
TWEETS COUNT AND DISTRIBUTION

5

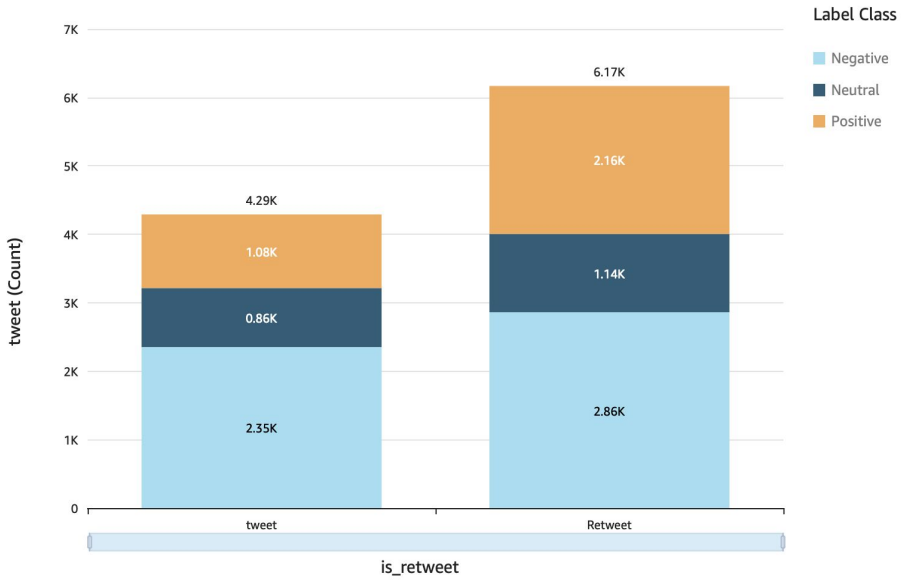
Tweets Sentiment Distribution



Count of first tweets and retweets



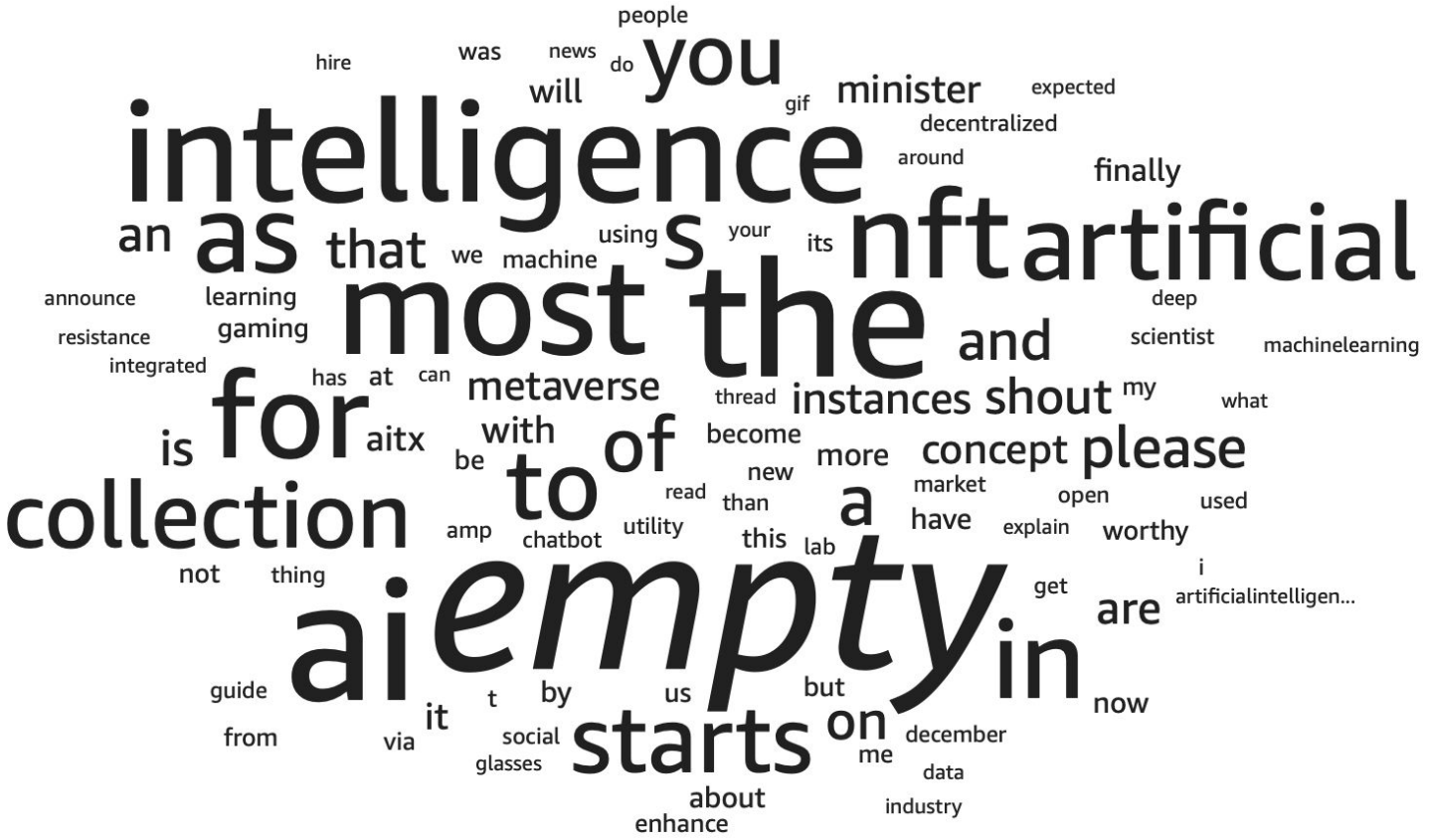
Count of tweets by type and label



WORDCLOUD

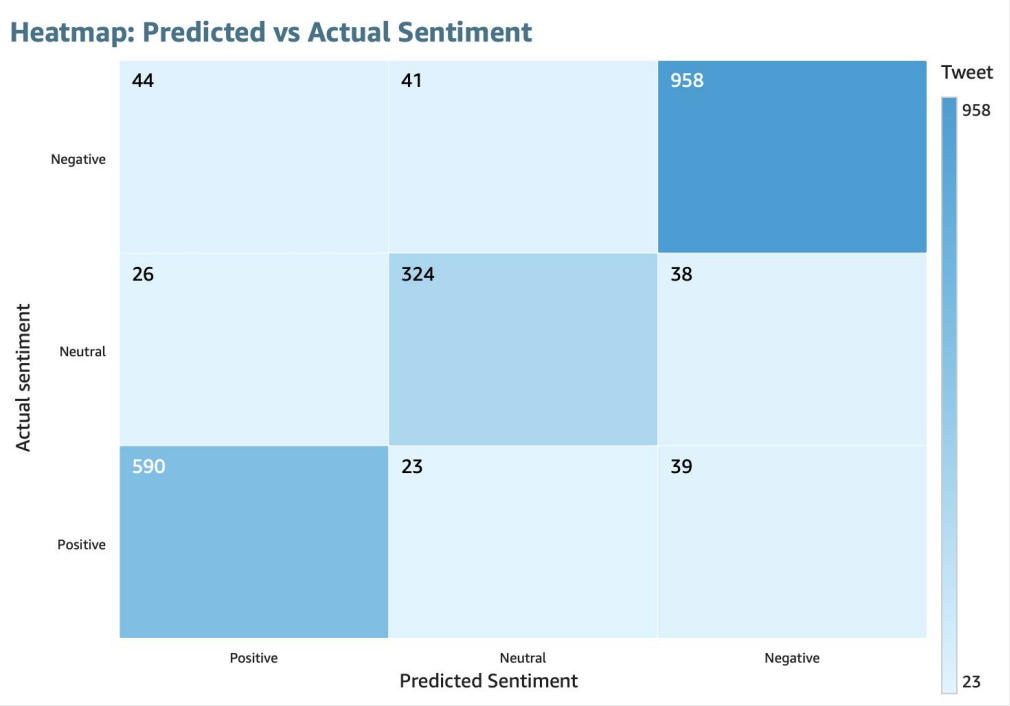
Word Cloud of most frequent words

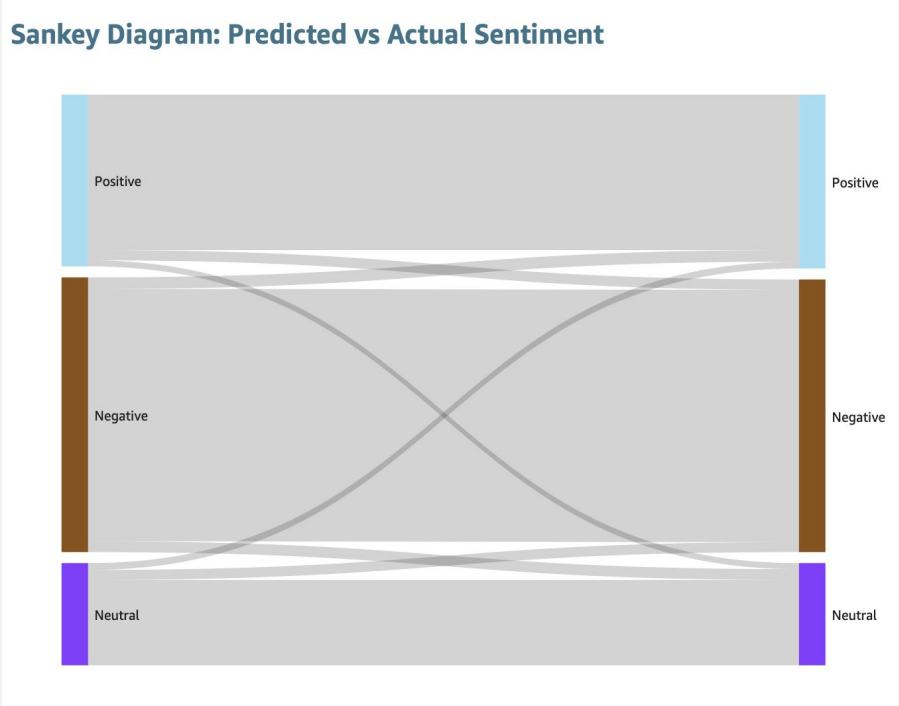
SHOWING TOP 100 IN WORDCOLUMN



ML ANALYSIS: PREDICTION EVALUATION

7





Model	Accuracy	F1 Score
Logistic Regression	0.898	0.898
Naive Bayes	0.821	0.819
Random Forest	0.639	0.584

CONCLUSIONS & CHALLENGES

8

CONCLUSIONS:

- Sentiment mostly negative
- Logistic Regression better than Random Forest
- Good performance on predicting sentiment even without tuning

CHALLENGES:

- Limitation on handling bigger dataset with multiclass: Cluster terminated
- Splitting the tweets into smaller chunks of words in Athena or QuickSight

THANK YOU