Big Data Project

Sentiment Analysis of tweets regarding AI

Introduction

- This projects objective was to build a sentiment analysis model using tweets retrieved from the web and present an analysis of the resulting dataset and model.
- The theme of the tweets used was ‘AI’, which refers to Artificial Intelligence.
- The data used was retrieved from one of Weclouddata’s public folders avaliable through Amazon Simple Storage Service (AWS S3).
- The date of the tweets analyzed range from December 08 to 09, 2022.

Workflow

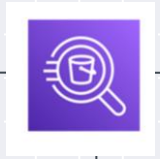
Amazon S3
WeCloudData's
Bucket

databricks
Cloud computing

Distributed data
processing

Machine
learning

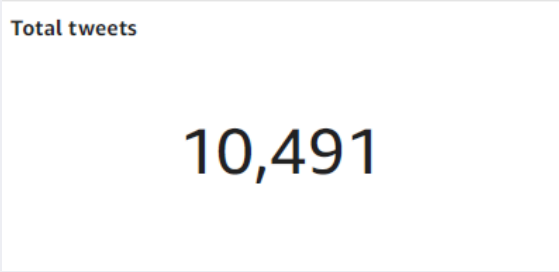
Amazon S3
Personal bucket

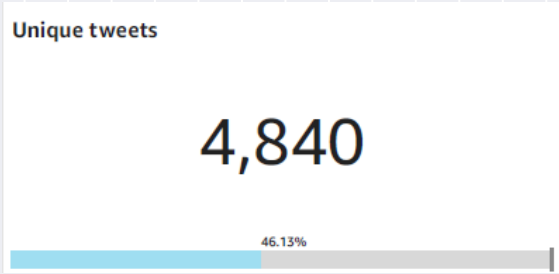


Amazon Athena
Organizing
tabular data

Build
visualizations

Analysis

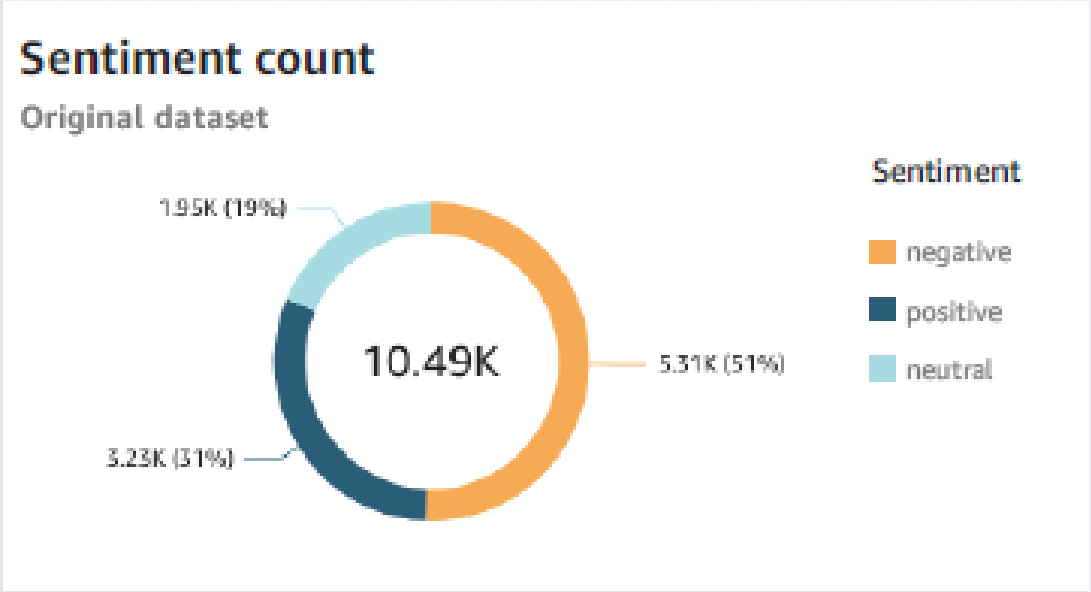




- The dataset had a total of 10,491 tweets.
- 4,840 of these (46,13%) were unique.

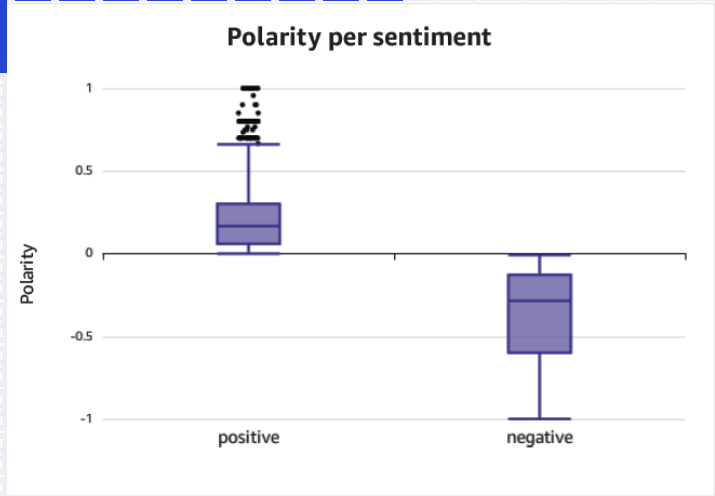
Analysis

- **Positive:** 31% of the data (~ 3230 tweets).
- **Neutral:** 19% of the data (~ 1950 tweets).
- **Negative:** 51% of the data (~ 5310 tweets).

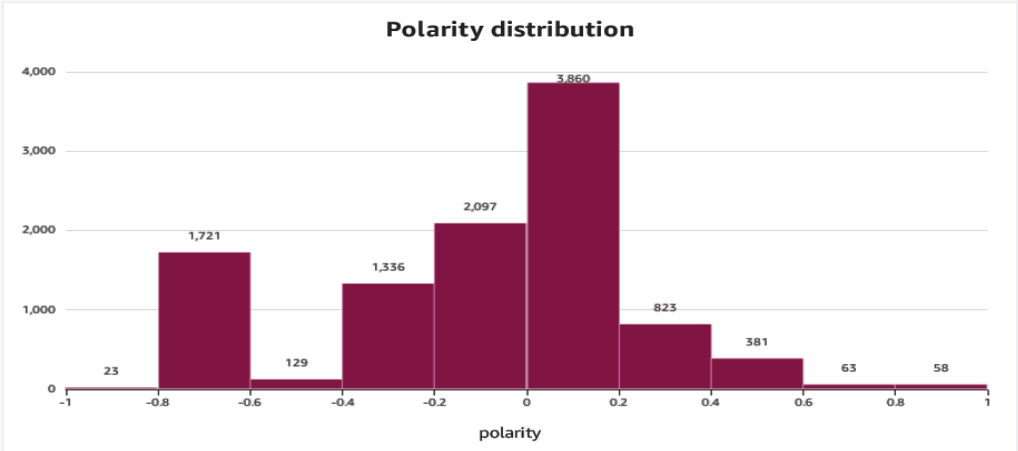


- 4,840 of these (46,13%) were unique.

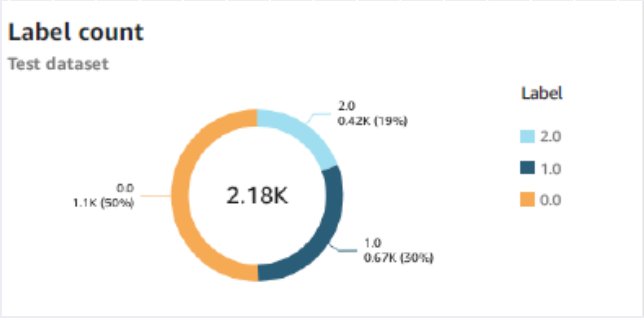
Analysis

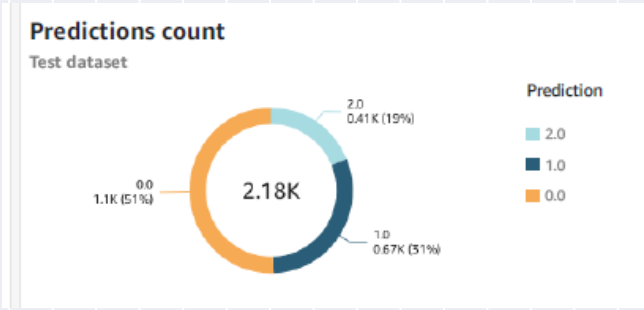


- Median absolute value for positive polarity was lower than the one for negative polarity.
- Extreme positive manifestation was way rarer than negative ones.



Analysis





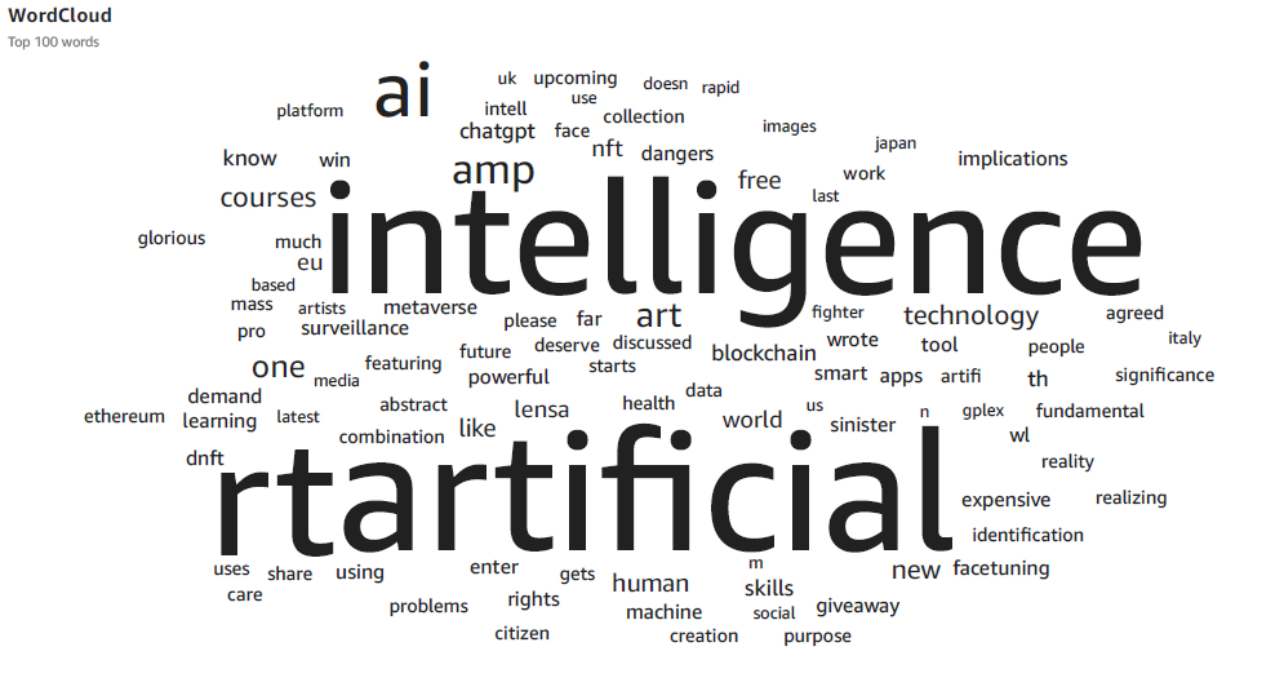
Three models were evaluated on a validation dataset:

- ➔ Logistic Regression (1 ngram tfidf)
- ➔ Decision Tree (1 ngram tf_idf)
- ➔ Random Forest (1 ngram tf_idf)

The Logistic Regression performed best on the validation set and was then scored on the test set.

LR Model scores on test data			
Acurracy	**Weighted precision**	**Weighted recall**	**F1**
0.9047	**0.9045**	**0.9047**	**0.9046**

Analysis


WordCloud
Top 100 words
ai
intelligence
rtartificial
amp
art
platform
uk
upcoming
doesn
rapid
use
intell
collection
chatgpt
face
images
japan
nft
dangers
know
win
free
work
implications
courses
last
glorious
much
eu
based
mass
artists
metaverse
please
far
fighter
technology
agreed
pro
surveillance
future
deserve
discussed
blockchain
wrote
tool
people
italy
one
media
featuring
powerful
starts
smart
apps
artifi
th
significance
demand
ethereum
learning
latest
abstract
lensa
health
data
us
n
gplex
fundamental
combination
like
world
sinister
wl
dnft
reality
expensive
realizing
identification
uses
share
using
enter
gets
human
m
new
facetuning
care
problems
rights
skills
social
giveaway
machine
citizen
creation
purpose

Chalenges

- Adapting code to pyspark context.
- Managing AWS functionalities.

Conclusions

Best Model

Logistic Regression was was the best model when predicting tweets' sentiments (accuracy= 90% f1 = 90%).

Sentiment frequency

Sentiments represented in the dataset were mostly negative (~51% negative, ~19% neutral, 31% positive).

Polarity

Negative manifestations of sentiments towards AI were more polarized than the expression of posivite sentiments.

Thanks!