PROJECT REPORT

SUBMITTED FOR

Conversational AI: Accelerated Data Science [Basics]
(UCS546)

BY

NAME OF THE STUDENT	ROLL NUMBER
Manav Singh	102283005
Nipun Tank	102153011
Prabhmehar Pal Singh Bedi	102165002

SUBMITTED TO

DR. JASMEET SINGH

COMPUTER SCIENCE ENGINEERING DEPARTMENT

THAPAR INSTITUTE OF ENGINEERING & TECHNOLOGY,
PATIALA-147004, PUNJAB
INDIA

JULY - DEC (2023)

Project Overview

In this sentiment analysis project, articles are systematically scraped from a website, and sentiment analysis techniques are applied to unveil the emotional tone of the content. The project culminates in the creation of an interactive dashboard that vividly displays sentiment analysis results.

Technique used for Data Collection

1. HTML Scraping:

- The code uses the ‘requests’ library to make HTTP requests to the specified URLs.

- For each URL, the HTML content of the web page is retrieved.

2. BeautifulSoup Parsing:

- The ‘BeautifulSoup’ library is employed to parse the HTML content and extract specific information.

- It searches for a ‘ ’ element with the class 'td-post-content' using BeautifulSoup's ‘find’ method.

3. File Storage:

- The HTML content is stored in individual files, where each file is named after a unique identifier extracted from the URL.

Technique used for Data Description

- The dataset being collected contain information from different articles and web pages, and each URL is associated with a specific piece of content.

- The columns in the dataset might include the following:

  - URL_ID: A unique identifier extracted from the URL (e.g., a part of the URL split).

  - URL: The web address of the article or content.

- The HTML content retrieved from each URL is stored in separate files within the 'data' folder. These HTML files presumably contain the text content, which is later processed in subsequent sections of the code and observations are stored in ‘convoproj.csv’.

Data Pre-Processing Technique used

1. Tokenization:

  - The text data is tokenized using the ‘word_tokenize’ function from the ‘nltk.tokenize’ module. Tokenization breaks down the text into individual words or tokens.

2. Stopword Removal:

  - Stopwords (common words like "the," "and," "is") are removed from the tokenized text using a list of English stopwords from the ‘nltk.corpus’ module.

3. Text Cleaning:

  - The code iterates through each row of the DataFrame, processes the text data, and creates a new DataFrame (‘df_processed’) with the processed text.

  - For each row, it tokenizes the text, removes stopwords, and joins the filtered tokens back into a sentence.

4. New DataFrame Creation:

- A new DataFrame (‘df_processed’) is created with columns ‘Title’ and ‘Text’, where ‘Text’ contains the processed and cleaned text data.

Resulting DataFrame (‘df_processed’)

- The DataFrame ‘df_processed’ contains two columns: ‘Title’ and ‘Text’.

- ‘Title’ corresponds to the titles of the articles.

- ‘Text’ contains the processed and cleaned text data after tokenization and stopword removal.

Snapshots of Dashboard

Link to Dashbaord - Click Here

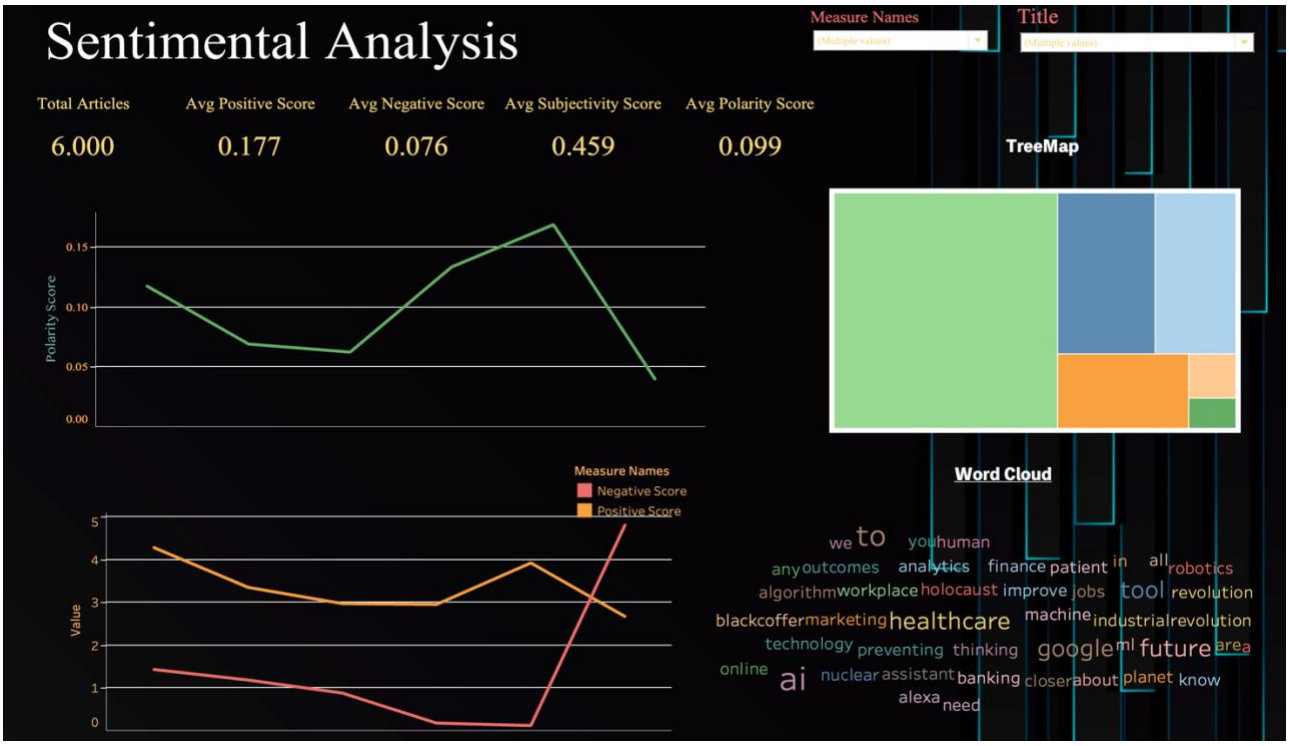
OR

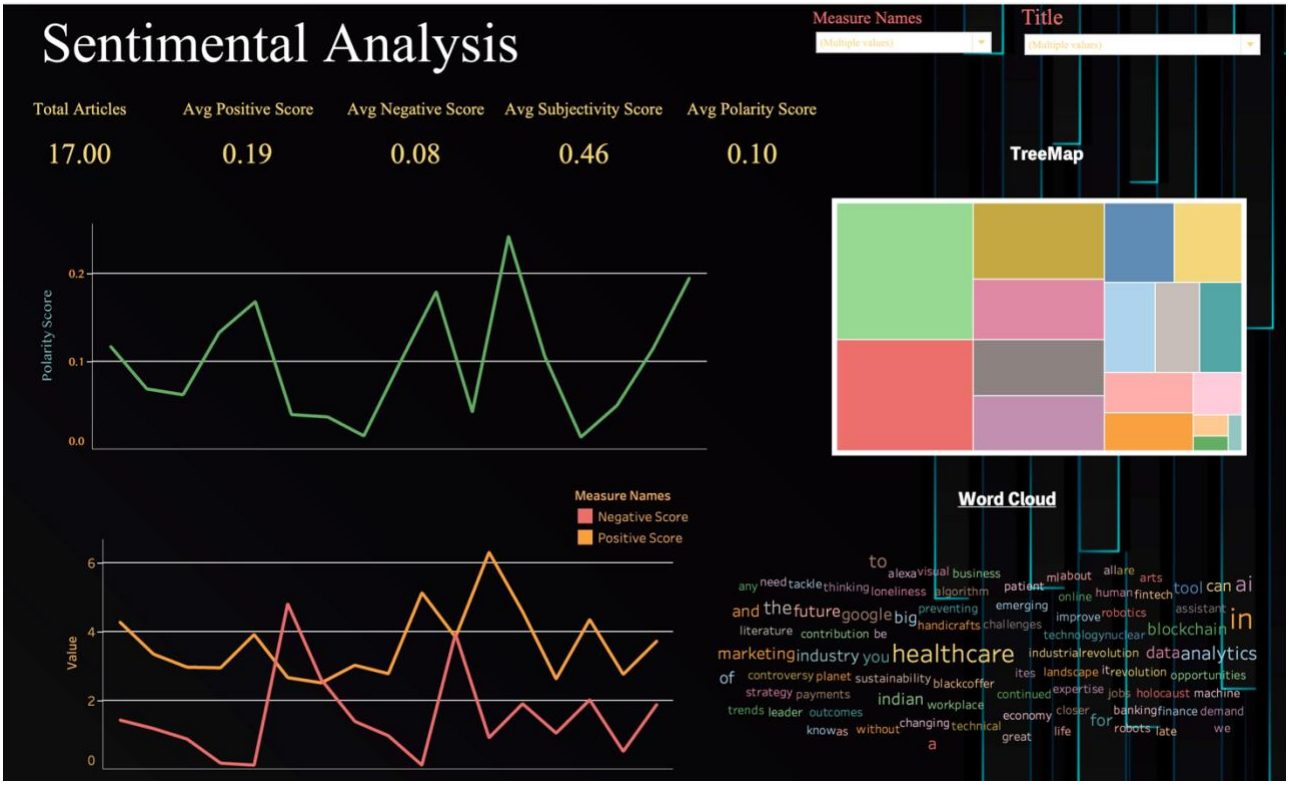
https://public.tableau.com/views/Book2_17031721966030/Dashboard1?:language=en-US&publish=yes&:display_count=n&:origin=viz_share_link

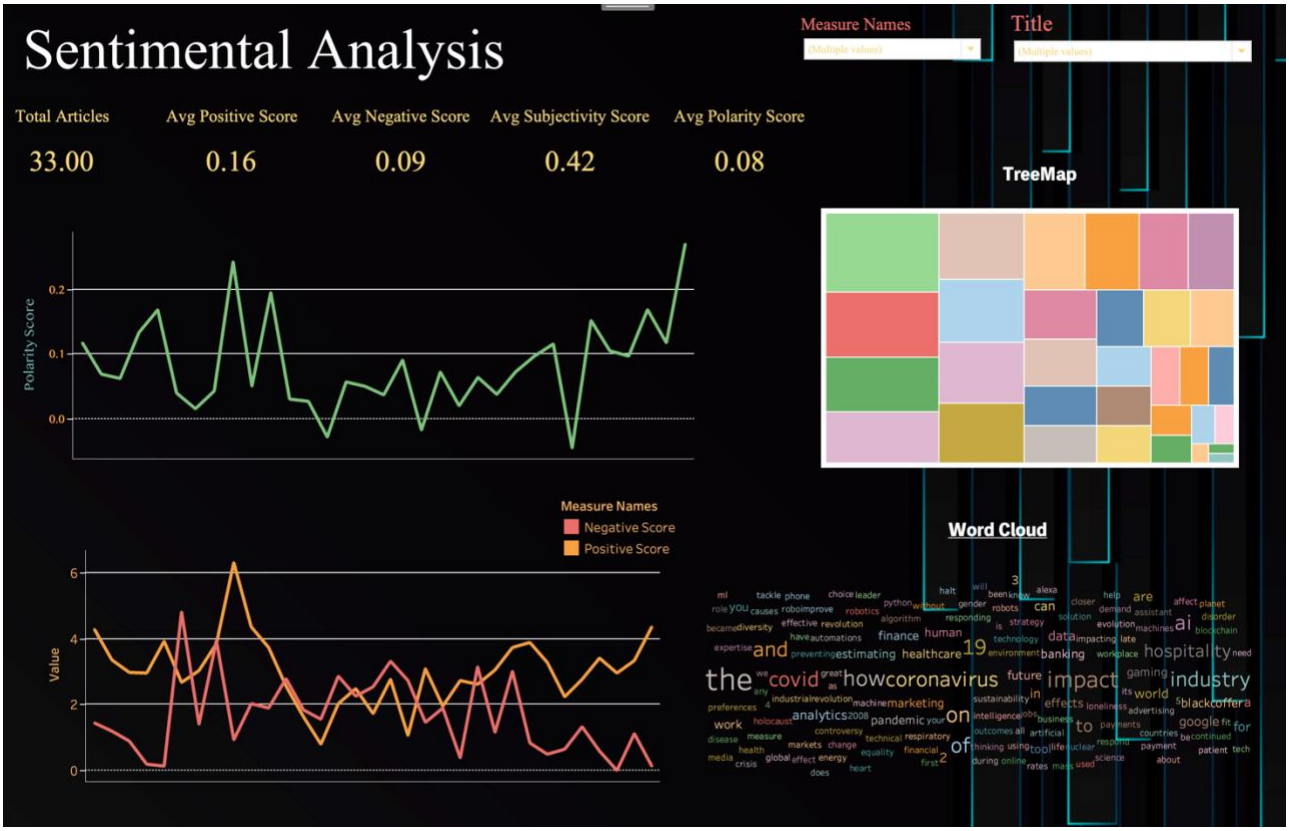
About Dashbaord -:

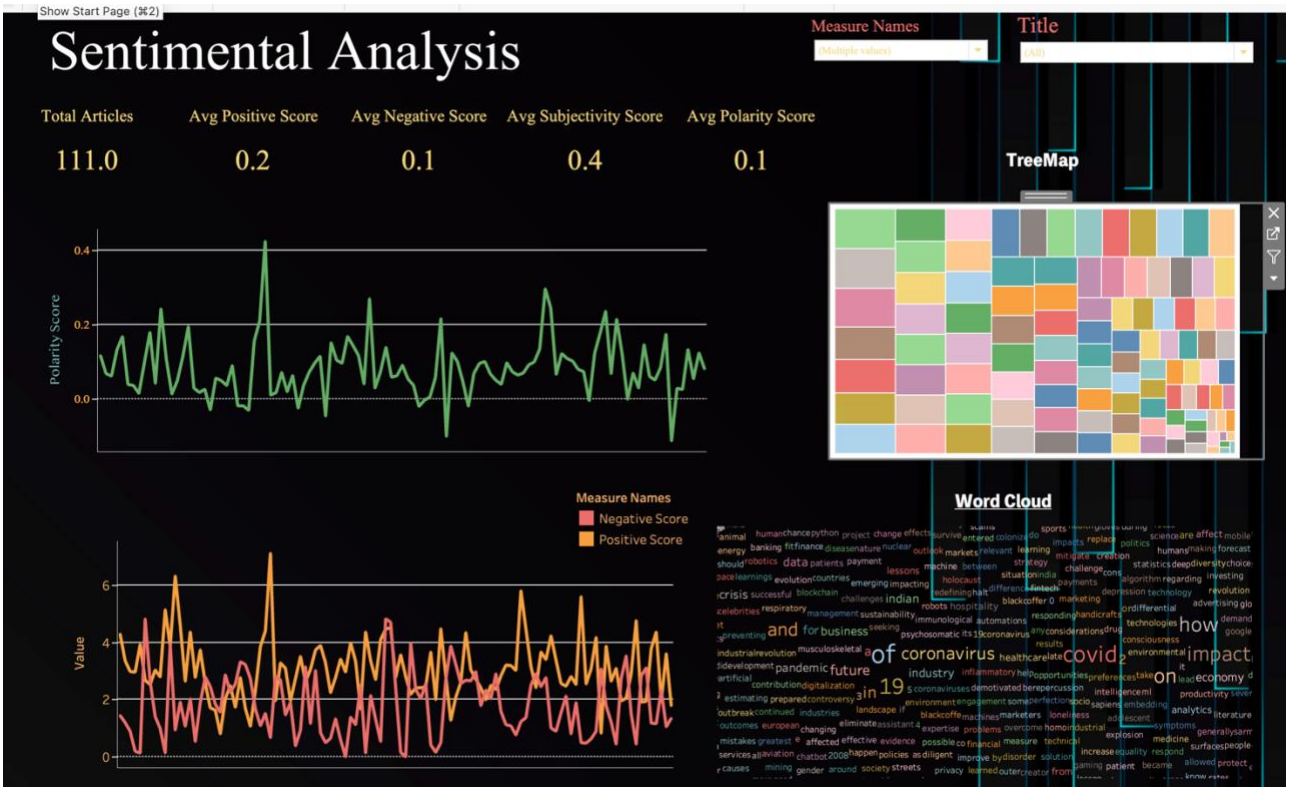
In Dashboard we have the option to choose articles based on their titles using a dropdown menu. Upon selecting a specific article, you will receive an overview of the sentiment analysis, including positive and negative scores, polarity scores, subjective score, a treemap, a word cloud, and plots comparing various articles and their corresponding scores.


Sentimental Analysis
Measure Names
Title
Total Articles
Avg Positive Score
Avg Negative Score
Avg Subjectivity Score
Avg Polarity Score
1.0000
0.2260
0.0760
0.5128
0.1178
TreeMap
Polarity Score
0.10
0.05
0.00
Measure Names
Negative Score
Positive Score
Value
4
3
2
1
0
Word Cloud
revolution
technology
ai
industrial
revolution
ml
lgorithm
healthcare


Sentimental Analysis
Measure Names
Title
Total Articles
Avg Positive Score
Avg Negative Score
Avg Subjectivity Score
Avg Polarity Score
6.000
0.177
0.076
0.459
0.099
TreeMap
Polarity Score
0.15
0.10
0.05
0.00
Measure Names
Negative Score
Positive Score
Value
5
4
3
2
1
0
Word Cloud
we
to
you
human
any
outcomes
analytics
finance
patient
in
all
robotics
algorithm
workplace
holocaust
improve
jobs
tool
revolution
blackcoffer
marketing
healthcare
machine
industrial
revolution
technology
preventing
thinking
google
ml
future
area
online
ai
nuclear
assistant
banking
closer
about
planet
know
alexa
need


Sentimental Analysis
Measure Names
Title
Total Articles
Avg Positive Score
Avg Negative Score
Avg Subjectivity Score
Avg Polarity Score
17.00
0.19
0.08
0.46
0.10
TreeMap
Polarity Score
Measure Names
Negative Score
Positive Score
Value
Word Cloud


Sentimental Analysis
Measure Names
Title
Total Articles
Avg Positive Score
Avg Negative Score
Avg Subjectivity Score
Avg Polarity Score
33.00
0.16
0.09
0.42
0.08
TreeMap
Polarity Score
Measure Names
Negative Score
Positive Score
Value
Word Cloud


Show Start Page (⌘2)
Sentimental Analysis
Measure Names
Title
Total Articles
Avg Positive Score
Avg Negative Score
Avg Subjectivity Score
Avg Polarity Score
111.0
0.2
0.1
0.4
0.1
TreeMap
Polarity Score
0.4
0.2
0.0
Measure Names
Negative Score
Positive Score
Value
6
4
2
0
Word Cloud