

**PROJECT REPORT**  
**SUBMITTED FOR**  
**Conversational AI: Accelerated Data Science [Basics]**  
**(UCS546)**

**BY**

**NAME OF THE STUDENT**

**ROLL NUMBER**

Manav Singh  
Nipun Tank  
Prabhmehar Pal Singh Bedi

102283005  
102153011  
102165002

**SUBMITTED TO**  
**DR. JASMEET SINGH**



**THAPAR INSTITUTE**  
OF ENGINEERING & TECHNOLOGY  
(Deemed to be University)

**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 '<div>' 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 Dashboard - [Click Here](#)**

**OR**

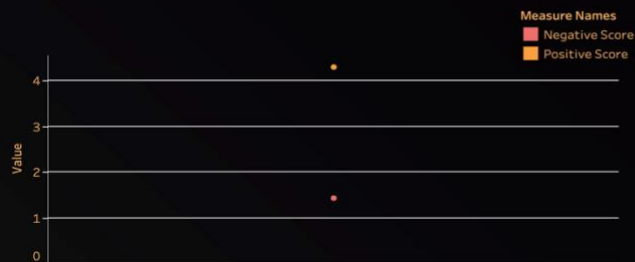
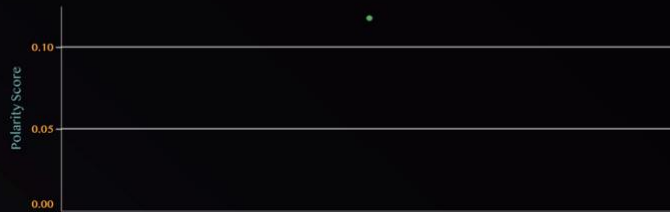
**[https://public.tableau.com/views/Book2\\_17031721966030/Dashboard1?:language=en-US&publish=yes&:display\\_count=n&:origin=viz\\_share\\_link](https://public.tableau.com/views/Book2_17031721966030/Dashboard1?:language=en-US&publish=yes&:display_count=n&:origin=viz_share_link)**

About Dashboard -:

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

Total Articles: 1.0000  
 Avg Positive Score: 0.2260  
 Avg Negative Score: 0.0760  
 Avg Subjectivity Score: 0.5128  
 Avg Polarity Score: 0.1178



Measure Names: (Multiple values) Title: (Multiple values)

TreeMap

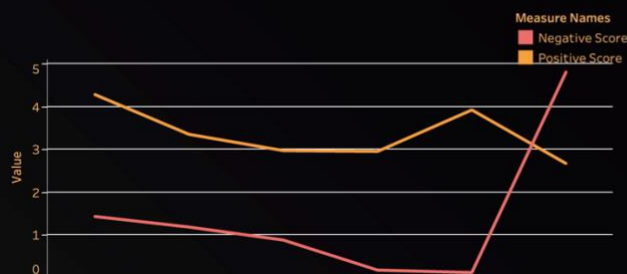
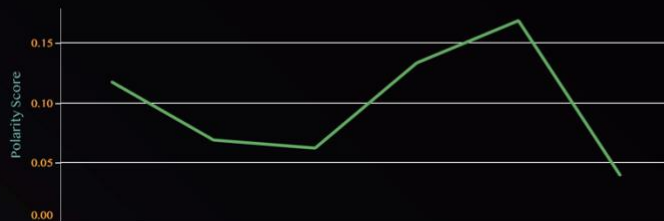


Word Cloud

google  
 revolution technology  
 ai industrial revolution ml  
 algorithm healthcare

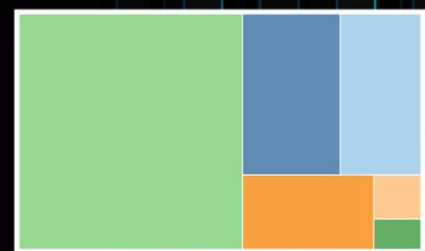
# Sentimental Analysis

Total Articles: 6.000  
 Avg Positive Score: 0.177  
 Avg Negative Score: 0.076  
 Avg Subjectivity Score: 0.459  
 Avg Polarity Score: 0.099



Measure Names: (Multiple values) Title: (Multiple values)

TreeMap



Word Cloud

we to you human  
 any outcomes analytics finance patient in all robotics  
 algorithm workplace holocaust improve jobs tool revolution  
 black coffee marketing healthcare machine industrial revolution  
 technology preventing thinking google ml future area  
 online ai nuclear assistant banking closer about planet know  
 alexa need

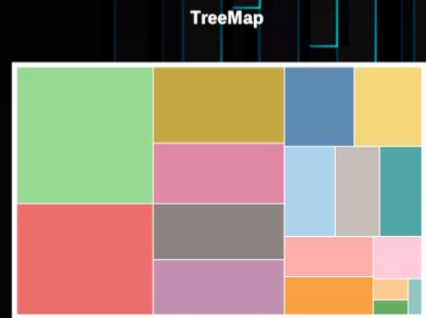
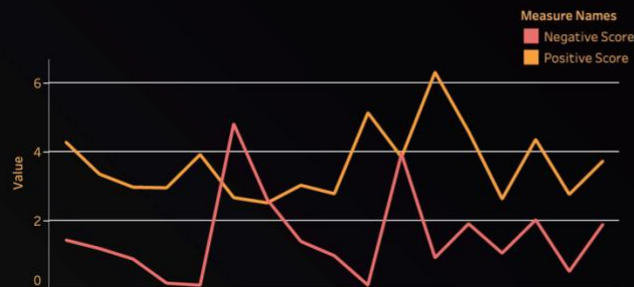
# Sentimental Analysis

Measure Names Title

(Multiple values) (Multiple values)

Total Articles Avg Positive Score Avg Negative Score Avg Subjectivity Score Avg Polarity Score

17.00 0.19 0.08 0.46 0.10



Word Cloud

to alexa visual business patient mabout allare arts tool can ai  
any need tackle thinking loneliness algorithm preventing emerging online human fintech assistant  
and the future google big handicrafts challenges improve robotics blockchain in  
literature contribution be technology nuclear data analytics  
marketing industry you healthcare ites landscape it revolution opportunities  
of controversy planet sustainability blackcoffer jobs holocaust machine  
strategy payments indian workplace economy closer banking finance demand  
trends leader outcomes knowas without changing technical a great life for robots late we

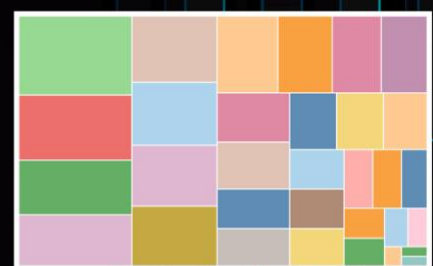
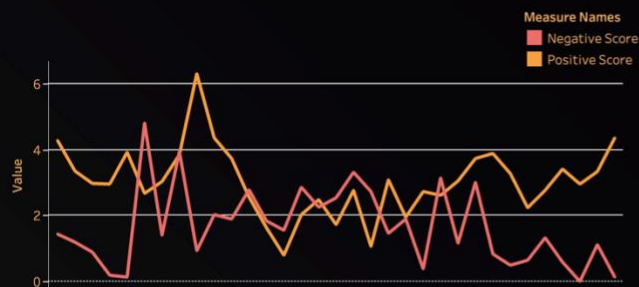
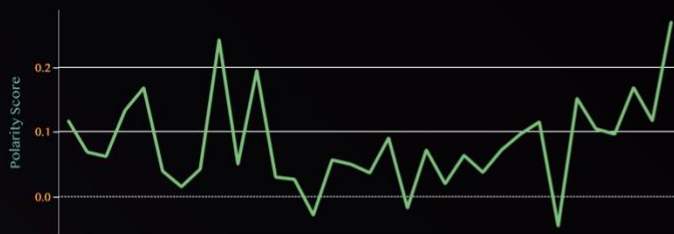
# Sentimental Analysis

Measure Names Title

(Multiple values) (Multiple values)

Total Articles Avg Positive Score Avg Negative Score Avg Subjectivity Score Avg Polarity Score

33.00 0.16 0.09 0.42 0.08



Word Cloud

ml tackle phone choice leader python without gender robots can closer help are affect planet disorder  
role YOU causes robot improve robotics responding is strategy solution evolution machines ai blockchain  
became diversity effective revolution finance human 19 environment banking workplace hospitality need  
expertise and preventing estimating healthcare 19 environment banking workplace hospitality need  
the covid great as how coronavirus sustainability in impact its world industry  
preferences 4 industrial revolution machine marketing on intelligence business to payments google fit for  
work holocaust analytics 2008 pandemic your on outcomes all artificial nuclear response countries continued  
disease measure markets change technical respiratory financial of thinking using top life science payment patient tech  
media health global effect energy equality financial first 2 during online rates made used about

Show Start Page (#2)

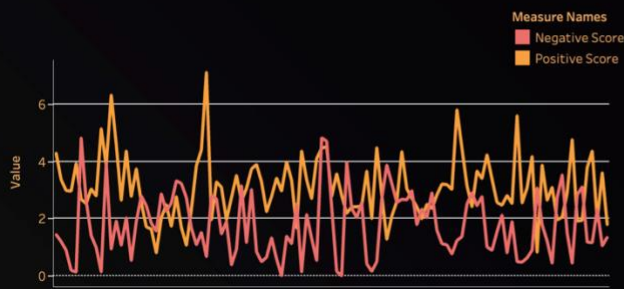
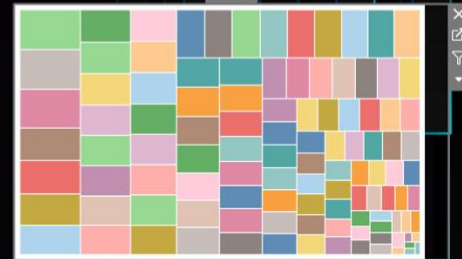
# Sentimental Analysis

Total Articles 111.0 Avg Positive Score 0.2 Avg Negative Score 0.1 Avg Subjectivity Score 0.4 Avg Polarity Score 0.1

Measure Names

Title

TreeMap



Word Cloud

animal human chance python project change effects survive entered colonize sports impacts replace science are affect mobile energy banking life financial depression nuclear outlook markets relevant learning mitigate creation human making forecast shout robotics data patients payment lessons machine between strategy challenge statistics deep diversity choice peak earnings evolution countries emerging impacting additio... blackpuffer 0 marketing depression technology advertising glo... / crisis successful blockchain challenges indian robots hospitality blackpuffer 0 marketing depression technology advertising glo... zentences respiratory management sustainability immunological automations responding handicraft technologies how demand... preventing and for business seeking psychosomatic its 1 coronavirus 0 considerations trust consciousness google... industrial revolution musculoskeletal of coronavirus healthcare covid2 environment impact... development pandemic future industry inflammatory heliographun his preference take on lead economy d... artificial contribution digitalization 31 n 19 s coronavirus demotivated beneperous on in the percent productivity lever... 7 estimating prepared controversy 31 n 19 s coronavirus demotivated beneperous on in the percent productivity lever... outbreak continued industries landscape if blackcoff machines marketers loneliness adolescent analytics literature... outcomes european changing eliminate assistant 4 expertise problems overcome homonidustrial explosion medicine generally am... mistakes greatest 4 affected effective evidence possible co financial measure technical increase equality respond surfaces people... services generation curbot 2008 happy policies as diligent improve by disorder solution gaming patient became allowed protect... causes mining gender arouse society streets privacy learned outer creator from know eyes