VISVESVARAYA TECHNOLOGICAL UNIVERSITY

BELAGAVI-590018

Internship Report (18EC185)

On

"TWITTER DATA SENTIMENTAL ANALYSIS"

Submitted in partial fulfillment of the requirements for the award of the degree of

Bachelor of Engineering

in

Electronics and Communication Engineering

Visvesvaraya Technological University, Belagavi

Submitted by

Aditi Jaiswal
1DT20EC003

Under the Guidance of

Dr. Ravikumar HC
Asst. Professor, Dept of ECE

Department of Electronics and Communication Engineering
Accredited by NBA, New Delhi.
DAYANANDA SAGAR ACADEMY OF TECHNOLOGY AND MANAGEMENT
Accredited by NAAC with Grade A+
Udayapura, Kanakapura Road, Bengaluru-560082

2023-2024

DAYANANDA SAGAR ACADEMY OF TECHNOLOGY AND MANAGEMENT

Udayapura, Kanakapura Road, Bengaluru-560082
Accredited by NAAC with Grade A+

2023-2024

Department of Electronics and Communication Engineering

Accredited by NBA, New Delhi

CERTIFICATE

This is to certify that the internship work on **Twitter Data Sentimental Analysis** carried out by **Aditi Jaiswal 1DT20EC003**, a bonafide student of **DAYANANDA SAGAR ACADEMY OF TECHNOLOGY AND MANAGEMENT, Bengaluru** in partial fulfilment for the award of the degree of **Bachelor of Engineering in Electronics and Communication Engineering** of the **Visvesvaraya Technological University, Belagavi,** during the year **2023-24**. It is certified that all corrections/suggestions indicated for internal assessment have been incorporated in the report deposited in the departmental library. The internship report has been approved as it satisfies the academic requirements with respect to internship work prescribed for the award of Bachelor of Engineering Degree.

Signature of the guide	Signature of the HoD	Signature of the Principal
Dr. Ravikumar HC	**Dr. Mallikarjun PY**	**Dr. M Ravishankar**

External Viva

Sl. No. Name of the Examiner Signature with date

1.
2.

ACKNOWLEDGEMENT

We extend our deepest gratitude and appreciation to **DAYANANDA SAGAR ACADEMY OF TECHNOLOGY AND MANAGEMENT** for providing us with the invaluable opportunity to pursue a B.E in Electronics and Communication Engineering, thereby playing a pivotal role in shaping our career paths. We are sincerely grateful to **Dr. M Ravishankar**, Principal, DSATM, Bengaluru, for his unwavering encouragement and support throughout our academic journey.

Our heartfelt thanks go to **Dr. Mallikarjun PY**, Professor and Head of the Department of Electronics and Communication Engineering, whose constant motivation and guidance have propelled us to strive for excellence and aim higher.

We would also like to express our sincere gratitude to our internship coordinator, **Dr. R. V. Manjunath**, Assistant Professor in the Department of Electronics and Communication Engineering, for their tireless motivation, unwavering support, and invaluable guidance.

Furthermore, we extend our utmost appreciation to our Guide, **Dr. Ravikumar HC**, Assistant Professor in the Department of Electronics and Communication Engineering, for his continuous guidance and mentorship.

We are also deeply thankful to all the esteemed faculty members of the Department of Electronics and Communication Engineering for their invaluable assistance and unwavering encouragement throughoutour academic pursuits.

Yours sincerely

Aditi Jaiswal (1DT20EC003)

EXECUTIVE SUMMARY

Purpose:

The purpose of this project is to delve into the vast trove of COVID-19-related tweets circulating on Twitter and conduct a comprehensive sentiment analysis. By harnessing the power of natural language processing (NLP) techniques and machine learning algorithms, this endeavor aims to provide valuable insights into the prevailing emotions, attitudes, and perceptions of Twitter users towards the ongoing COVID-19 pandemic. This analysis is essential for understanding the societal response to the crisis, identifying emerging trends and concerns, and informing decision-making processes in public health, policy-making, and crisis management.

Methods:

The methodology employed in this project revolves around the systematic collection, preprocessing, and analysis of COVID-19-related tweets sourced from Twitter's expansive database. Leveraging the Twitter API, tweets containing pertinent keywords, hashtags, and mentions related to COVID-19, such as "#COVID19," "coronavirus," and "pandemic," were retrieved in real-time. These tweets underwent rigorous preprocessing using a combination of NLP techniques, including tokenization to break down text into individual words, stop word removal to eliminate common, non-informative words, and sentiment lexicon-based analysis to assign sentiment scores to each tweet. Additionally, machine learning algorithms, such as Naive Bayes or LSTM networks, were deployed for sentiment classification, categorizing tweets into positive, negative, or neutral sentiments. This comprehensive approach allowed for a nuanced exploration of the emotional landscape surrounding the COVID-19 discourse on Twitter.

Importance:

The COVID-19 pandemic has catalyzed unprecedented global challenges, spanning public health, economic, and social domains. In this context, monitoring public sentiment towards the pandemic is paramount for several reasons. Firstly, understanding the prevailing sentiments and attitudes of the populace provides valuable feedback on the effectiveness of public health interventions, government policies, and crisis communication strategies. Moreover, sentiment analysis enables the detection of emerging concerns, misinformation, and societal tensions, thereby facilitating proactive responses and

mitigating potential risks. Additionally, insights gleaned from sentiment analysis can inform targeted public health messaging, foster community engagement, and foster trust and solidarity amidst uncertainty and adversity.

Findings/Results:

The sentiment analysis yielded a wealth of insights into the multifaceted and dynamic nature of public sentiment towards COVID-19 on Twitter. Analysis of sentiment polarity revealed a spectrum of emotions ranging from hope and solidarity to fear and frustration. Positive sentiments were often associated with expressions of gratitude towards frontline healthcare workers, messages of resilience, and news of scientific breakthroughs, such as vaccine developments. Conversely, negative sentiments were commonly expressed towards government responses, perceived misinformation, and the socio-economic impacts of the pandemic, including unemployment and mental health struggles. Neutral sentiments were observed in tweets disseminating factual information, updates on COVID-19 statistics, and public health advisories. Furthermore, sentiment analysis conducted on specific topics, keywords, or hashtags illuminated nuanced sentiment shifts and trends over time, providing valuable contextual insights into the evolving COVID-19 discourse.

Conclusions & Recommendations:

In conclusion, the sentiment analysis of COVID-19 tweets on Twitter underscores the complexity and fluidity of public sentiment in response to the pandemic. The findings highlight the critical role of social media platforms as barometers of public opinion and sentiment, offering real-time insights into societal reactions, concerns, and needs. To capitalize on these insights, several recommendations are proposed. Firstly, leveraging sentiment analysis findings to tailor public health messaging and crisis communication strategies to resonate with the prevailing sentiments and concerns of the populace. Secondly, enhancing efforts to counter misinformation and promote accurate, evidence-based information through targeted interventions and collaborative partnerships with social media platforms. Additionally, fostering proactive engagement with communities to address socio-economic disparities, mental health challenges, and other concerns exacerbated by the pandemic. Finally, advocating for continued monitoring of sentiment trends on Twitter and other social media platforms as part of a holistic approach to crisis monitoring and response. By embracing these recommendations, stakeholders can harness the power of sentiment analysis to navigate the complexities of the COVID-19 pandemic effectively, foster resilience, and build a more informed, empathetic response to future crises.

TABLE OF CONTENTS

List of Figures

List of Tables

Chapter 1

COMPANY PROFILE

Ideal Educraft Technologies Pvt Ltd

Founded in 2014, Educraftech is leading Research and Embedded Systems Company located in
BTM Layout, Bengaluru-560076. Research, Innovation, and Education are the three pillars of Educraftech. Design and Development of Embedded systems and IoT products, Research Solutions, and Global Standard Education are the major services of the company.

Educraftech was founded as a core Research Company. The Products and Services, later, were expanded to different domains. However, the research DNA of Educraftech remains solid. The Research Methodologies developed by Educraftech as a result of many years of experience are highly efficient for R&D and Industry projects. A unique combination of System Modelling and Design, Creative Learning, Psychometric Analysis, Agile/SCRUM, Instructional Design, and Theory of Innovation are used in the development of the Research Methods.

The uniqueness of any product depends on the creative brains that visualize the solution and materialize the thought. Educraftech is the brainchild of likeminded young professionals with Expertise and Experience in the fields of Electronics and Computer Science. We bring together state-of-the art technologies and Research Expertise to solve real-time problems.
Bridging the gap between Academia and Industry while imparting the Knowledge through creative methods is the philosophy of Educraftech.

Educraftech has successfully conducted Workshops and Internships that has influenced students to craft their career. They have conducted many Entrepreneurship Trainings, Research Session and internship sessions. It has collaborated with many institutions like RV college, DSCE, etc.
Educraftech has successfully contributed to various technical fields like-

Smart Plug- Educraftech has designed and developed an innovative model of smart plug with advanced networking and user interface.

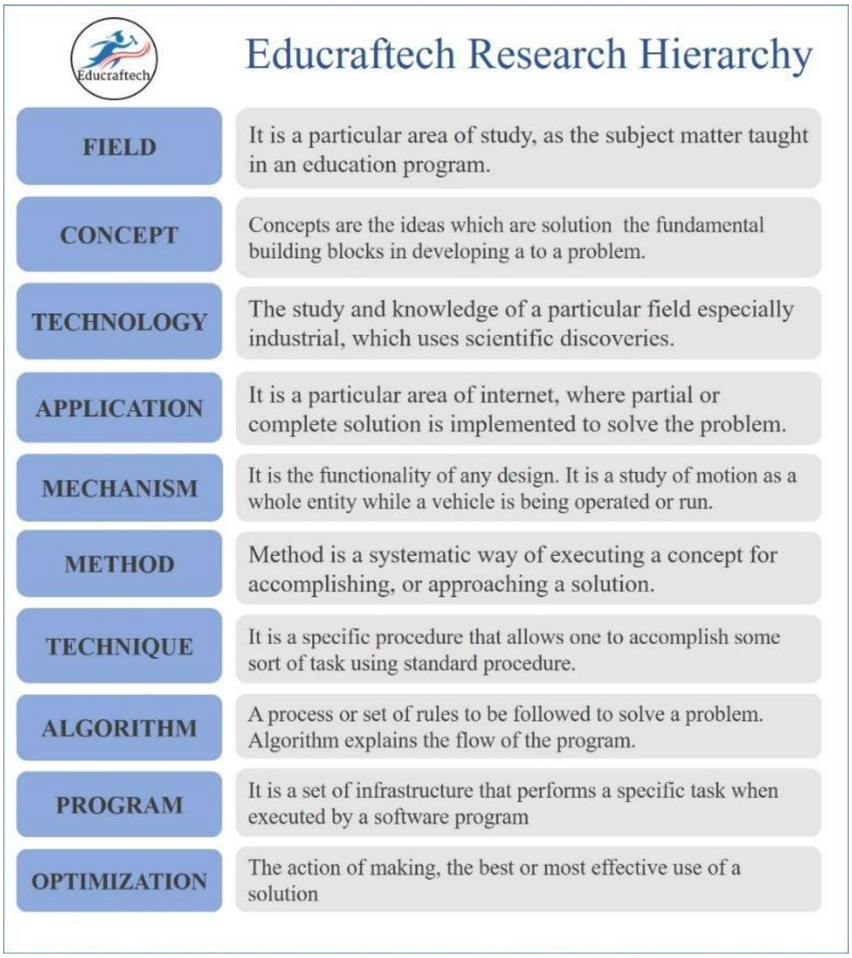
Home Automation- Educraftech offers customized Home Automation Solution such as Smart Doors, Smart Lighting, Smart Security System and IoT Products.

Research Kit- Educraftech offers customized IoT Research Kit which was developed by a team of young professionals and brilliant students. The research kit is first of its kind which can be used to conduct more than 50 Experiments.

R&D- Educraftech is actively engaged in developing solutions for Smart City Projects through Networking, Wireless Connectivity and IoT products.

Educraftech Milestones

2014- Founded as a Research Company
2015 - Embedded Systems product development
2016 - Collaborations with Companies and Universities
2017 - Recognition as a Startup by Govt. of India
2018 - Smart-Plug Design and Development
2019 - Smart City Project



Educraftech Research Hierarchy

FIELD	It is a particular area of study, as the subject matter taught in an education program.
CONCEPT	Concepts are the ideas which are solution the fundamental building blocks in developing a to a problem.
TECHNOLOGY	The study and knowledge of a particular field especially industrial, which uses scientific discoveries.
APPLICATION	It is a particular area of internet, where partial or complete solution is implemented to solve the problem.
MECHANISM	It is the functionality of any design. It is a study of motion as a whole entity while a vehicle is being operated or run.
METHOD	Method is a systematic way of executing a concept for accomplishing, or approaching a solution.
TECHNIQUE	It is a specific procedure that allows one to accomplish some sort of task using standard procedure.
ALGORITHM	A process or set of rules to be followed to solve a problem. Algorithm explains the flow of the program.
PROGRAM	It is a set of infrastructure that performs a specific task when executed by a software program
OPTIMIZATION	The action of making, the best or most effective use of a solution

Fig 1: Educraftech research hierarchy

Chapter 2

LEARNING DURING INTERNSHIP

2.1 TECHNICAL LEARNING

1. **Python Programming:**

   The internship provided an immersive learning experience in Python programming, which served as the foundation for all technical tasks and analyses. Python, being an open-source, versatile, and widely used programming language in data science, proved indispensable in various aspects of the project. Throughout the internship, I gained proficiency in Python syntax, data structures, and object-oriented programming concepts through practical application in data science projects. The utilization of Python's extensive ecosystem of libraries, such as NumPy, Pandas, Matplotlib, and Scikit-learn, facilitated data manipulation, analysis, visualization, and machine learning tasks. Additionally, I honed my skills in writing clean, efficient, and scalable code, following best practices and coding conventions. Python's readability, flexibility, and community support made it an ideal choice for implementing various data processing, analysis, and machine learning tasks in the project. Overall, the internship deepened my understanding of Python programming and its applications in data science and analytics, laying a solid foundation for future endeavors in the field.

2. **Data Collection and Preprocessing Techniques:**

   The internship provided hands-on experience in data collection from Twitter using the Twitter API. This involved setting up developer credentials, constructing API queries, and handling rate limits and pagination to efficiently retrieve large volumes of real-time data. The collected data comprised tweets containing relevant keywords, hashtags, and mentions related to COVID-19, such as "#COVID19," "coronavirus," and "pandemic." Additionally, I gained proficiency in preprocessing techniques to clean and prepare the collected data for sentiment analysis. This included tokenization using Python's NLTK library to split text into individual words, stop word removal to eliminate common, non-informative words, and sentiment lexicon-based analysis to assign sentiment scores to each tweet. These techniques were implemented using Python programming language and its associated libraries for data manipulation and analysis.

3. **Machine Learning Algorithms for Sentiment Analysis:**

   As part of the project, I explored various machine learning algorithms for sentiment analysis, including Naive Bayes and Long Short-Term Memory (LSTM) networks. I gained a deep understanding of the theoretical foundations of these algorithms and their applications in text classification tasks. Using Python's scikit-learn library, I trained and evaluated these algorithms on labeled datasets to categorize tweets into positive, negative, or neutral sentiments based on their content. Additionally, I experimented with hyperparameter tuning techniques to optimize the performance of the models. This hands-on experience allowed me to develop proficiency in implementing machine learning algorithms for sentiment analysis and interpret the results effectively.

4. **Natural Language Processing (NLP) Libraries and Tools:**
   Throughout the internship, I utilized a variety of NLP libraries and tools to preprocess and analyze text data. Python's NLTK (Natural Language Toolkit) library served as a versatile toolkit for text processing tasks such as tokenization, stemming, and part-of-speech tagging. I also leveraged Text Blob, a user-friendly NLP library built on the NLTK and Pattern libraries, for sentiment analysis and feature extraction. Moreover, I explored advanced NLP techniques such as word embeddings using libraries like spaCy and gensim, which enabled me to capture semantic relationships and context in text data more effectively. These tools and libraries were instrumental in implementing complex NLP tasks and extracting meaningful insights from the COVID-19 tweets collected from Twitter. Additionally, I used Python programming language and its associated libraries for data manipulation and analysis, including pandas for data processing, Matplotlib and Seaborn for data visualization, and scikit-learn for machine learning tasks.

5. **Web Scraping Techniques:**
   In addition to Twitter API, I also learned web scraping techniques to gather supplementary data from online sources. Using libraries such as BeautifulSoup and Scrapy in Python, I extracted information from various websites relevant to the project. Web scraping involved parsing HTML and XML documents, navigating through website structures, and extracting specific data elements. This experience broadened my skillset in data acquisition and provided alternative methods for gathering information beyond API access.

2.2 SOFT SKILLS LEARNING

1. **Communication Skills:**

   Effective communication played a crucial role throughout the internship, as I regularly communicated project progress, findings, and insights to team members and supervisors. I developed the ability to articulate complex technical concepts in a clear and concise manner, tailoring my communication style to suit the audience's level of technical expertise. This experience enhanced my presentation skills and equipped me with the confidence to effectively convey my ideas and findings to both technical and non-technical stakeholders.

2. **Problem-Solving Abilities:**

   The internship provided numerous opportunities to encounter and address challenges in data collection, preprocessing, and analysis. I developed strong problem-solving abilities by systematically identifying issues, exploring potential solutions, and implementing effective strategies to overcome obstacles. Collaboration with peers and mentors further enriched the problem-solving process, as we shared insights, brainstormed solutions, and offered support to one another. These experiences enhanced my critical thinking skills and prepared me to tackle complex problems in future projects.

3. **Time Management:**

   Managing multiple project tasks and deadlines required effective time management skills. I learned to prioritize tasks based on urgency and importance, allocate time and resources efficiently, and maintain focus and productivity amidst competing demands. Adopting strategies such as setting clear goals, breaking down tasks into manageable steps, and using productivity tools helped me stay organized and meet project deadlines consistently. This experience strengthened my ability to work effectively under pressure and deliver high-quality results within tight timelines.

4. **Adaptability and Flexibility:**

   The internship provided an environment conducive to learning and growth, where adaptability and flexibility were essential traits. As project requirements evolved and priorities shifted, I demonstrated flexibility in adjusting my approach and embracing change proactively. Adapting to new technologies, methodologies, and project dynamics enabled me to navigate uncertainty and ambiguity with confidence. This experience fostered resilience and agility, qualities that are valuable assets in today's fast-paced and dynamic work environment.

5. **Teamwork and Collaboration:**

   Collaboration was a cornerstone of the internship experience, as I worked closely with team members, supervisors, and external stakeholders to achieve project objectives. I actively contributed to team discussions, shared ideas and insights, and leveraged collective expertise to solve problems and drive innovation. By fostering a culture of collaboration and mutual support, we cultivated a positive team environment where everyone felt valued and empowered to contribute their unique perspectives and skills. This experience reinforced the importance of teamwork and collaboration in achieving shared goals and fostering a culture of continuous learning and improvement.

2.3 PROFESSIONAL DEVELOPMENT

1. **Professional Growth:**
   The internship served as a catalyst for professional growth and development, providing me with valuable skills, experiences, and insights that will shape my career trajectory in the field of data science and analytics. I gained a deeper understanding of data analysis techniques, machine learning algorithms, and natural language processing tools, laying a solid foundation for future projects and endeavors. Moreover, the internship allowed me to cultivate essential soft skills such as communication, problem-solving, and time management, which are integral to success in any professional role. Overall, the internship was a transformative experience that equipped me with the knowledge, skills, and confidence to pursue my career goals and make meaningful contributions to the field of data science and beyond.

2. **Networking Opportunities:**
   During the internship, I had the opportunity to connect with professionals in the field of data science and analytics. Through team meetings, networking events, and collaborative projects, I expanded my professional network and established valuable connections with industry experts, researchers, and fellow interns. These networking opportunities provided insights into industry trends, career pathways, and potential job opportunities in the field. Additionally, I received mentorship and guidance from experienced professionals, who shared their knowledge, expertise, and career advice, further enhancing my professional development.

3. **Continuous Learning and Skill Enhancement:**
   The internship encouraged a culture of continuous learning and skill enhancement, where I was encouraged to explore new technologies, methodologies, and best practices in data science and analytics. I actively participated in workshops, webinars, and online courses to broaden my skillset and stay updated on the latest developments in the field. Additionally, I pursued certifications and attended conferences related to data science and analytics, further deepening my knowledge and expertise. This commitment to continuous learning and skill enhancement has equipped me with the agility and adaptability to thrive in a rapidly evolving industry landscape.

4. **Career Planning and Goal Setting:**
   The internship provided a platform for me to reflect on my career aspirations, strengths, and areas for growth. Through discussions with mentors, supervisors, and career advisors, I developed a

clearer understanding of my career goals and formulated a plan to achieve them. I identified opportunities for further education, training, and professional development to enhance my skillset and position myself for success in the field of data science and analytics. Moreover, I received guidance on resume building, interview preparation, and job search strategies, laying the groundwork for a successful transition from intern to professional.

5. **Leadership and Mentorship:**
   As I progressed through the internship, I had the opportunity to take on leadership roles and mentorship responsibilities within the team. I led project initiatives, supervised junior team members, and provided guidance and support to peers facing technical challenges. These leadership and mentorship experiences honed my interpersonal skills, fostered teamwork and collaboration, and deepened my understanding of effective leadership practices. Moreover, they instilled a sense of responsibility and accountability, motivating me to make meaningful contributions to the team and organization.

6. **Personal Development:**
   Beyond professional growth, the internship also facilitated personal development and self-discovery. Through challenging projects, constructive feedback, and reflective practices, I gained insights into my strengths, weaknesses, and areas for personal growth. I developed resilience, adaptability, and confidence in navigating complex and dynamic work environments. Moreover, I cultivated a growth mindset, embracing failures and setbacks as opportunities for learning and improvement. These personal development experiences have shaped not only my professional journey but also my personal growth and fulfillment.

In conclusion, the internship provided a transformative experience that contributed to my overall professional development and prepared me for a successful career in the field of data science and analytics. Through technical learning, soft skills development, networking opportunities, continuous learning, career planning, leadership, mentorship, and personal growth, I have acquired the knowledge, skills, and mindset to excel in a dynamic and competitive industry landscape. I am grateful for the valuable experiences and opportunities afforded to me.

Chapter 3

WEEKLY INTERNSHIP ACTIVITIES

3.1 WEEK 1

Date	Day	Name of topic/Module covered
18-08-2023	**Friday**	**Group Activity on different systems of a car.**
19-08-2023	**Saturday**	**Project Topics were assigned.**
22-08-2023	**Tuesday**	**Session on Report writing**
24-08-2023	**Thursday**	**Communication Skill Session**

Table 1:Week1 Activities

On 18th August, an insightful group activity focused on developing communication skill, interpersonal skill, problem solving skills etc. was organized which helped us to understand the need of Team work.

On 19th August, a representative from Infosys divided us in a team of 4 according to the domain selected and assigned us the respective topic.

On 22nd August, a session on report writing was conducted which gave us the insights on how to write a proper report. It prepared us to document our research findings effectively.

On 24th August, a session on communication skill was conducted in which the essential aspects of resume writing & interview skills were covered.

3.2 WEEK 2

Date	Day	Name of topic/Module covered
25-08-2023	**Friday**	**Google meet on Project Progress update.**
26-08-2023	**Saturday**	**Submission of weekly report.**
28-08-2023	**Sunday**	**Communication Skill Session with Varsha Samtani**
28-08-2023	**Monday**	**Zoom meeting for project update and code**
29-08-2023	**Friday**	**Evaluation of weekly report.**

Table 2:Week2 Activities

On 25th August, an online meeting was organized to discuss the project progress and suggestions were given on how to follow up the project.

On 26th August, we were informed to submit the weekly report on Literature Survey and Introduction of the project topic assigned.

On 28th August, a session on communication skill was conducted with Varsha Samtani. She gave us the insight on how to write our resume, how to be prepared for a company interview etc. and in the afternoon a zoom meeting was held to discuss about the project progress.

3.3 WEEK 3

Date	**Day**	**Name of topic/Module covered**
31-08-2023	**Thursday**	**Industrial Talk with Ms. Anusha and Ms. Chanchal.**
02-09-2023	**Saturday**	**Submission of weekly report.**
04-09-2023	**Monday**	**Discussion on Project Progress.**
05-09-2023	**Tuesday**	**Industrial Talk from research scientist from NAL on ceramic products**

Table 3:Week3 Activities

On 31st August, an Industrial Talk with Ms. Anusha and Ms. Chanchal was organized which us the insights about the work culture of a company and how to enhance our technical skills. A group activity was also conducted to help us improve our problem-solving skills.

On 2nd September, weekly reports were submitted, showcasing our progress and findings, difficulties faced on the respective projects.

On 4th September, a discussion session was held to review and discuss the progress of ongoing projects, facilitating feedbacks from the Infosys Representative

.

On 5th September, a research scientist from the National Aerospace Laboratories (NAL) delivered an enlightening talk on ceramic products in the industrial context, providing valuable insights into this specialized field.

3.4 WEEK 4

Date	Day	Name of topic/Module covered
07-09-2023	Friday	Industrial Talk of Resource person from Automotive Industry
07-09-2023	Friday	Final Project Discussion
10-09-2023	Monday	Discussion on format of Final Report.

Table 4:Week4 Activities

On 7th September, an industry expert graced us with insights into the automotive sector, sharing valuable knowledge and experiences. This talk provided participants with a real-world perspective on the automotive industry's challenges. A critical discussion was conducted about the final projects, seeking feedback, sharing progress, and addressing any concerns.

On 10th September, discussion on the required format, structure, and content for final report format was done.

Chapter 4

PROJECT DESCRIPTION

4.1 PROBLEM STATEMENT

The project aims to address the challenge of comprehensively analyzing the sentiments expressed in COVID-19-related tweets on Twitter. With the global pandemic significantly impacting individuals and communities worldwide, understanding the public sentiment on social media platforms has become increasingly crucial. However, the sheer volume and complexity of COVID-19-related tweets present a formidable obstacle in extracting meaningful insights accurately.

The primary problem is to develop effective methodologies and algorithms capable of accurately categorizing tweets into positive, negative, or neutral sentiments despite the inherent challenges posed by natural language nuances. Additionally, the analysis must consider temporal fluctuations, geographical variations, and cultural contexts to ensure the relevance and reliability of the findings across diverse settings.

Furthermore, the project aims to identify prevalent themes, concerns, and priorities expressed in COVID-19 tweets. By uncovering underlying patterns and trends within the Twitter discourse, the analysis seeks to provide actionable insights that can inform decision-making processes and communication strategies for various stakeholders, including policymakers, healthcare professionals, businesses, and the general public.

In essence, the project revolves around conducting a robust sentiment analysis of COVID-19-related tweets on Twitter to gain a deeper understanding of public sentiment during this unprecedented global crisis and facilitate informed decision-making and communication strategies.

4.2 OBJECTIVES

1. **Collect a dataset of COVID-19-related tweets from Twitter:** The primary objective of this project is to gather a substantial dataset comprising tweets related to COVID-19, sourced from Twitter's API or other relevant data sources. This dataset will serve as the foundational bedrock for conducting comprehensive sentiment analysis and deriving meaningful insights. The collection process will involve defining and refining search queries to ensure the retrieval of relevant tweets containing keywords, hashtags, or mentions pertaining to COVID-19. Additionally, attention will be given to data integrity and quality assurance measures to ensure the reliability and accuracy of the collected dataset.

2. **Perform sentiment analysis to categorize tweets as positive, negative, or neutral:** Leveraging advanced natural language processing (NLP) techniques, the project aims to analyze the sentiment expressed in each tweet and categorize them into distinct categories of positive, negative, or neutral sentiments. This involves deploying sentiment analysis algorithms and methodologies to decode the underlying emotional tone and polarity embedded within the textual content of the tweets. By systematically categorizing tweets based on sentiment, the analysis seeks to unravel the prevailing emotional landscape surrounding COVID-19 discussions on Twitter, thereby offering valuable insights into public sentiment dynamics.

3. **Explore and visualize trends in COVID-19 sentiment over time and across regions:** Utilizing data visualization techniques, such as time series graphs, geographic heatmaps, and thematic maps, the project aims to elucidate how sentiments towards COVID-19 evolve temporally and vary spatially across different regions. Through the creation of visually compelling and informative visualizations, the analysis seeks to capture the nuanced nuances in sentiment trends, identify temporal patterns, and uncover geographical disparities in sentiment expression. This exploration will provide stakeholders with a comprehensive understanding of the temporal and spatial dynamics of public sentiment surrounding the pandemic, facilitating informed decision-making and targeted interventions.

4. **Identify common themes and concerns expressed in the tweets:** Conducting thematic analysis on the collected dataset, the project endeavors to identify recurring topics, issues, and concerns expressed in COVID-19-related tweets. By systematically analyzing the textual content of tweets, key themes and discourse patterns will be identified, shedding light on the prevalent concerns,

priorities, and focal points within the Twitter discourse on COVID-19. This thematic analysis will offer valuable insights into the key areas of focus and public sentiment surrounding the pandemic, thereby informing strategic communication efforts and policy responses.

5. **Assess the impact of major events or policy changes on sentiment:** Investigating the correlation between major events, policy changes, or significant milestones in the pandemic and fluctuations in public sentiment on Twitter, the project aims to assess the impact of external stimuli on sentiment dynamics. By systematically analyzing sentiment fluctuations in response to specific events or interventions, the analysis seeks to unravel the intricate interplay between external factors and public sentiment. This assessment will provide stakeholders with valuable insights into the drivers of sentiment change and enable them to adapt communication strategies and policy responses accordingly.

6. **Develop a sentiment lexicon tailored to COVID-19 tweets:** Given the unique nature of sentiment expressed in COVID-19-related tweets, the project aims to develop a specialized sentiment lexicon specifically tailored to capture the nuances of emotions and attitudes surrounding the pandemic. This involves compiling a comprehensive list of sentiment-bearing words and phrases relevant to COVID-19 discourse and assigning sentiment scores based on the intensity and valence of emotions expressed. The development of a custom sentiment lexicon will enhance the accuracy and relevance of sentiment analysis results, enabling more nuanced insights into public sentiment.

7. **Integrate machine learning algorithms for sentiment classification:** In addition to traditional lexicon-based approaches, the project seeks to leverage machine learning algorithms for sentiment classification of COVID-19 tweets. By training supervised learning models on labeled datasets of COVID-19 tweets, the analysis aims to enhance the accuracy and robustness of sentiment classification, thereby enabling more precise identification of positive, negative, and neutral sentiments. This integration of machine learning algorithms will empower the analysis to handle the complexities and nuances of sentiment expression in COVID-19-related tweets more effectively.

8. **Conduct comparative analysis across social media platforms:** Beyond Twitter, the project aims to extend its analysis to other social media platforms, such as Facebook, Instagram, and Reddit, to gain a comprehensive understanding of sentiment dynamics across different digital channels. By conducting comparative analysis across multiple platforms, the project seeks to identify similarities and differences in sentiment expression, audience demographics, and thematic trends, offering

valuable insights into the broader landscape of public sentiment surrounding COVID-19.

9. **Evaluate the efficacy of sentiment analysis methodologies:** Throughout the project, the efficacy and performance of various sentiment analysis methodologies, including lexicon-based approaches and machine learning algorithms, will be rigorously evaluated and compared. By assessing the accuracy, precision, recall, and computational efficiency of different methodologies, the project aims to identify best practices and optimize the sentiment analysis workflow for maximum effectiveness and reliability.

10. **Provide actionable insights for stakeholders:** Ultimately, the project aims to translate the findings of sentiment analysis into actionable insights and recommendations for stakeholders across various sectors. By distilling complex sentiment data into clear, actionable recommendations, the analysis seeks to empower decision-makers to formulate targeted communication strategies, policy interventions, and public health initiatives that are responsive to the prevailing sentiment dynamics surrounding COVID-19.

4.3 METHODOLOGY

BLOCK DIAGRAM:

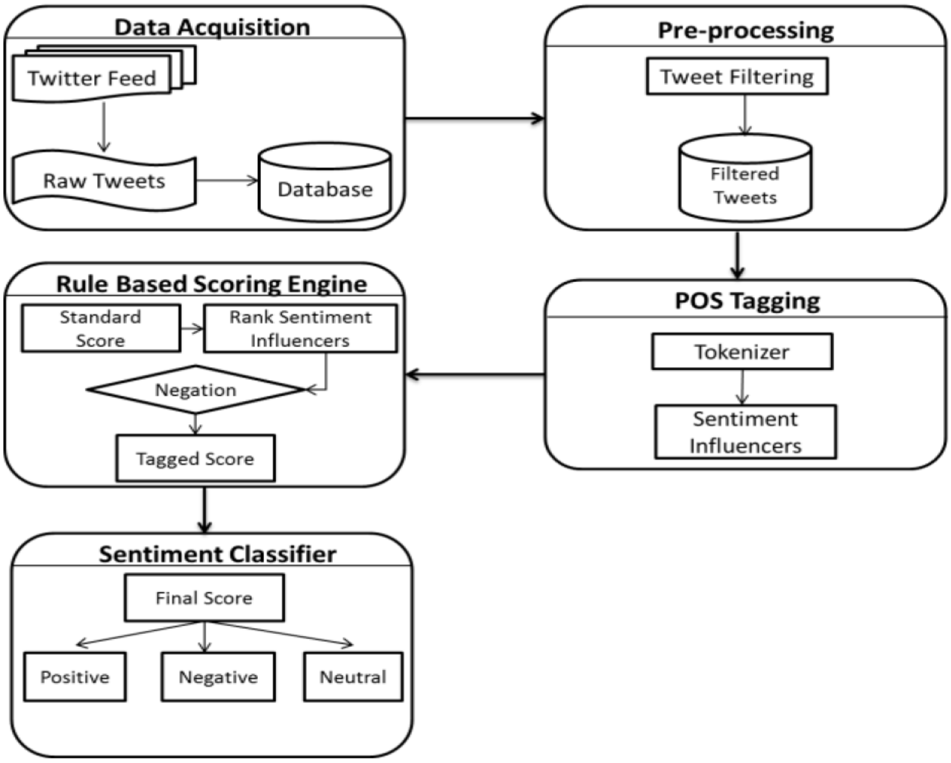


Figure 2 - Abstract level Diagram of Proposed Framework

FLOWCHART:

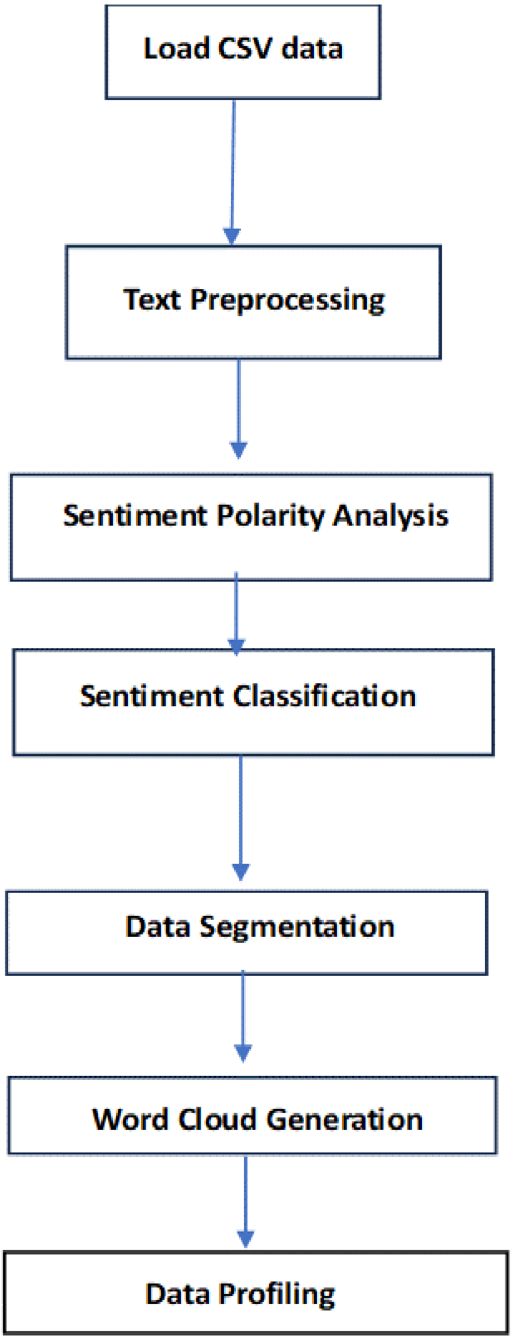


Figure 3 - Flowchart for sentiment analysis of the tweets

PROPOSED METHODOLOGY:

DATA PREPARATION:

1. **Collecting the Dataset:**

- In the initial phase of the sentiment analysis project, the foremost task is to gather a comprehensive dataset containing text data suitable for analysis. This dataset should ideally encompass a diverse range of text samples relevant to the subject of interest, which in this case is COVID-19-related discussions on Twitter.
- To ensure the dataset's relevance and representativeness, it is crucial to select a variety of sources, including reputable repositories, public datasets, and possibly real-time data streams from Twitter's API. This diverse sourcing strategy helps capture different perspectives, opinions, and sentiments expressed across various demographics, geographic regions, and timeframes.
- Careful attention should be paid to the data collection process to maintain data integrity and quality. Data sources should be verified for credibility and accuracy, and steps should be taken to address any potential biases or limitations inherent in the dataset.

SENTIMENTAL ANALYSIS:

1. **Preprocessing the Text Data:**

- Once the dataset has been collected, the next step is to preprocess the text data to prepare it for sentiment analysis. This preprocessing stage involves a series of steps aimed at cleaning, standardizing, and structuring the text data to ensure consistency and accuracy in the subsequent analysis.
- Common preprocessing techniques include removing punctuation, special characters, and irrelevant information; converting text to lowercase to standardize text casing; tokenizing the text into individual words or tokens for analysis; and handling issues such as encoding errors and non-standard formatting.
- Additionally, it may be necessary to address specific challenges related to text data, such as handling multilingual texts, non-English characters, or domain-specific jargon. Specialized preprocessing techniques may be required to address these challenges effectively and ensure the integrity of the text data.

2. **Utilizing Sentiment Analysis Tools:**

- After preprocessing the text data, sentiment analysis tools can be employed to analyze the sentiment expressed in each text sample. One commonly used tool for sentiment analysis is Text Blob, a Python library that provides a simple interface for performing sentiment analysis.

- on textual data.
- Text Blob calculates sentiment polarity scores for each text, indicating the degree of positive or negative sentiment expressed in the text. The polarity scores range from -1 (indicating strong negative sentiment) to +1 (indicating strong positive sentiment), with scores close to 0 indicating neutral sentiment.
- By leveraging sentiment analysis tools like Text Blob, researchers can efficiently analyze large volumes of text data and derive valuable insights into the prevailing sentiment trends and patterns.

3. **Categorizing Texts:**

- Once sentiment polarity scores have been calculated for each text sample, the next step is to categorize the texts into distinct sentiment categories, such as positive, negative, or neutral. This categorization enables a more granular analysis of sentiment dynamics and facilitates the identification of sentiment trends and patterns.
- Categorization can be based on predefined polarity thresholds, where texts with polarity scores above a certain threshold are classified as positive, those below another threshold are classified as negative, and those within a neutral range are classified as neutral. These thresholds can be adjusted based on the specific requirements and objectives of the analysis.

DATA SEGMENTATION:

1. **Separating the Dataset:**

- To facilitate further analysis and visualization, the dataset can be segmented into two subsets: one containing text classified as positive sentiment and another containing texts classified as negative sentiment. This segmentation allows for a more focused examination of sentiment dynamics within each sentiment category.
- Separating the dataset into distinct sentiment subsets enables researchers to identify common themes, trends, and keywords associated with positive and negative sentiment, respectively. This segmentation also facilitates the generation of targeted visualizations and insights tailored to each sentiment category.

WORD CLOUD GENERATION:

1. **Combining Text Data:**

- Before generating word clouds for positive and negative sentiment subsets, the text data from each sentiment category needs to be combined into separate text strings. This aggregation consolidates the textual content associated with each sentiment category, providing a comprehensive overview of the predominant themes and keywords.

- The combined text strings serve as input for generating word clouds, which visually represent the frequency of words in the text by displaying them in varying sizes based on their occurrence frequency. Word clouds offer an intuitive and visually appealing way to identify prominent themes, keywords, and patterns within the text data.

2. **Generating Word Clouds:**

- Using a word cloud generation library such as Word Cloud in Python, researchers can create word clouds for both positive and negative sentiment categories. These word clouds provide visual representations of the most frequently occurring words in each sentiment subset, offering insights into the predominant themes and topics associated with positive and negative sentiment.
- By examining the word clouds, researchers can identify key words and phrases that are characteristic of each sentiment category, enabling a deeper understanding of the underlying sentiments and concerns expressed by Twitter users in relation to COVID-19.

DATA PROFILING:

1. **Generating Profile Reports:**

- To gain a comprehensive understanding of the characteristics and properties of the sentiment datasets, detailed profile reports can be generated using data profiling tools. These profile reports provide valuable insights into various aspects of the datasets, including data distributions, summary statistics, missing values, and data quality assessments.
- By analyzing the profile reports of the positive and negative sentiment datasets, researchers can identify any patterns, anomalies, or inconsistencies that may impact the quality and reliability of the data. This information can inform further preprocessing steps and data cleaning efforts to ensure the integrity and accuracy of the sentiment analysis results.

TOOLS:

The several tools and libraries for various tasks, primarily related to data analysis, visualization, and natural language processing (NLP) used are:

Jupyter Notebooks is used to implement the Twitter Data Sentimental analysis.

1. **Python:** Python is the programming language used for writing the code.
2. **pandas:** The `pandas` library is used for data manipulation and analysis. It is used to read and process the CSV data and create DataFrames for further analysis.
3. **TextBlob:** The `TextBlob` library is used for sentiment analysis. It provides simple methods for analysing text data and determining sentiment polarity.
4. **pandas-profiling:** The `pandas-profiling` library is used for generating profile reports of data. It helps in exploring and summarizing the dataset, including statistical analysis and data visualization.
5. **word cloud:** The `wordcloud` library is used to create word clouds. Word clouds are visual representations of word frequency in a text, often used for textual data analysis and visualization.
6. **matplotlib:** The `matplotlib` library is used for creating various types of plots and visualizations, including bar charts and word clouds.
7. **nltk:** The `nltk` library (Natural Language Toolkit) is used for natural language processing tasks. In this code, it’s used for downloading NLTK data, which may be used for NLP tasks, although the code snippet provided does not include explicit NLTK functions.
8. **IPython. display:** The `IPython. display` library is used to display HTML content and images in Jupyter Notebooks. It's used to display the profile reports and word cloud images.

The code provided demonstrates a workflow for analyzing sentiment in a dataset of COVID-19-related tweets, generating profile reports for positive and negative sentiments, and creating visualizations to represent the data. Additionally, it includes a separate section for analysing tweet counts by user and visualizing the top 10 users by tweet count.

These tools and libraries are commonly used in data analysis and visualization tasks in Python, making it easier to work with data, perform analyses, and create meaningful visualizations.

DATASET EXPLANATION:

The dataset used in the provided code is a collection of tweets related to COVID-19. Each row in the dataset represents a single tweet, and the columns contain various attributes and information about the tweets and the Twitter users who posted them.

COVID-19 Tweet Dataset :

Columns in the Dataset:

1. **`user_name`:** This column contains the name or username of the Twitter user who posted the tweet. Usernames are unique identifiers for Twitter users.
2. **`user_location`:** It represents the location mentioned in the Twitter user's profile. Users may provide information about where they are from or where they are currently located. This information can vary widely in its format.
3. **`user_description`:** This column contains the description or bio provided in the Twitter user's profile. Users often use this space to provide additional information about themselves, their interests, or their affiliations.
4. **`user_followers`:** The number of followers the Twitter user has. Followers are other Twitter users who have subscribed to receive updates from this user. A higher follower count may indicate a more influential user.
5. **`user_friends`:** The number of friends or accounts followed by the Twitter user. These are the accounts that the user follows on Twitter. It represents the user's network.
6. **`user_favourites`:** This column represents the number of favorites (likes) that the Twitter user has given. It indicates how active the user is in interacting with tweets by liking them.
7. **`user_verified`:** A boolean column that indicates whether the Twitter user's account is verified. Verified accounts typically belong to well-known individuals, organizations, or public figures. Verification adds credibility to the user.
8. **`text`:** The text content of the tweet itself. This column contains the actual message or statement posted by the user. It can include information, opinions, hashtags, mentions, or links.
9. **`hashtags`:** If the tweet contains any hashtags (keywords or topics preceded by the '#' symbol), they are listed in this column. Hashtags help categorize tweets and make them discoverable.

DATA ANALYSIS AND VISUALIZATION:

- Sentiment Analysis: The code performs sentiment analysis on the tweet texts from "text" column using the `TextBlob` library. Sentiment analysis determines whether the sentiment of each tweet is positive, negative, or neutral based on the text's content. The sentiment is quantified using polarity scores, where positive values indicate positive sentiment, negative values indicate negative sentiment, and values near zero indicate neutrality.

- Data Profiling: Profile reports are generated for both positive and negative sentiments using the `pandas-profiling` library. These reports provide detailed statistics and visualizations about the data, helping to understand its characteristics. The reports include information such as data types, missing values, and summary statistics for each column.

- Word Clouds: Word clouds are created to visualize the most common words in tweets with positive and negative sentiments. Word size in the cloud represents word frequency. This visualization technique highlights frequently occurring terms in the analyzed text data.

- Sentiment Distribution: A pie chart is created to visualize the distribution of sentiment categories (positive, negative, and presumably neutral) among the analyzed tweets. The chart shows the proportion of each sentiment category within the dataset.

PURPOSE AND INSIGHTS:

The goal of this code is to gain insights into the sentiment expressed in COVID-19-related tweets and to provide a visual representation of the data. It helps in understanding how Twitter users are discussing COVID-19 and whether the sentiment is predominantly positive or negative. Additionally, it provides insights into the users posting these tweets and their tweeting behaviors

ALOGORITHM:

1. Import necessary libraries and download NLTK data.
2. Load the COVID-19-related tweets dataset.
3. Define the DataFrame column containing tweet text.
4. Initialize variables for sentiment analysis.
5. Loop through each row in DataFrame 'df' to perform sentiment analysis using TextBlob.
6. Create two DataFrames to store rows with positive and negative sentiments.
7. Generate profile reports for both sentiment DataFrames using pandas-profiling.
8. Combine tweet text with positive and negative sentiments and create word clouds.
9. Save word cloud images.
10. Calculate 'neutral_count' based on 'positive_count' and 'negative_count' and create a pie chart representing sentiment distribution.
11. Create a bar chart to visualize the top 10 Twitter users with the most tweets.
12. Display results in a Jupyter Notebook environment using IPython.display.

The goal of this project is to provide insights into the sentiment expressed in tweets related to COVID-19 and visualize the data to better understand how Twitter users are discussing the pandemic.

4.4 HARDWARE AND SOFTWARE REQUIREMENTS

The successful execution of the sentiment analysis project relies on the availability and compatibility of various software components. These components are essential for data processing, analysis, and visualization tasks, ensuring a seamless workflow throughout the project lifecycle. The following software requirements are necessary for conducting the sentiment analysis project effectively:

1. **Operating System:**

   The sentiment analysis project can be conducted on multiple operating systems, including Windows, macOS, and various Linux distributions such as Ubuntu and CentOS. The choice of operating system is based on personal preference, compatibility with required software, and familiarity with the environment.

2. **Python**:

   Python serves as the primary programming language for implementing the sentiment analysis project. It provides a versatile and extensive ecosystem of libraries and tools for data manipulation, natural language processing (NLP), and visualization. It is imperative to ensure Python is installed on the system, preferably the latest version (Python 3.x), to leverage its capabilities fully.

3. **Integrated Development Environment (IDE):**

   An integrated development environment (IDE) facilitates coding, execution, and debugging of Python scripts, enhancing productivity during development. Popular IDE options for Python development include Jupyter Notebook, PyCharm, and Visual Studio Code. These IDEs offer features such as code highlighting, auto-completion, and interactive code execution, contributing to a smooth development experience.

4. **Libraries**:

   Python libraries play a crucial role in data analysis, NLP, and visualization tasks within the sentiment analysis project. It is essential to install the necessary libraries using package managers such as pip or conda to manage dependencies efficiently. Key libraries include:

- **Pandas**: For data manipulation and analysis.
- **NumPy**: For numerical computing and array operations.
- **NLTK** (Natural Language Toolkit): For NLP tasks such as tokenization, stemming, and sentiment analysis.
- **TextBlob**: A simplified NLP library for sentiment analysis and text processing.
- **Matplotlib**: For creating static, interactive, and animated visualizations.
- **Seaborn**: For statistical data visualization based on Matplotlib.
- **WordCloud**: For generating word clouds from text daTexto

5. **Text Editor:**

   A text editor provides a lightweight environment for editing configuration files, code snippets, or documentation related to the sentiment analysis project. Popular text editor options include Sublime Text, Atom, and Notepad++. These editors offer features tailored to coding tasks, enhancing the editing and management of code files.

6. **Web Browser:**

   Access to a web browser is necessary for downloading datasets, accessing online documentation, or referencing resources related to Python libraries and tools. Modern web browsers such as Google Chrome, Mozilla Firefox, or Microsoft Edge ensure optimal compatibility and performance for accessing online resources.

In conclusion, the availability and compatibility of these software components are essential for establishing a robust development environment and conducting effective sentiment analysis projects using Python. These components empower users to manipulate, analyze, and visualize data efficiently, facilitating informed decision-making and insights generation.

4.5 RESULTS AND ANALYSIS:

The code reads a pre-defined CSV dataset of COVID-19 related tweets, performs sentiment analysis on the text data, generates data profiling reports for positive and negative sentiments, and creates word clouds to visualize the most frequent words in tweets with different sentiment polarities. It's a comprehensive analysis of text data to understand the sentiment and content of the tweets.

Word Clouds: The code creates word clouds to visualize the most frequent words in tweets with positive and negative sentiments. It first concatenates all the positive and negative sentiment text data into two separate strings (positive_text and negative_text).

Word Cloud for Positive Sentiments: It generates a word cloud for positive sentiments using the WordCloud library, customizing its appearance, and displaying it using Matplotlib.

Word Cloud for Negative Sentiments: Similarly, it generates a word cloud for negative sentiments.

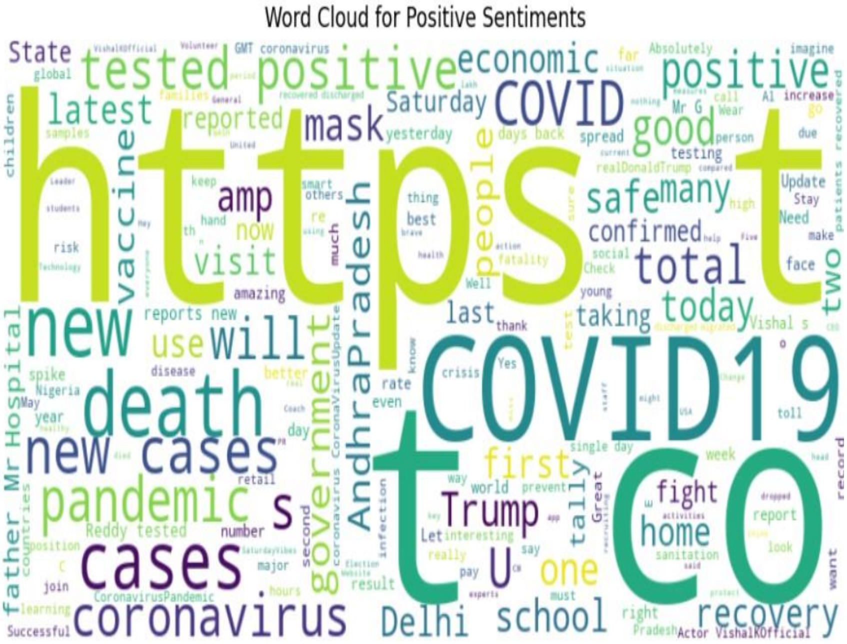


Fig 4:World cloud for positive sentiments

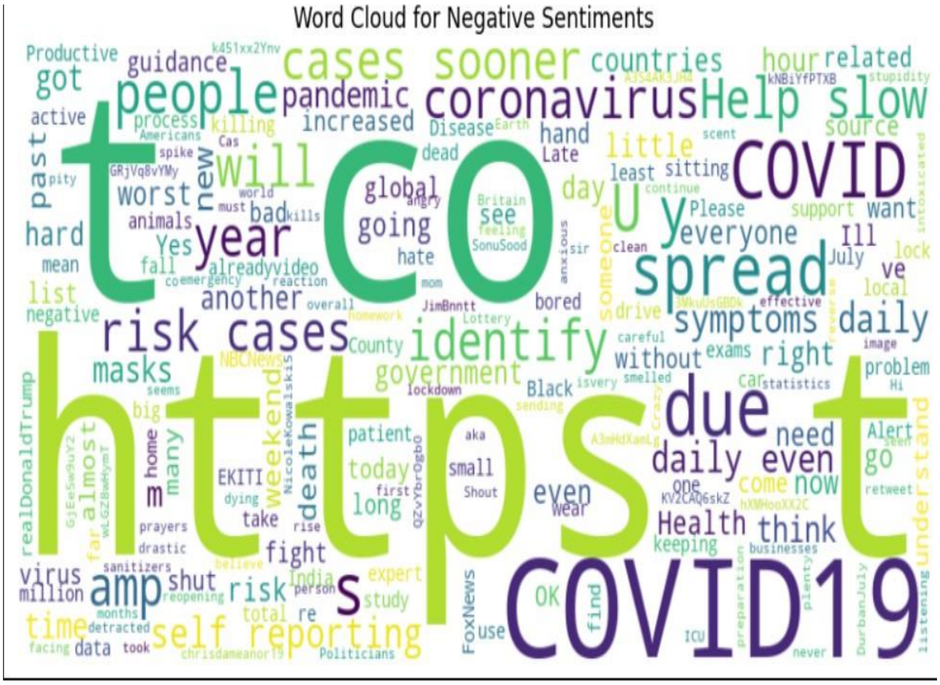


Fig 5:World cloud for negative sentiments

Next, we created a pie chart that visualizes the distribution of sentiment categories (positive, negative, and neutral) within a dataset of 500 tweets related to COVID-19.

The code calculates and visualizes the distribution of sentiment categories (positive, negative, and neutral) within a dataset of 500 COVID-19-related tweets. The pie chart provides a clear visual representation of the proportions of each sentiment category in the dataset.

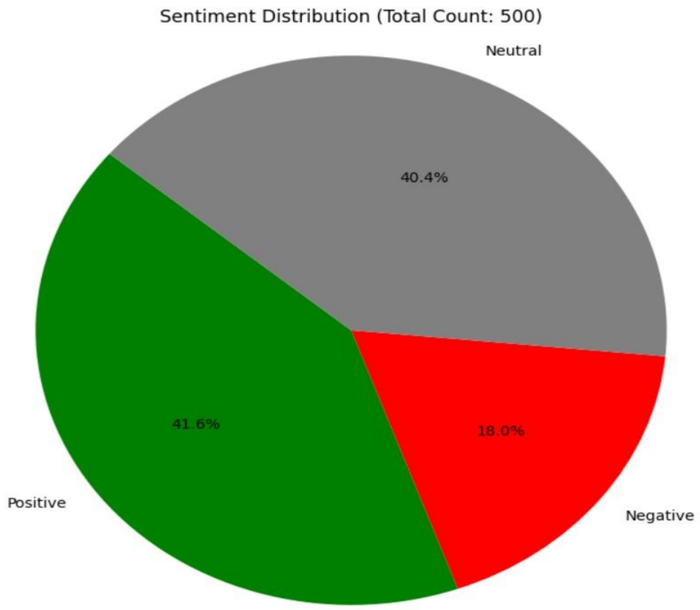
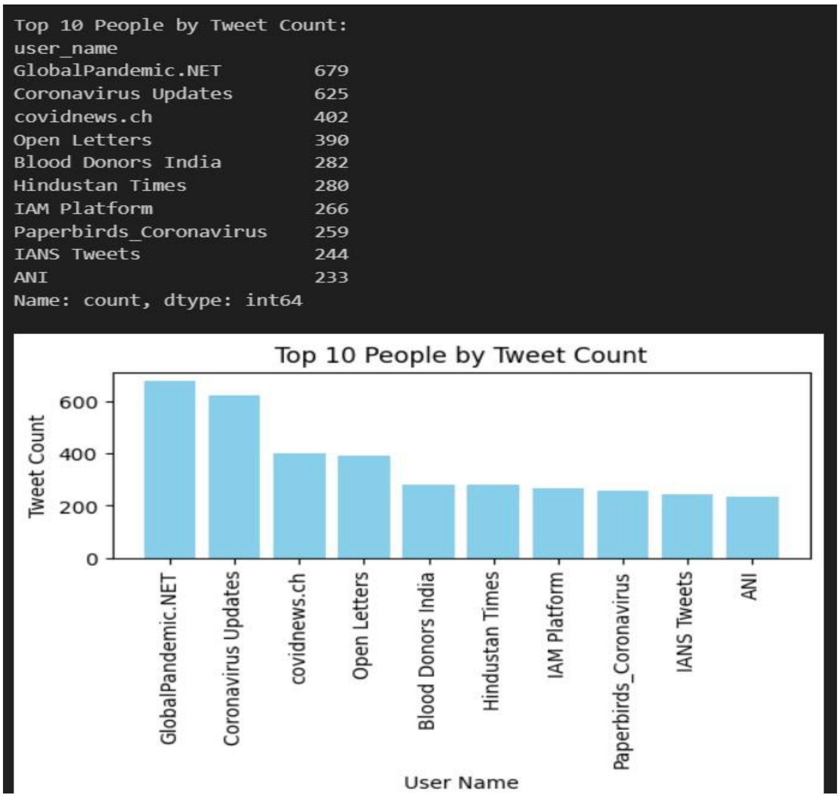


Fig 6 : Pie Chart visualizing the positive , negative and neutral tweets

After that, we load a CSV dataset containing information about COVID-19-related tweets, processes the data to determine the top 10 users with the most tweets, and then creates a vertical bar chart to visualize the tweet counts for these top users.

In summary, this code analyzes a dataset of COVID-19-related tweets, identifies the top 10 users with the most tweets, and visualizes their tweet counts using a bar chart for easy interpretation and comparison.

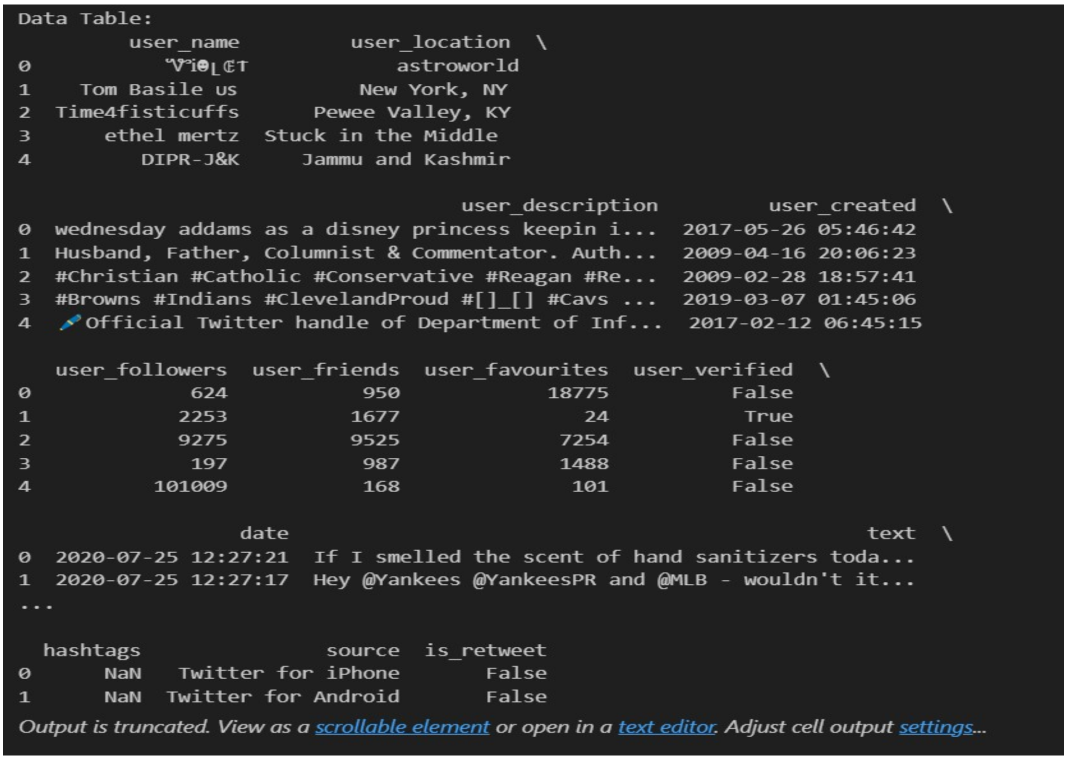


```
Top 10 People by Tweet Count:
user_name
GlobalPandemic.NET        679
Coronavirus Updates       625
covidnews.ch              402
Open Letters              390
Blood Donors India        282
Hindustan Times           280
IAM Platform              266
Paperbirds_Coronavirus    259
IANS Tweets               244
ANI                       233
Name: count, dtype: int64
```

Fig 7: Top 10 handles with a greater number of tweets

Next the code loads and processes data from two CSV files: one containing COVID-19 statistics (such as confirmed cases, deaths, and recoveries) and another containing tweets related to COVID-19. It also renames columns and converts date columns to datetime format.

The code loads and preprocesses data from two CSV files: one containing COVID-19 statistics and another containing tweets related to COVID-19. It ensures proper data types, column renaming, and datetime conversion for relevant columns. Finally, it prints a preview of the loaded data to verify its correctness and structure.



```
Data Table:
         user_name         user_location  \
0          ᏉᎥ☻ʟℭт            astroworld
1    Tom Basile us         New York, NY
2  Time4fisticuffs      Pewee Valley, KY
3      ethel mertz  Stuck in the Middle
4        DIPR-J&K     Jammu and Kashmir

                                 user_description         user_created  \
0  wednesday addams as a disney princess keepin i...  2017-05-26 05:46:42
1  Husband, Father, Columnist & Commentator. Auth...  2009-04-16 20:06:23
2  #Christian #Catholic #Conservative #Reagan #Re...  2009-02-28 18:57:41
3  #Browns #Indians #ClevelandProud #[]_[] #Cavs ...  2019-03-07 01:45:06
4  🖊Official Twitter handle of Department of Inf...  2017-02-12 06:45:15

   user_followers  user_friends  user_favourites  user_verified  \
0             624           950            18775          False
1            2253          1677               24           True
2            9275          9525             7254          False
3             197           987             1488          False
4          101009           168              101          False

                  date                                              text  \
0  2020-07-25 12:27:21  If I smelled the scent of hand sanitizers toda...
1  2020-07-25 12:27:17  Hey @Yankees @YankeesPR and @MLB - wouldn't it...
...

  hashtags               source  is_retweet
0      NaN   Twitter for iPhone       False
1      NaN  Twitter for Android       False
Output is truncated. View as a scrollable element or open in a text editor. Adjust cell output settings...
```

Fig 8:Processed Data

Next the code processes the dataset containing information related to COVID-19 tweets. It extracts and visualizes the top 15 user locations mentioned in the tweets.

The code loads tweet data from a CSV file, identifies the most frequently mentioned user locations in the tweets, and creates a horizontal bar chart to visualize the top 15 user locations and their frequencies. This visualization provides insights into where Twitter users are tweeting from in relation to COVID-19.

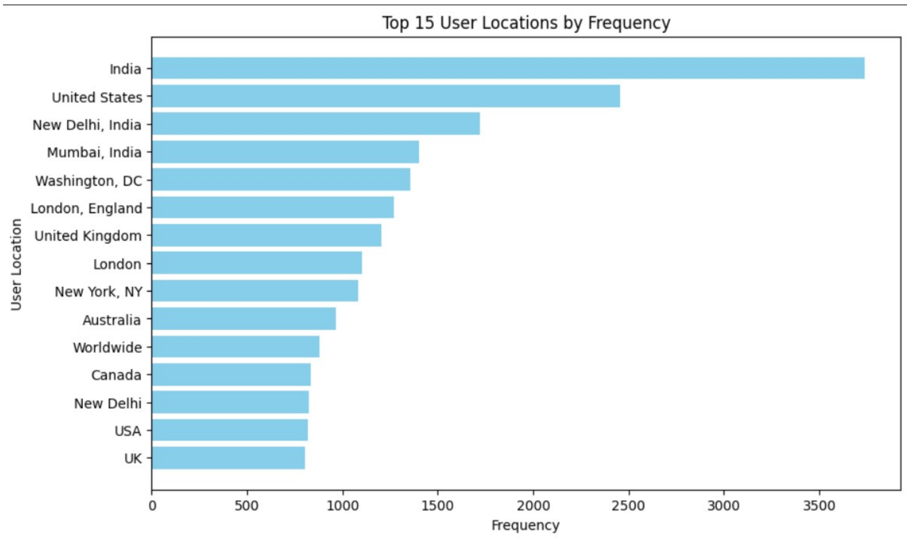


Fig 9: Top 10 Twitter Users

Next the code is designed to create a line plot that visualizes the number of tweets over time.

The code prepares and visualizes time-based data by creating a line plot that shows the number of tweets over time. It includes customizations for plot appearance, such as titles and labels, to make the visualization informative and easy to understand.

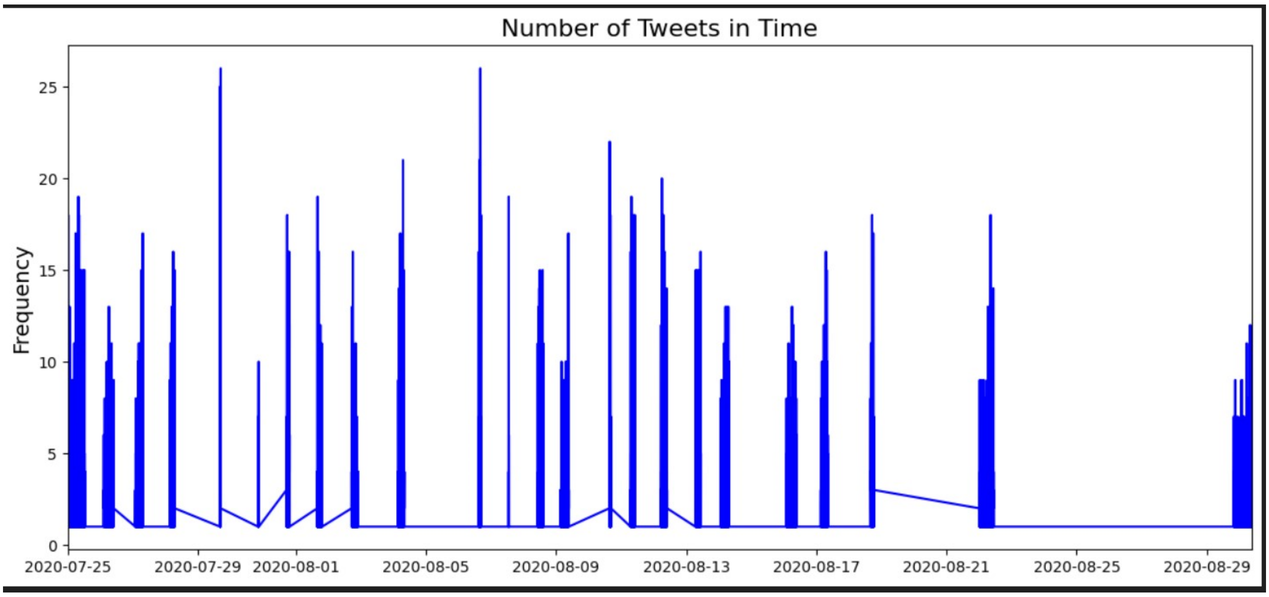


Fig 10 : Timeline of Tweets

Chapter 5

REFLECTION ON THE INTERNSHIP

My internship experience served as a transformative journey, immersing me in the dynamic and multifaceted world of data analysis, sentiment analysis, and natural language processing (NLP). Throughout this enriching experience, I encountered a multitude of challenges, seized numerous opportunities for growth, and emerged with a deeper understanding of the intricacies involved in deriving insights from textual data. Reflecting on the internship content, I am struck by the breadth and depth of knowledge I have acquired, as well as the invaluable skills I have honed:

1. Python Programming Proficiency:

The internship provided an immersive and hands-on learning experience in Python programming, which proved to be indispensable in the field of data science. From mastering Python's syntax and data structures to understanding advanced concepts such as object-oriented programming, I traversed a comprehensive learning path that equipped me with a solid foundation in software development. Through a series of challenging projects and exercises, I cultivated the ability to write clean, efficient, and maintainable code, adhering to industry best practices and coding standards. Moreover, I gained insights into the importance of modularization, documentation, and version control, laying the groundwork for effective collaboration and code maintenance in real-world projects.

2. Data Collection and Preprocessing Techniques:

A significant portion of the internship was devoted to mastering data collection and preprocessing techniques, particularly from social media platforms like Twitter. I embarked on a journey of discovery, learning to navigate the intricacies of Twitter's API, construct complex queries, and efficiently retrieve real-time data pertaining to the COVID-19 pandemic. Furthermore, I grappled with the complexities of preprocessing textual data, tackling challenges such as noise removal, text normalization, and the application of sentiment lexicons. Through iterative experimentation and refinement, I developed a keen eye for detail and a systematic approach to data wrangling, ensuring that raw data is transformed into a clean and structured format suitable for analysis.

3. Natural Language Processing (NLP) Tools and Libraries:

The internship provided a comprehensive exploration of NLP tools and libraries, empowering me to unlock the hidden insights buried within textual data. From NLTK and TextBlob to spaCy and genism, I delved into a diverse array of tools and techniques for text processing and analysis. Through hands-on experimentation and practical applications, I gained proficiency in tasks such as tokenization, part-of-speech tagging, sentiment analysis, and word embeddings. This experiential learning approach enabled me to unravel the intricate nuances of language and extract valuable insights from unstructured text data. By harnessing the power of NLP, I acquired the ability to decipher complex patterns, discern underlying trends, and derive actionable intelligence from textual sources.

4. Soft Skills Development:

In addition to technical proficiency, the internship fostered the development of essential soft skills crucial for professional success. Regular interactions with team members and mentors provided a fertile ground for honing communication skills, enabling me to articulate complex ideas with clarity and precision. Moreover, I embraced challenges as opportunities for growth, cultivating resilience, adaptability, and problem-solving acumen in the face of adversity. By fostering a collaborative and supportive environment, the internship nurtured a spirit of teamwork and camaraderie, enhancing my ability to collaborate effectively with peers and stakeholders. Furthermore, I gained insights into the importance of time management, organization, and prioritization, enabling me to balance multiple tasks and deadlines efficiently.

5. Adaptability and Professional Growth:

Perhaps the most profound lesson gleaned from the internship was the importance of adaptability and continuous learning in the ever-evolving field of data science. As project requirements evolved and new challenges emerged, I embraced change proactively, eagerly exploring new technologies, methodologies, and tools to overcome obstacles and achieve project objectives. This growth mindset fueled my professional development, empowering me to navigate the complexities of the data science landscape with confidence and agility. By staying abreast of emerging trends and innovations, I cultivated a forward-thinking approach that will serve as a cornerstone of my career advancement.

In summary, the internship content provided a rich tapestry of learning experiences and opportunities for growth, shaping me into a more proficient, adaptable, and resilient data scientist. The practical exposure to real-world data analysis tasks and NLP techniques has not only expanded my technical repertoire but also ignited a passion for leveraging data-driven insights to solve complex problems and drive positive change. As I reflect on the invaluable lessons learned and experiences gained during the internship.

Chapter 6

CONCLUSION

This exploratory analysis uses four ways of studying Tweet data: using correlation to determine relationship, identifying most used hashtags and tagged users, analyzing text to identify frequency of words, and using outlier daily

Tweet frequencies to determine possible cause. In this process, 2 datasets were created for each city: one for tweet sentiment, cases and deaths, and one for tweet text. The sentiment analysis was done on the negative compound sentiments of the tweets, averaged per day and textual examination was done for all the text in the tweet. The data should be explored more in order to obtain more information about user sentiment during Covid-19. Since the original dataset is extraordinarily large, analyzing the sentiment of all the tweets, localized to city, state, country, or even worldwide, would result in a more reliable conclusion regarding the relationship between sentiment and Covid-19 casualties. Similarly, classifying the tweets by concerns, such as healthcare, politics, economy, etc. can clarify the connection between sentiment and type of concern. Additionally, the retweet data can produce more reliable trackers for sentiment as they contain the favorite count and retweet count.

This research work aimed at analyzing the sentiments and emotions of the people during the pandemic COVID19. During the study, it was revealed that countries like Belgium, India and Australia were tweeting about COVID19 with a positive sentiment, people in China had negative sentiments about the same. Similarly, while analyzing the word clouds of different countries, it was concluded that people are tweeting words like Pandemic, Death, Quarantine, Hope, Stay Safe, Government, Political, Fight and Masks with different emotions. The name of the USA President, Donald Trump was amongst one of the most tweeted words not only in USA, but across all the twelve countries considered for the study. The tweets which were collected for this study were in English language which might serve as a limitation for the study. Also, while NRC Lexicon used for this study analyzed the tweets for eight different emotions, it does not count the emotions of sarcasm and irony. For the future works, this study can be used to analyze the changing emotions and sentiments of people from these countries and check whether there are major shifts in them over the period of time. It is expected that as the spread of this pandemic will increase, the sentiments and emotions in the tweets may change on the lines of what was seen in the case of China.

REFERENCES

[1] D. Antonakaki, P. Fragopoulou, and S. Ioannidis, "A Survey of Twitter Research: Data Model, Graph Structure, Sentiment Analysis, and Attacks," 15 September 2020.

[2] N. Srivastava, N. K. Singh, P. Singh, and V. Gupta, "Sentiment Analysis using Twitter Data."

[3] I. Kaur, M. N. Qureshi, A. Iqubal, and H. Adeeb, "Twitter Data Sentiment Analysis," in International Journal for Research in Applied Science & Engineering Technology (IJRASET), vol. 11, no. 4, 2023.

[4] K. Pabreja, "GST Sentiment Analysis Using Twitter Data."

[5] A. P. Singh, A. Singh, and G. J. Pandey, "Sentiment Analysis Using Twitter Data," in International Research Journal of Engineering and Technology (IRJET), 2020.

[6] A. Mahmoud, "Political Sentiment Analysis Using Twitter Data."

[7] A. Pathare, A. Nishad, S. Dubey, S. Malve, and U. Goradiya, "Airlines Twitter Sentiment Analysis Using EDA AND ML," in International Journal of Scientific Research in Computer Science, Engineering and Information Technology, vol. 10, no. 2, 2023.

[8] A. Mishra, M. S. Wajid, and U. Dugal, "A Comprehensive Analysis of Approaches for Sentiment Analysis Using Twitter Data on COVID-19 Vaccines," in Journal of Informatics Electrical and Electronics Engineering, 2021.

[9] A. Modi, K. Shah, S. Shah, S. Patel, and M. Shah, "Sentiment Analysis of Twitter Feeds Using Flask Environment: A Superior Application of Data Analysis," 2022.

[10] Padmanayana and B. K. Bhavya, "Stock Market Prediction Using Twitter Sentiment Analysis," in International Journal of Scientific Research in Computer Science, Engineering and Information Technology, vol. 7, no. 4, 2021.

[11] A. Ragothaman and C.-Y. Huang, "Sentiment Analysis on Covid-19 Twitter Data," in International Journal of Computer Theory and Engineering, vol. 13, no. 4, 2021.

[12] P. Sumathy and S. M. Muthukumari, "Sentiment Analysis of Twitter Data Using Multi Class Semantic Approach," in International Journal of Scientific Research in Computer Science, Engineering, and Information Technology.

[13] S. Patil, B. Wagh, A. Bhinge, A. Sahal, and M. Ingale, "Twitter Sentiment Analysis on Government Law Using Real-Time Data," in International Journal of Scientific Research in Science and Technology.

[14] M. M. Bhajibhakare, A. Borkar, S. Naik, S. Solase, and P. Kunjir, "Sentiment Analysis on YouTube & Twitter Data using Machine Learning," in International Journal for Research in Applied Science & Engineering Technology (IJRASET).

[15] A. H. Ali, H. Kumar, and P. J. Soh, "Big Data Sentiment Analysis of Twitter Data," in Mesopotamian Journal of Big Data, 2021.

[16] N. G. Bailur and M. Meleet, "Sentiment Analysis on Twitter Data using ML," in International Research Journal of Engineering and Technology (IRJET), July 2020.

[17] R. L. Chouhan, "Sentiment Analysis of Pulwama Attack Using Twitter Data," in Springer Nature Singapore Pte Ltd., 2021.

[18] M. Khan, A. Malviya, and S. K. Yadav, "Big Data Approach of Sentiment Analysis of Twitter Data Using K-Mean Clustering Approach," in International Journal of Mechanical and Production Engineering Research and Development (IJMPERD), vol. 10, no. 3, June 2020.

[19] V. S. Saketh, Y. K. Guntupalli, and D. S. Vaishnav, "Sentiment Analysis of Live Twitter Data using Apache Spark," in International Research Journal of Engineering and Technology (IRJET), vol. 7, no. 8, August 2020.

Appendix 1 – Code of the Project

```
import nltk
import pandas as pd
from textblob import TextBlob
from pandas_profiling import ProfileReport
from wordcloud import WordCloud
import matplotlib.pyplot as plt
from IPython.display import HTML, display, Image

# Set Matplotlib backend to display inline in the notebook
%matplotlib inline

# Download NLTK data (if not already downloaded)
nltk.download("punkt")

# Load the CSV dataset (replace 'covid19_tweets.csv' with your file path)
df = pd.read_csv('covid19_tweets.csv')

# Specify the name of the column containing the text to analyze
text_column = 'text'  # Replace 'text' with the actual column name

def analyze_sentiment(text):
    analysis = TextBlob(text)
    # Sentiment polarity ranges from -1 (negative) to 1 (positive)
    return analysis.sentiment.polarity

# Sentiment Analysis
max_rows = 500
row_count = 0
neutral_sentiments= []
positive_sentiments = []
negative_sentiments = []
```

```
for index, row in df.iterrows():
    if row_count >= max_rows:
        break  # Exit the loop after processing 200 rows

    text = row[text_column]

    polarity = analyze_sentiment(text)

    if polarity > 0:
        sentiment = "Positive"
        positive_sentiments.append(row)
    elif polarity < 0:
        sentiment = "Negative"
        negative_sentiments.append(row)
    elif polarity == 0:
        sentiment = "Neutral"
        neutral_sentiments.append(row)



    print(f"Row {index + 1}: '{text}'")
    print(f"Sentiment: {sentiment}")
    print(f"Polarity: {polarity}\n")

    row_count += 1

# Create DataFrames containing rows with positive and negative sentiments
df_positive_sentiments = pd.DataFrame(positive_sentiments)
df_negative_sentiments = pd.DataFrame(negative_sentiments)

# Data Profiling for positive sentiments
profile_positive = ProfileReport(df_positive_sentiments)
profile_positive.to_file(output_file="Positive_Sentiments_Profile_Report.html")

# Data Profiling for negative sentiments
```

```
profile_negative = ProfileReport(df_negative_sentiments)
profile_negative.to_file(output_file="Negative_Sentiments_Profile_Report.html")

# Create and display the Word Clouds for positive and negative sentiments
positive_text = " ".join(df_positive_sentiments[text_column].tolist())
negative_text = " ".join(df_negative_sentiments[text_column].tolist())

# Word Cloud for Positive Sentiments
wordcloud_positive = WordCloud(width=800, height=400,
background_color='white').generate(positive_text)
plt.figure(figsize=(10, 5))
plt.imshow(wordcloud_positive, interpolation='bilinear')
plt.axis("off")
plt.title("Word Cloud for Positive Sentiments")
plt.show()

# Word Cloud for Negative Sentiments
wordcloud_negative = WordCloud(width=800, height=400,
background_color='white').generate(negative_text)
plt.figure(figsize=(10, 5))
plt.imshow(wordcloud_negative, interpolation='bilinear')
plt.axis("off")
plt.title("Word Cloud for Negative Sentiments")
plt.show()

# Save the Word Clouds as image files and display them
wordcloud_positive.to_file("positive_wordcloud.png")
wordcloud_negative.to_file("negative_wordcloud.png")

positive_count = len(df_positive_sentiments)
negative_count = len(df_negative_sentiments)
# Assuming you have already calculated positive_count and negative_count as the counts of positive
and negative sentiments
# Calculate the count of the opposite sentiment
```

```
total_count = 500
neutral_count = total_count - (positive_count + negative_count)

# Create data for the pie chart
sentiments = ['Positive', 'Negative', 'Neutral']
counts = [positive_count, negative_count, neutral_count]
colors = ['green', 'red', 'gray']

# Create a pie chart
plt.figure(figsize=(8, 8))
plt.pie(counts, labels=sentiments, colors=colors, autopct='%1.1f%%', startangle=140)
plt.title('Sentiment Distribution (Total Count: 500)')
plt.axis('equal')  # Equal aspect ratio ensures that pie is drawn as a circle.
plt.show()

import pandas as pd
import matplotlib.pyplot as plt
# Load the CSV dataset (replace 'covid19_tweets.csv' with your file path)
df = pd.read_csv('covid19_tweets.csv')
# Extract the tweet counts for each user
tweet_counts = df['user_name'].value_counts()
# Select the top 10 people with the most tweets
top_10_people = tweet_counts.head(10)

# Display the list of top 10 people by tweet count
print("Top 10 People by Tweet Count:")
print(top_10_people)

# Create a vertical bar chart for the top 10 people
plt.figure(figsize=(6, 2))
plt.bar(top_10_people.index, top_10_people.values, color='skyblue')
plt.xticks(rotation=90)  # Rotate x-axis labels for better readability
plt.xlabel('User Name')
plt.ylabel('Tweet Count')
plt.title('Top 10 People by Tweet Count')
plt.show()

import pandas as pd
# Load data
data = pd.read_csv("covid19_tweets.csv",
                 dtype={'Country/Region': str,
                        'Province/State': str,
                        'Latitude': float,
                        'Longitude': float,
                        'Confirmed': float,
                        'Recovered': float,
                        'Deaths': float})

# Rename columns
data = data.rename(columns={"Country/Region": "Country", "Province/State": "State"})
# Load tweet data
tweets = pd.read_csv("covid19_tweets.csv",
                   dtype={'user_name': str,
                          'user_location': str,
                          'user_description': str,
                          'user_followers': float,
                          'user_friends': float,
                          'user_favourites': float,
                          'user_verified': bool,
                          'text': str,
                          'hashtags': str,
                          'source': str,
                          'is_retweet': bool})

# Convert date columns to datetime
tweets['user_created'] = pd.to_datetime(tweets['user_created'])
tweets['date'] = pd.to_datetime(tweets['date'])
```

```
# Inspect data
print("Data Table:")
print(data.head(5))
# Inspect tweet data
print("\nTweet Data Table:")
print(tweets.head(2))

import pandas as pd
import matplotlib.pyplot as plt
# Load the CSV dataset (replace 'covid19_tweets.csv' with your file path)
df = pd.read_csv('covid19_tweets.csv')

# Assuming you have a column in your dataset that contains user locations, let's call it 'user_location'
user_location_column = 'user_location'  # Replace 'user_location' with the actual column name
# Extract all user locations
user_locations_all = df[user_location_column]
# Get the top 15 unique user locations and their frequencies
top_15_user_locations = user_locations_all.value_counts().head(15)

# Create a DataFrame to store the data
top_user_locations_df = pd.DataFrame({'User Location': top_15_user_locations.index, 'Frequency':
top_15_user_locations.values})

# Create a bar chart for the top 15 user locations
plt.figure(figsize=(10, 6))
plt.barh(top_user_locations_df['User Location'], top_user_locations_df['Frequency'], color='skyblue')
plt.xlabel('Frequency')
plt.ylabel('User Location')
plt.title('Top 15 User Locations by Frequency')
plt.gca().invert_yaxis()  # Invert the y-axis to display the top location at the top
plt.show()

import pandas as pd
import matplotlib.pyplot as plt
```

```
# Set the figure size
plt.rcParams['figure.figsize'] = (14, 6)

# Assuming you have a DataFrame called 'tweets' with a column 'date'
# Convert the 'date' column to datetime
tweets['date'] = pd.to_datetime(tweets['date'])
# Group by date and count the number of tweets
tweet_counts = tweets.groupby('date').size().reset_index(name='n')
# Create a line plot
plt.plot(tweet_counts['date'], tweet_counts['n'], linewidth=1.5, color='blue')
# Set the plot limits to not clip data points
plt.xlim(tweet_counts['date'].min(), tweet_counts['date'].max())

# Customize the plot appearance (you can define 'my_theme' and 'my_colors' separately)
plt.title("Number of Tweets in Time", fontsize=16)
plt.xlabel("Date", fontsize=14)
plt.ylabel("Frequency", fontsize=14)

# Remove x-axis title
plt.gca().set_xlabel('')

# Show the plot
plt.show()
```