AI Presentation

Sentiment Analysais On Future Post

Team Lead :Rohma Talpur

Team Collaboration

- 1. Research Team: Wisal and Shahzaib
- 2. Dataset Team: Shaheer and Wajiha
- 3. Preprocessing Team: Essa and Suliman
- 4. Model Training Team: Hasnian and Rafay
- 5. Content Writers: Sakhawat and Rayyan
- 6. Testing: Wisal and Hassan

Agenda

Defining the Problem

Research

Dataset Collection

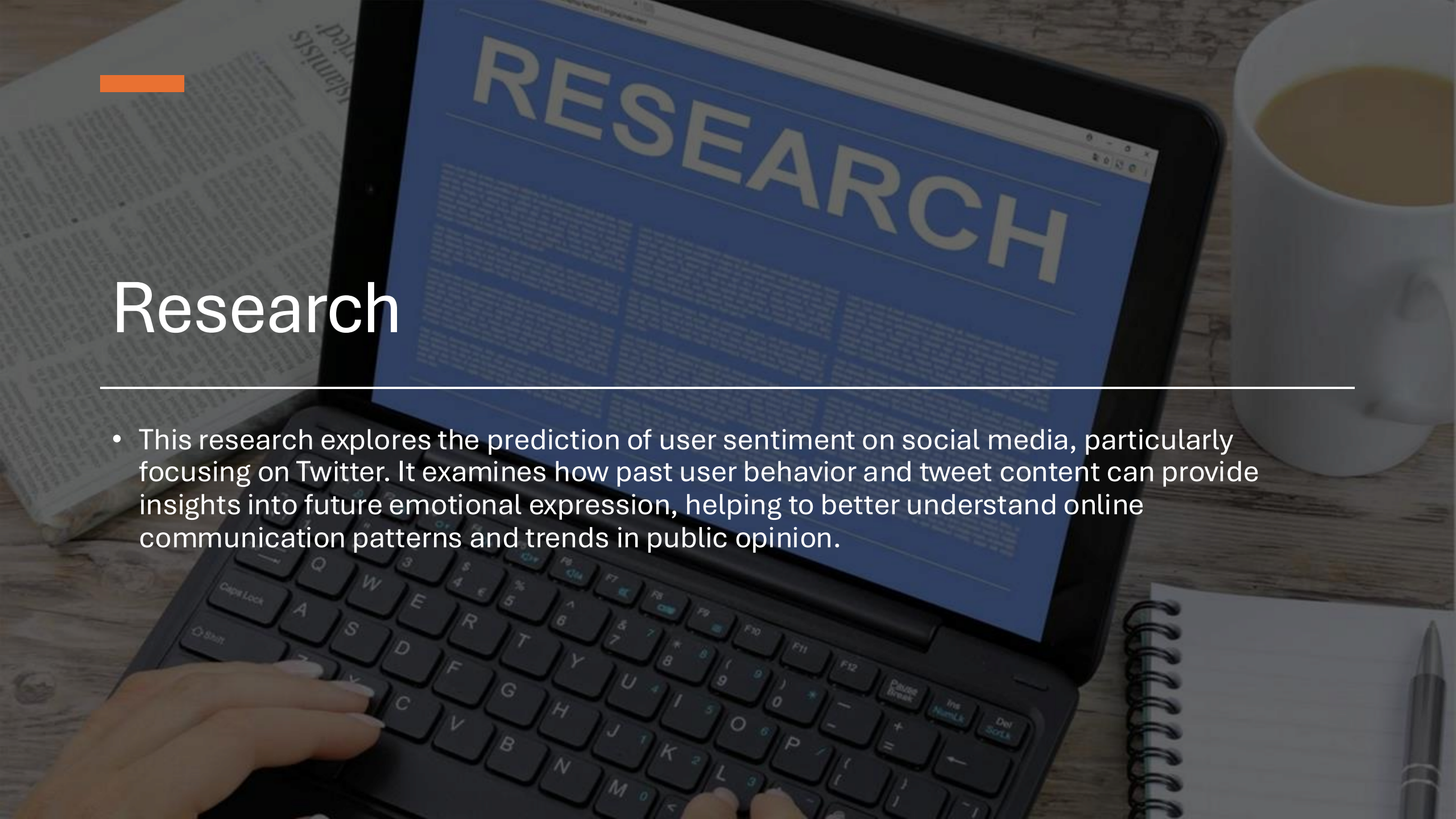
Pre Processing

Model Training

Testing

Defining the Problem

- The problem is to predict the sentiment (tone) of a user's next tweet based on the previous tweet's content and the user's identity. This can provide valuable insights into user behavior, social media trends, and online communication patterns.



Research

- This research explores the prediction of user sentiment on social media, particularly focusing on Twitter. It examines how past user behavior and tweet content can provide insights into future emotional expression, helping to better understand online communication patterns and trends in public opinion.

Dataset

- The dataset used in this research was collected from Kaggle and was also scraped from Twitter. It includes historical tweet data, providing insights into user behavior and sentiment patterns, which are leveraged to predict the tone of future tweets.

Initial Attempt: Scraping Tweets from Twitter

1. Initially, we decided to collect tweets directly from Twitter using Tweepy.

2. Tweepy is a Python library that provides access to the Twitter API for scraping and managing tweets.

3. However, due to Twitter being banned in Pakistan, we had to use a VPN to access the platform.

4. Despite using a VPN, we encountered frequent connection issues, making the process frustratingly slow.

5. The data retrieval rate was significantly impacted by these limitations, making direct scraping infeasible.

Shift to Kaggle: Searching for Pre-existing Datasets

1. Recognizing the inefficiency of scraping, we turned to Kaggle to search for pre-existing Twitter datasets.

2. While Kaggle provided a variety of datasets, most were generic and did not align with our specific requirements.

3. Our goal was to collect data from diverse users, but many datasets focused on a narrow range of users or topics.

4. The scarcity of data from regular users tweeting frequently led us to refine our approach to target specific categories.

Refining Focus: Targeting Specific User Groups

1. To overcome the lack of consistent data from regular users, we decided to target specific groups like politicians.

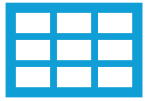
2. We aimed to collect data from Pakistani politicians but faced challenges due to limited activity, except for Imran Khan.

3. To expand our dataset, we included international figures like Donald Trump and other world leaders.

4. Additionally, we targeted famous personalities such as Cristiano Ronaldo and Lionel Messi for a diverse dataset.

5. This approach required careful manual collection to ensure each user contributed at least 1,000 tweets.

Preprocessing Overview



Objective:

Combine multiple tweet datasets into a single cohesive dataset for analysis

Tools Used:

Python

Pandas library

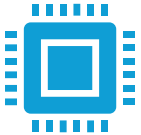
Google Colab

Challenges:

Handling different data formats and encodings

Ensuring data integrity and avoiding duplicates

Managing large datasets efficiently



Solution Approach:

Automated the merging process using a Python script

Implemented error handling to manage read failures and encoding issues

Used Pandas for efficient data manipulation and merging

Data Merging Process and Code Snippet

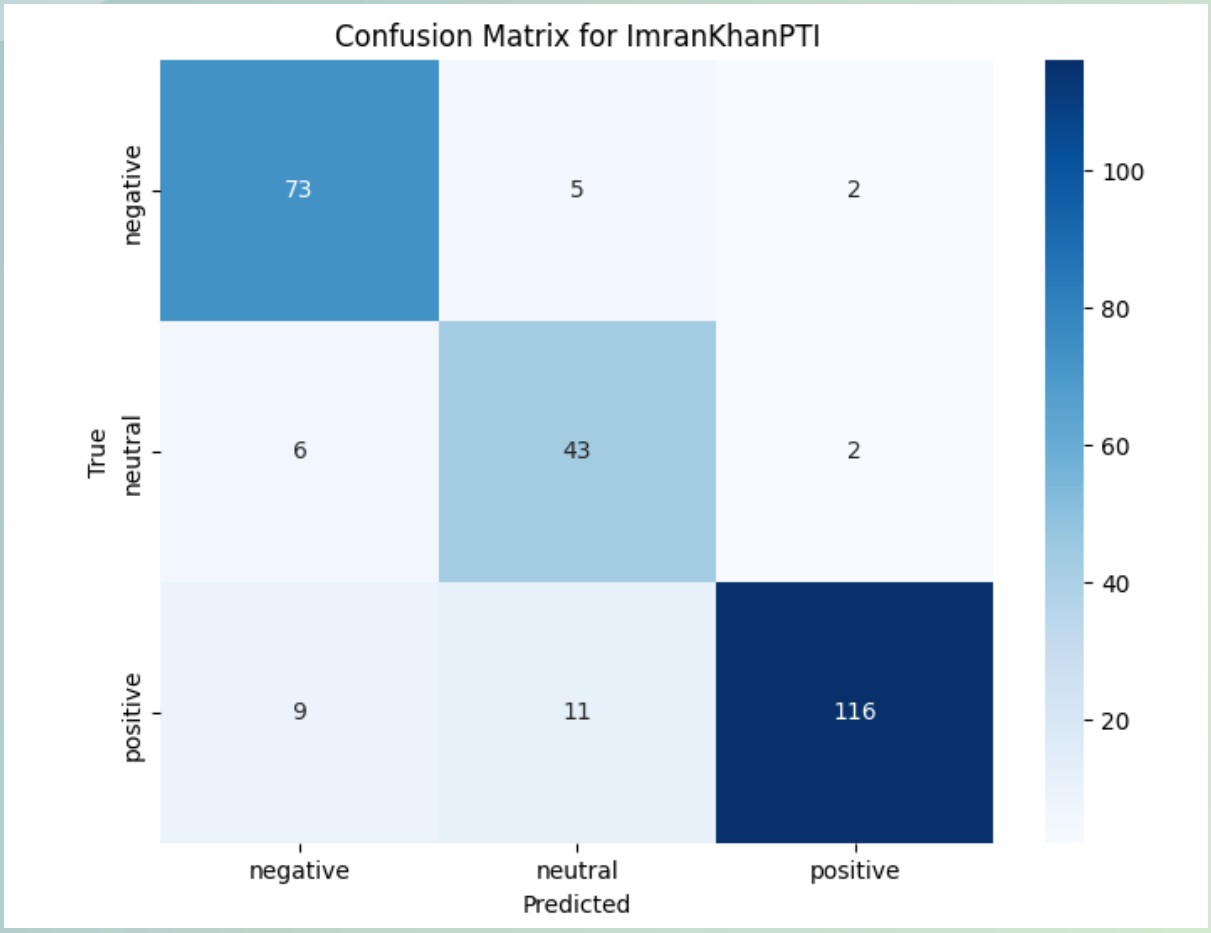
Step-by-Step Process:

- **Import Libraries**: pandas, os, google.colab.drive
- **Mount Google Drive**: Access datasets stored on Google Drive
- **List CSV Files**: Created a list of all CSV files to be merged
- **Read CSV Files**: Loaded each CSV into a Pandas DataFrame with error handling
- **Append DataFrames**: Collected all DataFrames into a list
- **Concatenate DataFrames**: Merged all DataFrames into one using pd.concat()
- **Export Merged Data**: Saved the merged DataFrame to a new CSV file

Training the LSTM Model for Sentiment Analysis

- **Objective:** Classify sentiments (positive, neutral, negative) from tweets.
- **Dataset:** Preprocessed tweets (cleaned text, tokenized, padded).
- **Model:**
  - LSTM with Embedding and Dense layers.

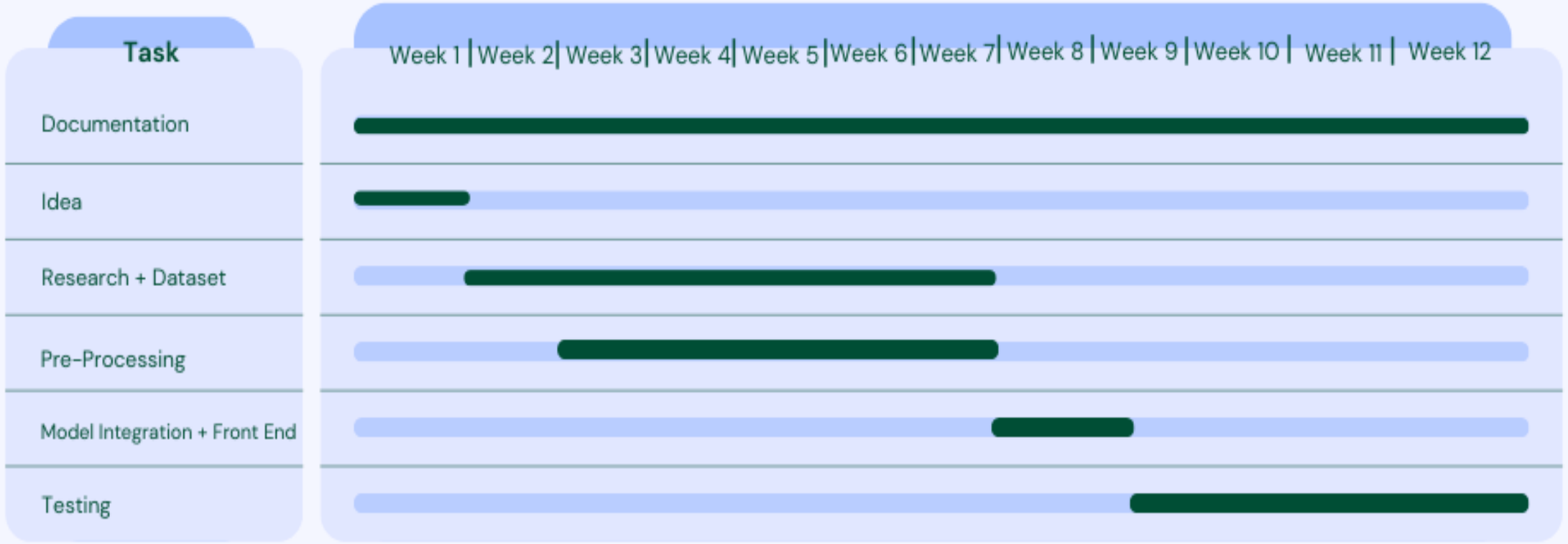
Model Training & Performance



Testing

Test Case ID	User	Tweet Content	Topic	Expected Sentiment	Pass/Fail
SA-001	Imran khan	Nawaz Sharif's policies have done more harm than good for the economy.	Nawaz shreef	Negative	Pass
SA-002	Imran khan	Climate change initiatives are crucial for the future.	Climate Change	Positive	Pass
SA-003	Imran khan	Rising inflation is a burden on every family, and it's unacceptable	Economy	Positive	Pass
SA-004	Imran khan	Cryptocurrencies are unpredictable and risky for beginners.	Cryptocurrency	Netural	Pass
SA-005	Elon Musk	Launching Starship for its first orbital flight today!	Space Exploration	Netural	pass
SA-006	Elon Musk	Climate change initiatives are crucial for the future	Climate Change	Positive	Fail
SA-007	Elon Musk	Cryptocurrencies are unpredictable and risky for beginners.	Cryptocurrency	Negative	Pass
SA-008	Elon Musk	It's always exciting to see SpaceX pushing the boundaries of space exploration	Space	Positive	pass
SA-009	Randio	Tax reforms are necessary to create a fairer economy, but execution is key	Economic Policies	Neutral	pass
SA-010	Randio	Rising inflation is a burden on every family, and it's unacceptable	Economy	Negative	fail
SA-011	Randio	Attended a community event to discuss mental health awareness	Mental Health	Netural	pass

Project Timeline: Gantt Chart



Thank You