Phu Qui Dang (22095834)

Latrobe university

Sprint 2 – NLP Insights: SDG5\_Gender\_Equality Dataset

|  |
| --- |
| Table of Contents  [Task 1: N-Gram Analysis 2](#_Toc197473531)  [Task 2: Temporal Analysis 3](#_Toc197473532)  [Task 3: Sentiment Analysis 4](#_Toc197473533)  [Task 4: Topic Modelling 5](#_Toc197473534)  [Task 5: Generative AI Insight 7](#_Toc197473535) |

# Task 1: N-Gram Analysis

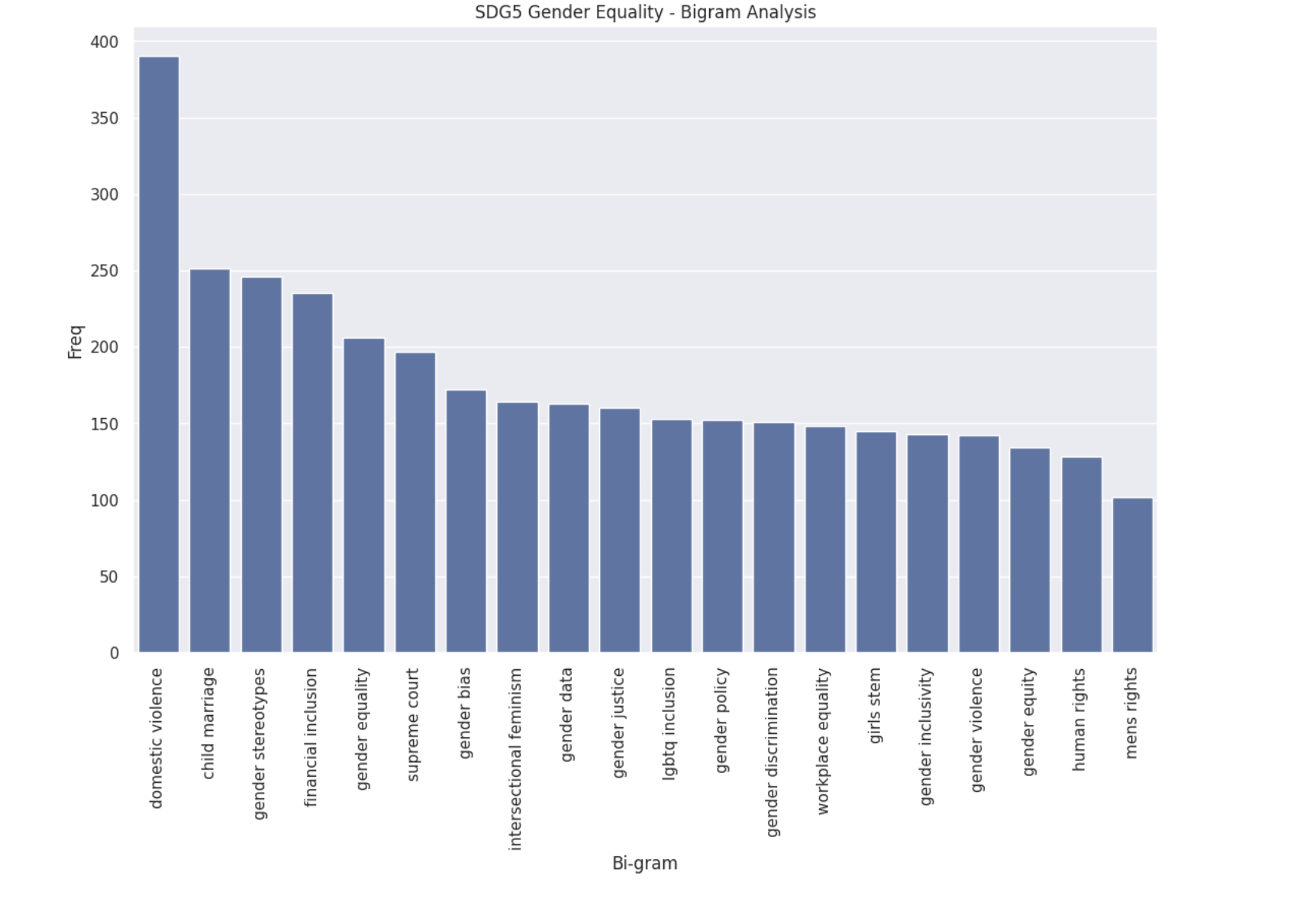
**Description:**

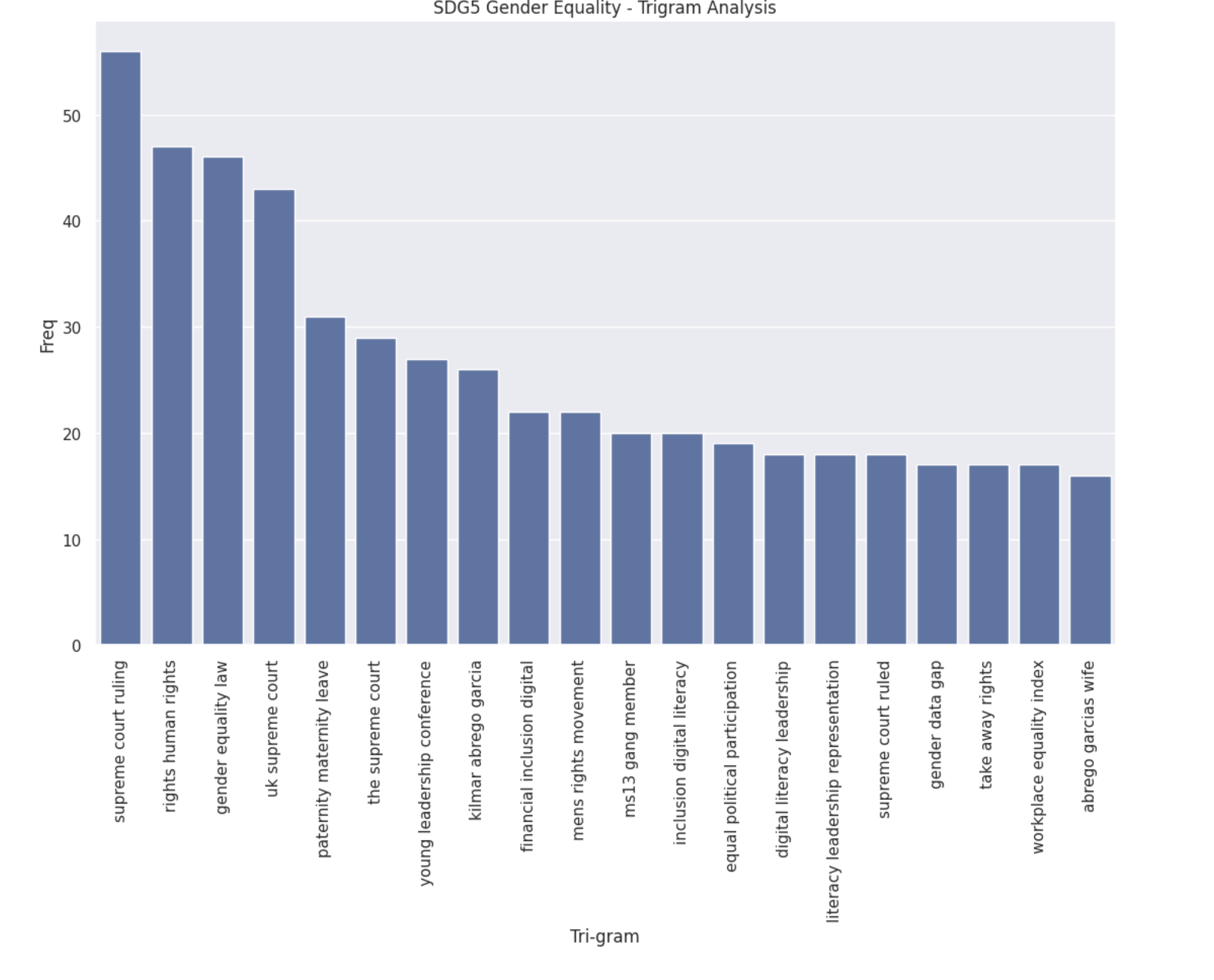
N-gram analysis is a Natural Language Processing technique used to identify frequent sequences of words in a text corpus. In this task, we generated **bigrams (2-word combinations)** and **trigrams (3-word combinations)** to uncover common phrases and patterns in tweets related to SDG5 – Gender Equality. This helps reveal dominant themes and issues in public discussions.

**Input Data Used:**

The dataset used for this analysis was imported from the file SDG5\_Gender\_Equality.xlsx, which contains tweets specifically related to Sustainable Development Goal 5: Gender Equality. The relevant text data for N-gram extraction was taken from the Text column of the dataset. This column includes raw tweet content discussing various gender-related issues. The text entries were converted into a list using the .tolist() method to prepare them for vectorization. The processed text list was then passed into a custom get\_ngrams() function, which utilized scikit-learn’s CountVectorizer to generate the most frequent bigrams (2-word phrases) and trigrams (3-word phrases) within the dataset.

**Outcomes:**



**Insights:**

Bigrams with the highest frequencies in the dataset include "domestic violence," "child marriage," and "gender stereotypes," thereby suggesting these as some of the most prominent issues relating to gender equality that find expression on Twitter. Other issues that garner much talk include "financial inclusion," "gender equality," and "lgbtq inclusion," denoting considerations about security, rights, and representation.

The most used trigrams include "supreme court ruling," "rights human rights," and "gender equality law," indicating that the target of legal and policy conversations. Other phrases used frequently point to concerns about workplace equity and granting access to digital resources such as "paternity maternity leave" and "financial inclusion digital.

# Task 2: Temporal Analysis

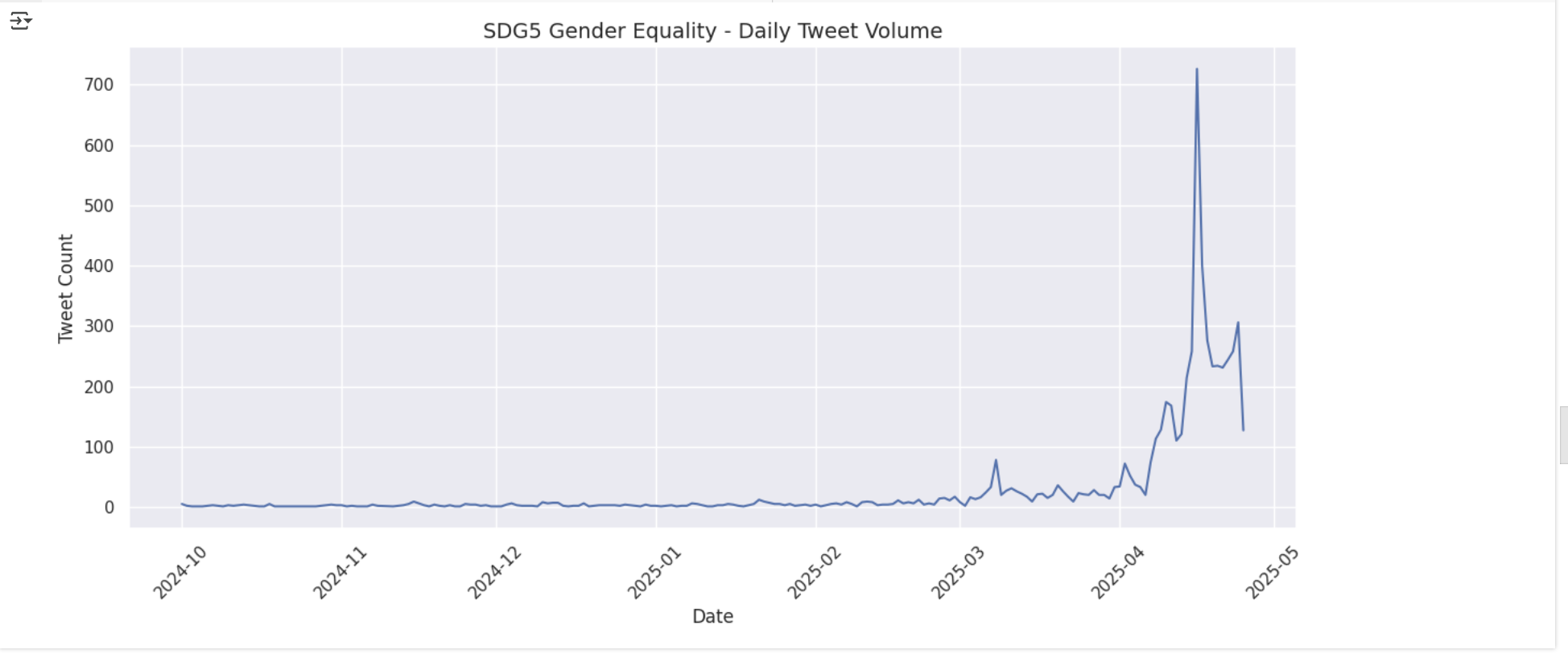
**Description:**

The temporal analysis aimed to spot trends, anticipate spikes, or watch the patterns of emergent phenomena over time. This task involves looking at daily tweet volumes on SDG5 (Gender Equality) to comprehend the alterations in public discussion during the observed period.

**Input Data Used:**

The analysis used the day and Text columns from SDG5\_Gender\_Equality.xlsx. Tweets were grouped by posting dates, enabling the calculation of their daily frequency to be used for engagement visualization through time.

**Outcomes:**

**Insights:**

# Until March 2025, tweet activity on SDG5, Gender Equality remained consistently low; thereafter, a sharp increase came into being in early April 2025. The volume of tweets reached its maximum between 700 and more in a single day.

# Task 3: Sentiment Analysis

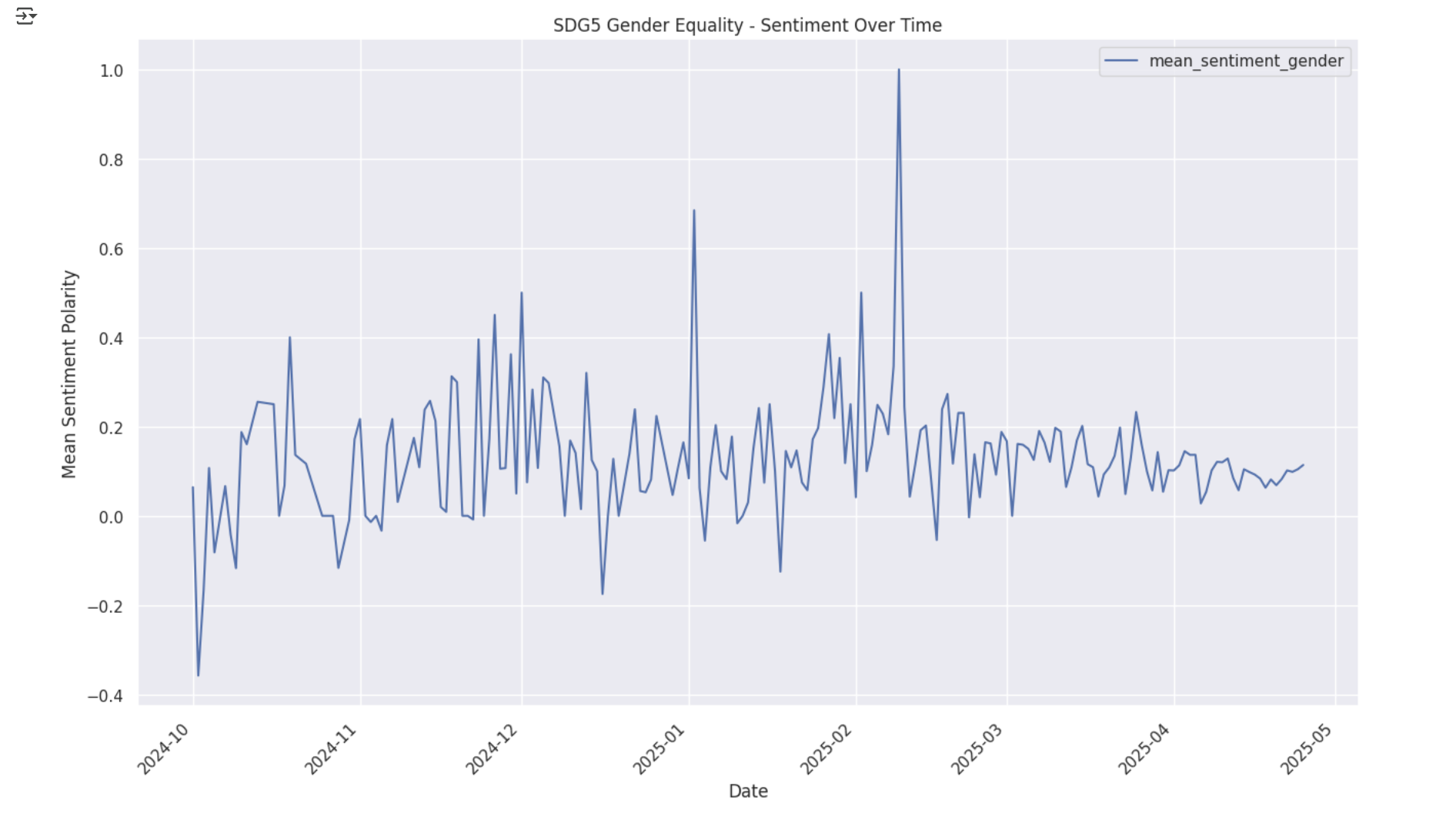
**Description:**

This analysis used TextBlob to evaluate the sentiment polarity of each tweet in the SDG5\_Gender\_Equality dataset. TextBlob calculates a sentiment score between -1 (very negative) and 1 (very positive) based on the polarity of the text. By aggregating these scores daily, the analysis uncovers changes in public sentiment toward gender equality topics over time.

**Input Data Used:**

The analysis used the Text column from the df\_gender dataframe, which contains tweet content related to SDG5 (Gender Equality). Each tweet was evaluated using TextBlob(x).sentiment.polarity, and the results were grouped by the day column to compute daily average sentiment.

**Outcomes:**



**Insights:**  
The sentiment timeline shows steady fluctuations in the public opinion on gender equality. From October 2024 through early 2025, the sentiment mostly remained from moderately positive to occasional negative dips. The most remarkable feature was the sharp spike in sentiment in February 2025, a near 1.0-level surge, which indicates an explosion of highly positive tweets following some big event, announcement, or campaign. Following this spike, sentiment remains buoyant but gradually recedes in April 2025.

**Recommendations:**

Governments and policymakers can seize opportunities arising from peaks in positive sentiment to discuss or strengthen gender equality policies and campaigns. The spike witnessed in February 2025 suggests a great deal of public attention, and such a time could prove critical for policy action. Wherever sentiment declines, review of social discourse or events presumably causing things of concern to the public must commence. NGOs would benefit from coordinating their outreach efforts with the sentiment trends by intensifying efforts during upbeat phases and tackling issues in sentiment downturns. For the entire public, continuous engagement and support for gender equality, especially when sentiments take a downturn-will help maintain the momentum and counter misinformation.

# Task 4: Topic Modelling

**Description:**

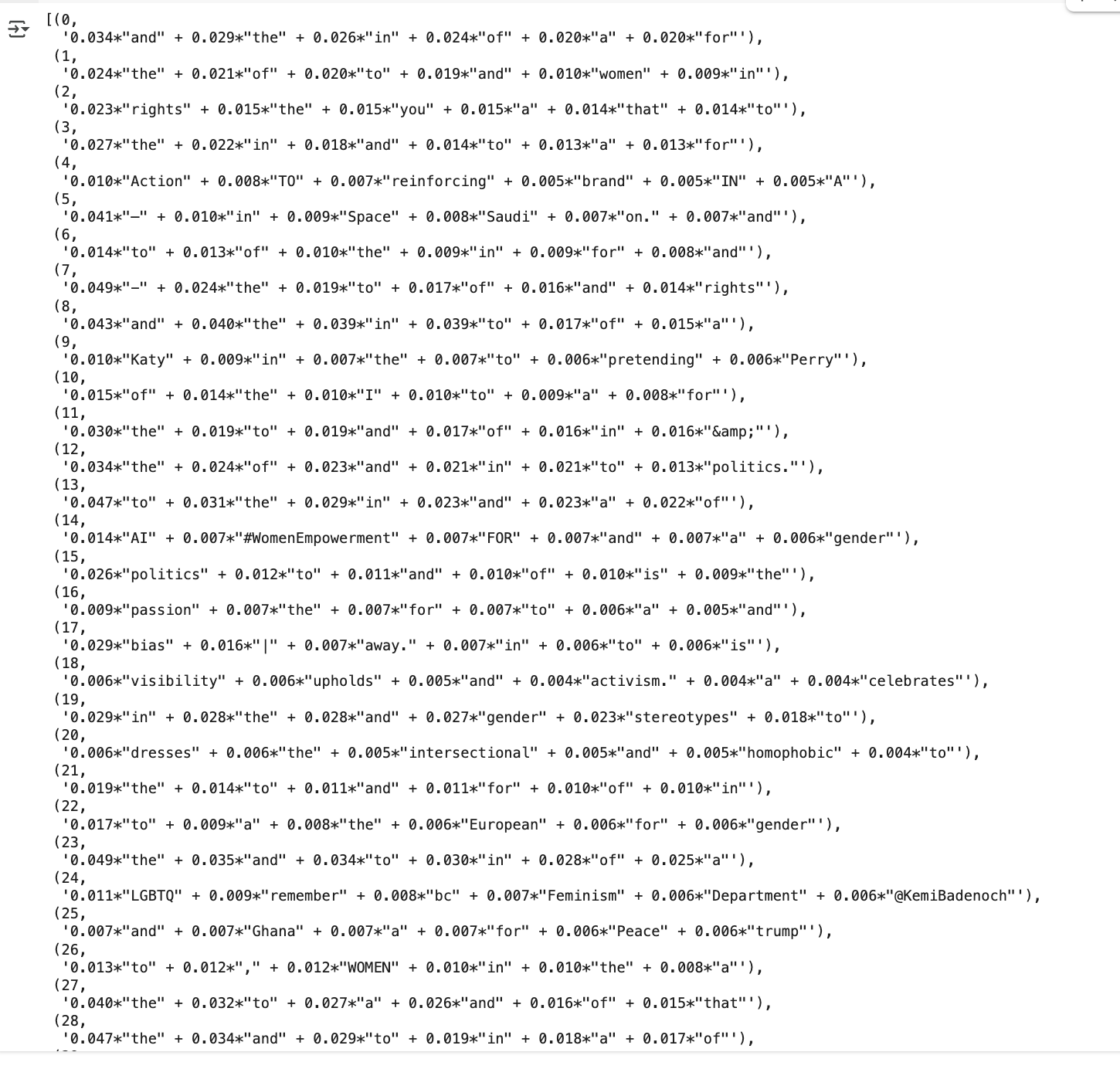
Topic modelling is an unsupervised machine learning technique used to discover hidden thematic structures within a large corpus of text. In this analysis, Latent Dirichlet Allocation (LDA) was employed to identify latent topics in SDG5 gender equality-related tweets. The model identifies groups of co-occurring words that form coherent topics, which can then be manually interpreted.

**Input Data Used:**

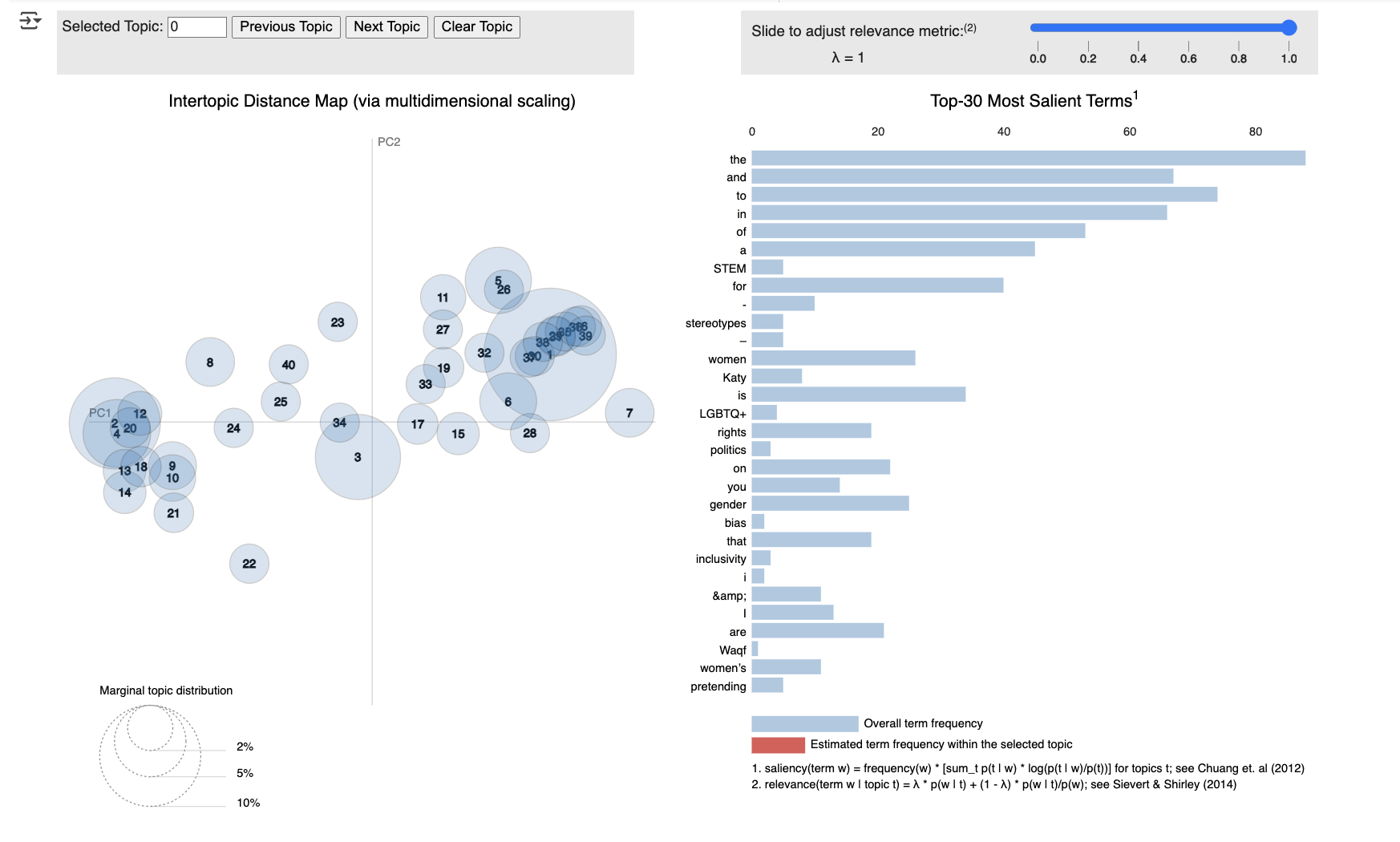
The input corpus consists of preprocessed tweets from the SDG5\_Gender\_Equality.xlsx dataset. The ‘Text’ column was tokenized and converted into a bag-of-words (BoW) representation using Gensim’s Dictionary and doc2bow() functions. The LDA model was built with 40 topics, and the pyLDAvis library was used for interactive visualization and interpretation.

**Outcomes:**

Generative the topic model:



Interactive Topic Analyzer:



**Insights:**

Generative the topic model Insight:  
Example: Topic 7

Keywords: "rights", "of", "the", "to", "and"

Interpretation: This topic likely centers around human rights or gender rights discourse. While many top words are common stopwords, the presence of “rights” as a weighted term suggests the topic relates to discussions about entitlements, equality, or activism around rights.

Interactive Topic Analyzer Insight:

The intertopic distances map depicts a well-distributed set of topics with minimal overlaps, thus representing clearly delineated thematic groupings in the SDG5 gender equality dataset. On the right, the most salient keywords include "STEM", "stereotypes", "women", "rights", "LGBTQ+", "politics", "bias", and "inclusivity", signifying the main areas of discourse. These keywords imply that public conversations on gender equality revolve around social inclusion, systemic discrimination, political advocacy, and gender representation in the STEM fields. This visualization further strengthens the argument, and also the LDA model has captured meaningful themes, with keywords such as "rights", "bias", and "LGBTQ+" emphasizing the major concerns expressed in the tweets.

# Task 5: Generative AI Insight

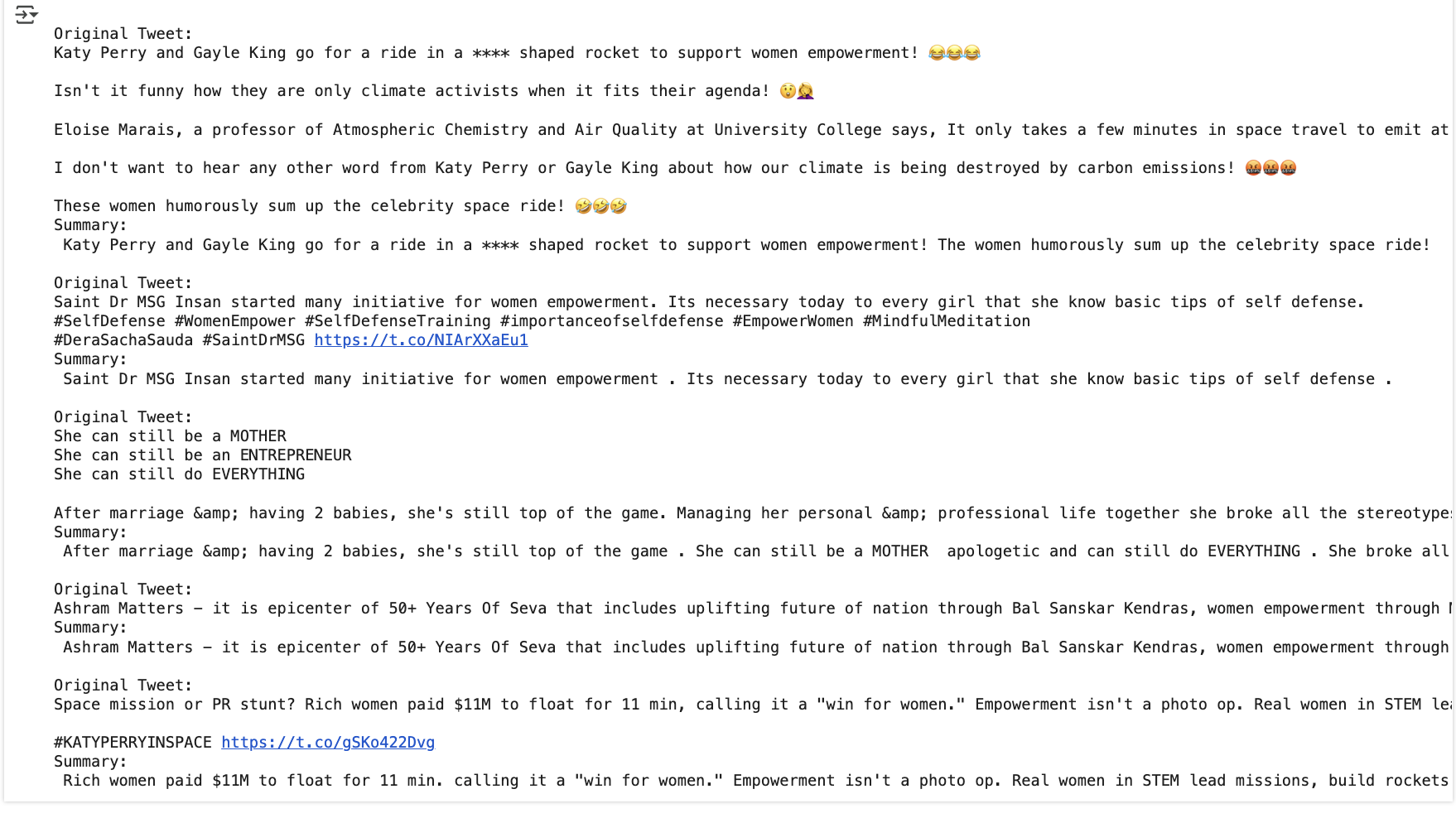
**Description:**

This technique uses a pre-trained transformer-based summarization model, distilbart-cnn-12-6, from the Hugging Face library. It condenses long SDG5-related tweets into short summaries, preserving key ideas and sentiments. This enables analysts to quickly extract the essence of gender equality discussions from large volumes of social media data.

**Input Data Used:**

Tweets from the SDG5\_Gender\_Equality.xlsx dataset were filtered to include only those with more than 30 words. The tweet text column ('Text') served as input to the summarization pipeline. The summarizer generated condensed versions of the tweets, which were later used for further insight generation.

**Outcome:**

****

**Insights:**

The summarization output reveals dominant themes in gender equality discourse, especially related to representation, empowerment, and public perception. Notably:

With such tweets referring to Katy Perry and Gayle King's "space ride" as an iconic act, the critics deride shallow shows of empowerment and allude to activism being performative and marketing-oriented at the expense of honest change.

Women’s Empowerment: Multiple tweets emphasize genuine empowerment narratives, including the importance of self-defense education for girls and highlighting real women breaking stereotypes in motherhood, leadership, and STEM.

Cultural and Institutional Support: References to initiatives like Ashram Matters and women-led efforts suggest institutional support plays a key role in sustained gender equality efforts.

Skepticism and Satire: Some tweets use sarcasm or skepticism to challenge the legitimacy of empowerment events, especially when involving media spectacle or political symbolism.