SENTIMENT ANALYSIS AND TOPIC MODELING ON SOCIAL MEDIA AND NEWS PLATFORMS USING VADER AND RoBERTa

PROJECT REPORT

Submitted by

HARIHARAN SRIRAM

310623205044

Easwari Engineering College

Ramapuram, Chennai – 600089

Submitted as a part of:

Internship at

SIFY TECHNOLOGIES LIMITED

Tidel Park, Tharamani, Chennai - 600113

July 2025

ACKNOWLEDGEMENT

I sincerely express my gratitude to my mentors and colleagues at Sify Technologies and my academic institution who provided me continuous support and guidance during the course of this project.

I am deeply grateful for the accessible resources from Google Colab, HuggingFace Transformers, NLTK, YouTube's API and Reddit's API, which have been instrumental in the successful completion of this project.

Finally, I would also like to thank my peers and reviewers for their valuable feedback during this internship.

ABSTRACT

This project explores the implementation of **Sentiment Analysis** and **Topic Modeling** on data gathered from various digital platforms including **YouTube, Reddit, Medium, and Bing News**. The goal is to offer a consolidated tool capable of collecting, analyzing, and visualizing public sentiment and topic trends across these platforms.

For sentiment analysis, two well-established models were used:

- **VADER** (Valence Aware Dictionary and Sentiment Reasoner)
- **RoBERTa** (Robustly Optimized BERT Approach)

The project additionally integrates **Topic Modeling using NMF (Non-Negative Matrix Factorization)** to extract meaningful insights from the gathered data.

Outcomes include:

- Clean, deduplicated datasets.
- Sentiment analysis, classification and scores.
- Topic clusters with key discussion themes.
- Sentiment versus Topic intensity split-up
- Informative visualizations.

TABLE OF CONTENTS

INTRODUCTION

In today's digital landscape, **public sentiment analysis** is critical for businesses, researchers, and governments to understand public opinion trends across social media and news platforms. This project aims to solve the gap of collecting data across multiple sources, analyzing it with robust AIML models, and presenting it in a user-friendly manner.

Platforms Scraped:

- **YouTube** (comments)
- **Reddit** (posts)
- **Medium** (articles)
- **Bing News** (articles)

Models Used:

- **VADER** (Valence Aware Dictionary for Sentiment Reasoner
- **RoBERTa** (Robustly Optimized BERT Approach)

This project combines **scraping, sentiment classification, topic extraction, and visualization** into a unified pipeline for efficient analysis.

PROBLEM STATEMENT

In the current digital era, a massive volume of user-generated content is produced daily across various online platforms such as **YouTube, Reddit, Medium, and news portals**. This content reflects valuable public opinions, sentiments, and trends which, if analyzed effectively, can provide deep insights for businesses, researchers, and individuals. However, **manually analyzing such vast amounts of textual data is not only impractical but also highly inefficient and time-consuming**.

The challenge lies in gathering reliable, deduplicated, and valid content from varied platforms and applying sophisticated models to extract actionable insights.

Traditional tools:

- Often limited to a single platform.
- Lack flexibility for deep sentiment or topic analysis.

This project bridges these gaps through:

- **Cross-platform data extraction.**
- **Advanced NLP models.**
- **Readable and insightful visualizations.**

Furthermore, the current solutions rarely provide a **seamless end-to-end pipeline that integrates scraping, deduplication, cleaning, and comparative analysis of sentiment results from multiple models**.

Additionally, platforms like Medium, Reddit, and YouTube each have their unique formats and restrictions for accessing data, making unified sentiment analysis across them challenging. Users seeking meaningful insights from these platforms often lack a unified, flexible, and easy-to-use tool that can automate this process.

OBJECTIVE

This project aims to **develop a comprehensive sentiment analysis framework** that:

- Scrapes and collects data from platforms such as **YouTube, Reddit, Medium, and Bing News**.
- Cleans and deduplicates the collected data to ensure only **valid and unique articles/posts/comments are considered**.
- Applies both **VADER (rule-based)** and **RoBERTa (transformer-based)** models to perform sentiment analysis, offering users a comparative understanding of results from different approaches.
- Presents insights through **visualizations like count plots and heatmaps** to enhance interpretability.
- Provides the user with **an option to select the sentiment model** based on their specific needs, through an intuitive interface.

TECHNOLOGIES USED

➢ Languages & Tools:

- Python
- Google Colab

➢ Libraries:

NLTK, BeautifulSoup, requests, pandas, matplotlib, seaborn, transformers, tqdm, praw, sklearn, newspaper3k, torch

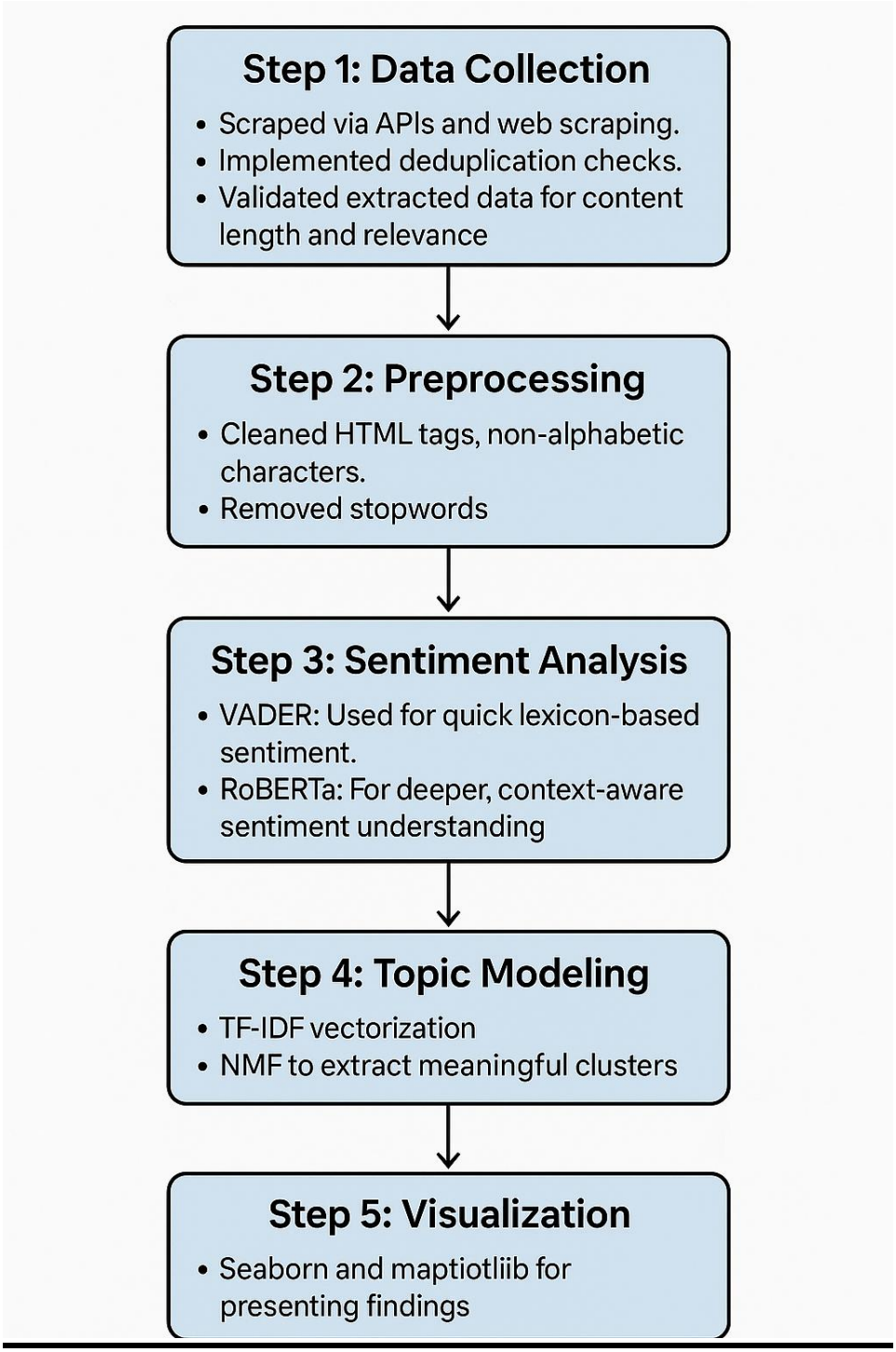
➢ Models:

- **VADER** (Lexicon-based)
- **RoBERTa** (Transformer-based via HuggingFace)

➢ APIs:

- YouTube Data API
- Reddit PRAW

METHODOLOGY



DATA COLLECTION

Platform	Type of Data	Reason for Selection
Reddit	Posts & Comments	Community-driven discussions covering diverse topics.
YouTube	Video Comments	Rich in informal user opinions, sentiment-heavy content.
Bing News	News Articles	Professional articles, global perspectives, informative tone.
Medium	Articles	Opinionated, long-form, structured articles.

Collection Methods

Reddit Data Collection

- Employed **PRAW (Python Reddit API Wrapper)** to gather posts and comments from relevant subreddits.
- Suggested subreddits based on keywords.
- Collected only **meaningful, non-duplicate content** with sufficient length.

YouTube Data Collection

- Used **YouTube Data API v3** to fetch video IDs and extract comments.
- Searched videos with **user-defined keywords** and ensured a **minimum number of comments** for quality.
- Handled API rate limits and removed duplicates.

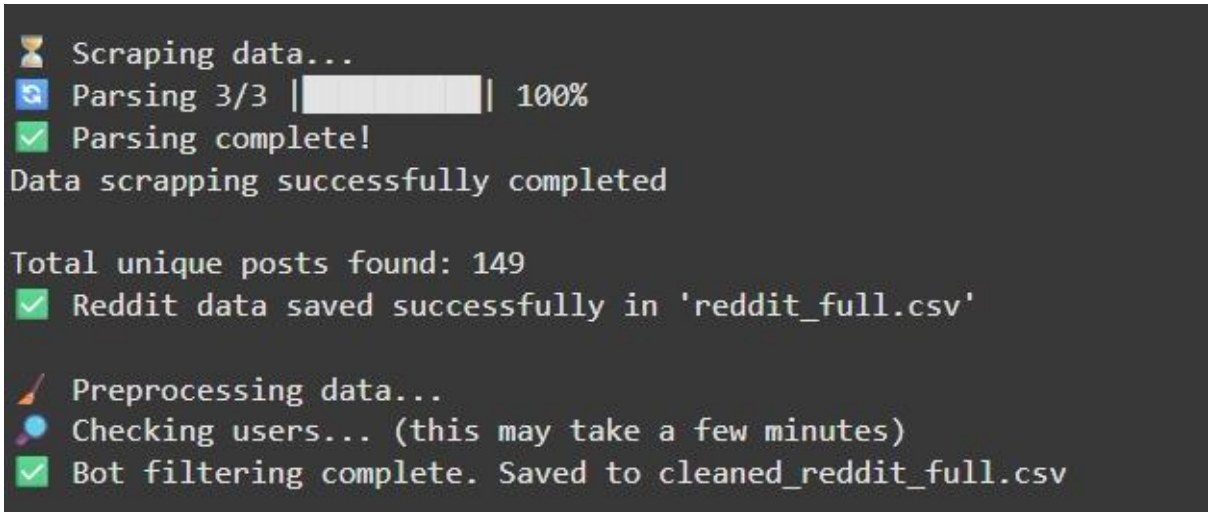
Bing News Data Collection

- Scraped articles via **Bing Search** and **BeautifulSoup**.
- Focused on **titles, URLs, and article text**.
- Deduplicated to avoid repetitions and ensured content validity.

Medium Data Collection

- Collected articles using **Google Custom Search Engine (CSE)** and **BeautifulSoup**.
- Extracted **titles, URLs, and content bodies**.
- Ensured articles were unique and of sufficient length.

Scraping and preprocessing working (Reddit)



Columns displaying the raw data (Text) and the pre-processed data (Final_Text).

Title	Text	Upvotes	Comments	Date	Author	URL	Cleaned_Title	Cleaned_Text	NoStop_Title	Final_Text
ing about the future of this country	So i was just thinking about what this country...	31	67	2025-07-13	aRandomGuy_xdd	https://www.reddit.com/r/india/comments/1lytfz...	thinking about the future of this country	so i was just thinking about what this country...	thinking future country	thinking country could amount futureeconomical...
India Slashes le Production maround Ti...	NaN	40	0	2025-07-09	kingterrytheterrific	https://www.reddit.com/r/india/comments/1lvoav...	india slashes missile production turnaround ti...	NaN	india slashes missile production turnaround ti...	NaN
king a school t app, and my lad says it's...	Hey, I'm a 15 year old, in class 11 CBSE, and ...	199	99	2025-07-06	Antinityfx	https://www.reddit.com/r/india/comments/1lss26...	making a school mgmt app and my dad says its n...	hey im a year old in class cbse and ive been d...	making school mgmt app dad says worth india im	hey im year old class cbse ive web development...
tuck Between Financial endence and My Pa...	I'm 24, working in IT (4 years now), recently ...	87	48	2025-06-21	Big-Entertainment605	https://www.reddit.com/r/india/comments/1lh09x...	stuck between financial independence and my pa...	im working in it years now recently switched f...	stuck financial independence parents neverendi...	im working years recently switched service pro...
ers Abroad in Aerospace vs tinue to wor...	Background:\n\nHave done my B. Tech in Mechani...	7	11	2025-06-28	WeirdVeterinarian629	https://www.reddit.com/r/india/comments/1lmi0c...	masters abroad in aerospace vs continue to wor...	background have done my b tech in mechanical e...	masters abroad aerospace vs continue work poli...	background done b tech mechanical engineering ...
...	...	...	...	...	...	...	...	...	...	...

SENTIMENT ANALYSIS

Sentiment analysis is a core component of this project, helping to transform collected textual data into actionable insights by determining the polarity (positive, negative, or neutral) of opinions across different platforms. This project employs **two distinct approaches for sentiment analysis**:

1. **Rule-Based Analysis (VADER)**
2. **Transformer-Based Deep Learning Model (RoBERTa)**

By combining both approaches, the project ensures comprehensive coverage of sentiment detection, balancing between speed and accuracy.

VADER Sentiment Analysis

Overview:

VADER (Valence Aware Dictionary and Sentiment Reasoner) is a **lexicon and rule-based sentiment analysis tool** specifically designed for analyzing sentiments expressed in social media content. VADER is known for its simplicity and efficiency in handling informal language, slangs, and emojis common in platforms like Reddit and YouTube.

Why VADER?

- ✔ Lightweight and fast.
- ✔ Well-suited for short texts (comments, posts).
- ✔ No GPU dependency; ideal for large datasets needing quick analysis.

Implementation Details:

- **Tool Used:** vaderSentiment library in Python.
- **Process:**
  - Cleaned and pre-processed text from all platforms.
  - Used SentimentIntensityAnalyzer to compute sentiment scores.
  - Classified texts as **Positive, Negative, or Neutral** based on the compound score thresholds.
- **Output:** Each comment/post/article was tagged with a sentiment label and score.

RoBERTa Sentiment Analysis

Overview:

RoBERTa (A Robustly Optimized BERT Pretraining Approach) is a **state-of-the-art transformer-based model** from HuggingFace designed for natural language understanding tasks, including sentiment analysis. Specifically, this project used cardiffnlp/twitter-roberta-base-sentiment, a model fine-tuned for social media sentiment tasks.

Why RoBERTa?

- ✔ More accurate and nuanced understanding of complex text.
- ✔ Better at detecting sarcasm, context, and varied sentence structures.
- ✔ Suitable for both short and long texts.

Implementation Details:

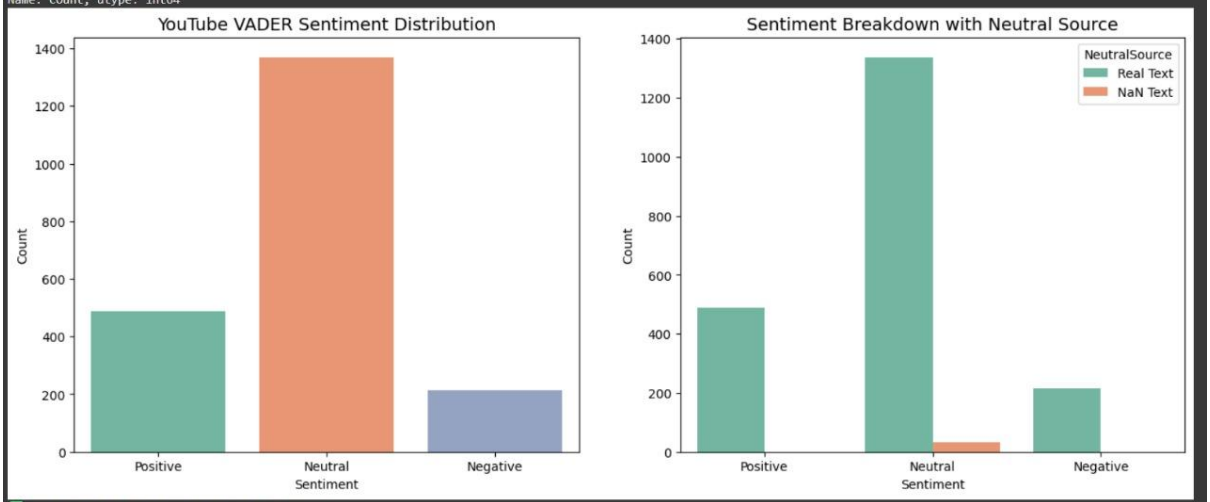
- **Tools Used:** transformers library, AutoTokenizer, and AutoModelForSequenceClassification.
- **Hardware:** Enabled GPU acceleration (cuda) to ensure faster inference in Google Colab.
- **Process:**
  - Tokenized texts with a maximum length of 512 tokens.
  - Inference performed using pre-trained RoBERTa sentiment model.
  - **Softmax function** applied to logits to obtain probability distributions.
  - Sentiment determined by the class with the highest probability: **Positive, Neutral, or Negative**.
- **Output:** Sentiment labels and confidence scores stored for each text sample.

Aspect	VADER	RoBERTa
Type	Rule-Based (Lexicon)	Deep Learning (Transformer)
Ideal For	Short social media texts	Context-rich texts of varying lengths
Speed	Very Fast	Slower (requires GPU for best performance)
Accuracy	Good for simple language	Superior for complex & nuanced language
Output	Compound score & label	Class label & probability scores

Sentiment Analysis performed on the same dataset scrapped

VADER Sentiment Analysis

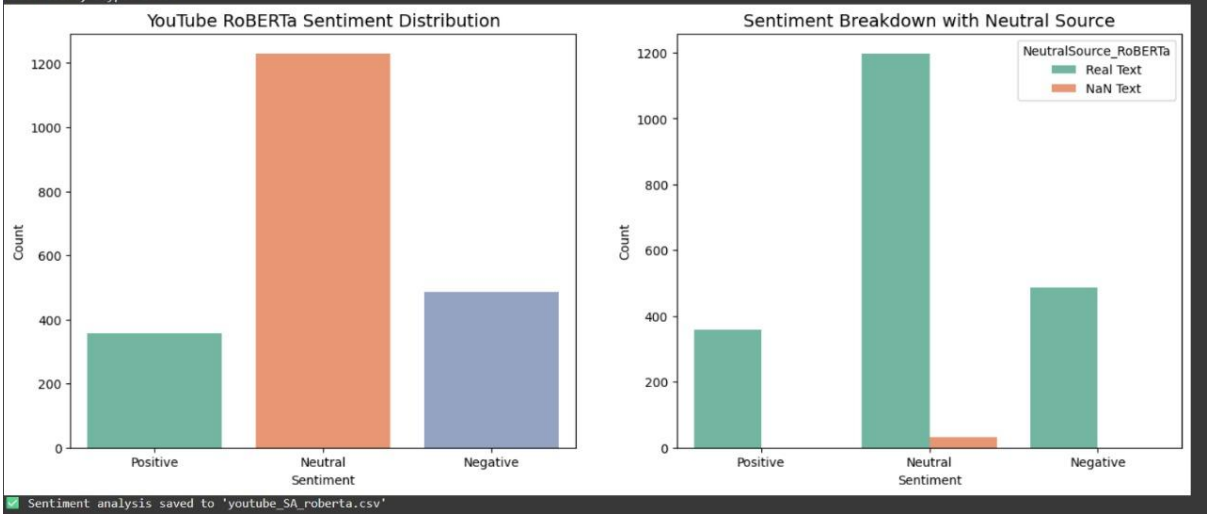
```
What would you like to do?
1 Sentiment Analysis
2 Topic Modeling
3 Sentiment vs Topics
4 Return to Home
5 Exit
Enter your choice: 1
Choose Sentiment Model (vader/roberta): vader

Sentiment
Positive     489
Negative     215
Neutral     1370
Name: count, dtype: int64
```




```
Sentiment analysis saved to 'youtube_SA_vader.csv'
```

RoBERTa Sentiment Analysis

```
Enter your choice: 1
Choose Sentiment Model (vader/roberta): roberta

Sentiment_RoBERTa
Positive     358
Negative     486
Neutral     1230
Name: count, dtype: int64
```




```
Sentiment analysis saved to 'youtube_SA_roberta.csv'
```

TOPIC MODELING

Overview:

Topic modeling was employed in this project to extract latent themes and topics from the large corpus of collected textual data. This unsupervised machine learning technique automatically identifies hidden semantic structures within the data and groups related words under coherent topics.

The goal was to understand **what people are commonly discussing** across platforms (YouTube, Reddit, Medium, Bing News) beyond just sentiment.

The project used **TF-IDF Vectorization** combined with **Non-Negative Matrix Factorization (NMF)** for topic modeling.

TF-IDF (Term Frequency-Inverse Document Frequency):

- Converts textual data into a numerical matrix, highlighting words that are **important in a specific document but less frequent globally**.
- Helps focus on meaningful and discriminative words for topic modeling.

NMF (Non-Negative Matrix Factorization):

- Decomposes the TF-IDF matrix into a predefined number of topics and associated keywords.
- Suitable for identifying **clear, non-overlapping topics** from sparse datasets like short comments and posts.

Implementation Steps:

1. **Preprocessing:**

- Text cleaning: Removed punctuations, special characters, URLs, and stopwords.
- Tokenized words and performed lemmatization.

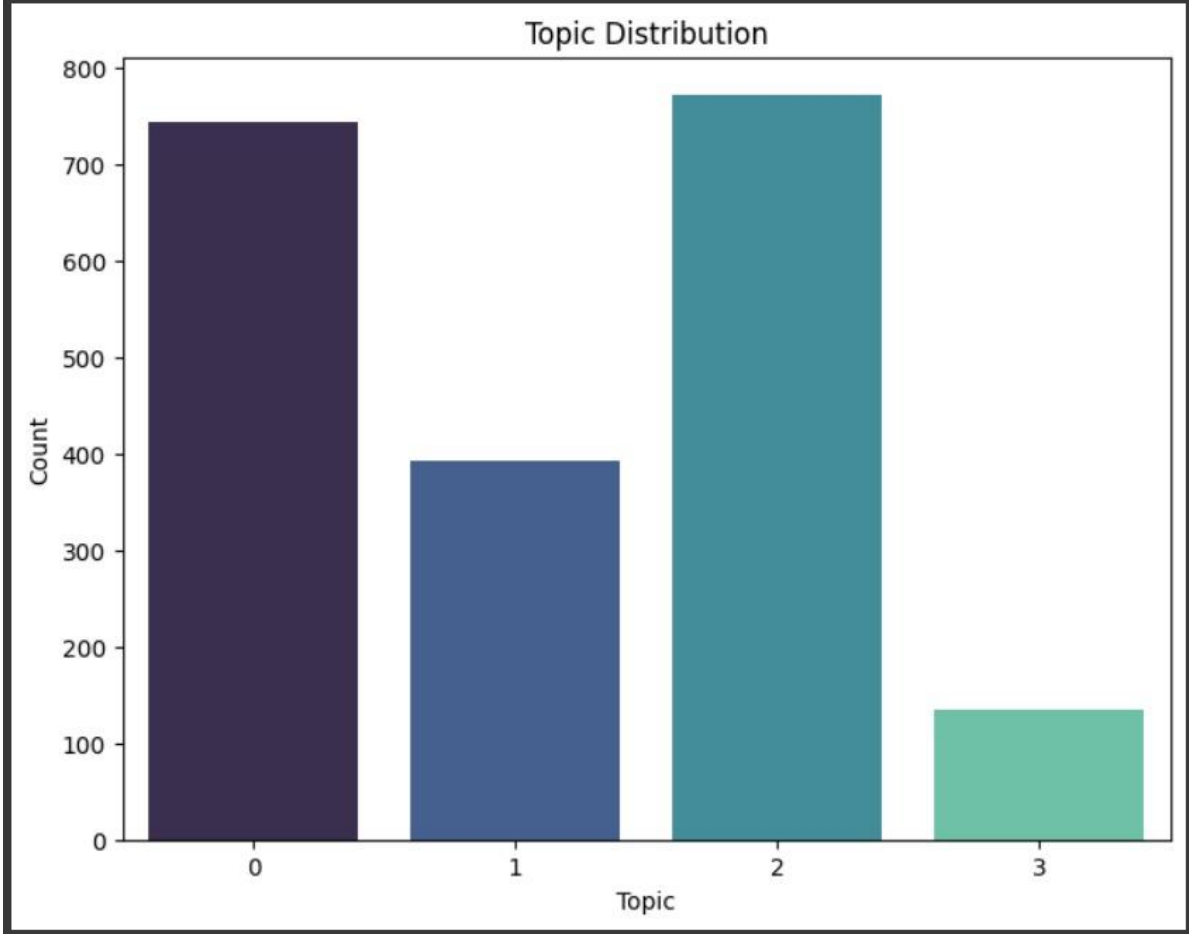
2. **Vectorization:**

- Applied TfidfVectorizer from scikit-learn to convert preprocessed text into a weighted numerical representation.
- Set parameters such as max_features and ngram_range to optimize performance.

3. **Modeling:**

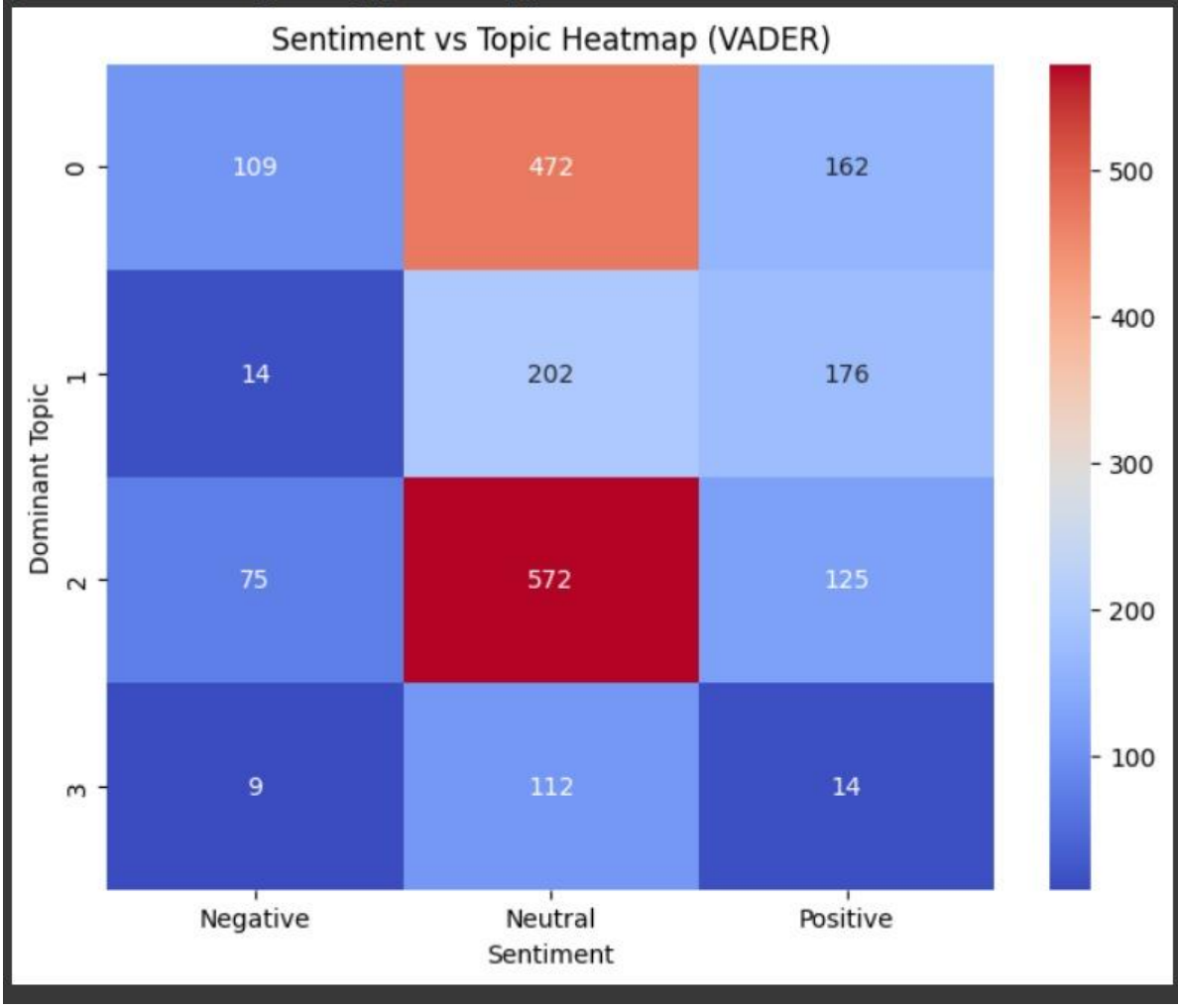
- Used NMF to extract topics and keywords.
- The number of topics was set dynamically based on user input (commonly 5-10 topics).
- Extracted **top 10 keywords per topic** to label and understand each theme.

```
🔹 Topic 0:
india, make, china, like, failed, world, tesla, need, people, great

🔹 Topic 1:
video, good, nice, great, job, excellent, sir, thanks, presentation, work

🔹 Topic 2:
hai, modi, manufacturing, assembling, se, china, indian, ke, bhi, ho

🔹 Topic 3:
map, correct, india, indian, france, use, dont, showing, pls, kashmir
```




```
Data saved to topic_modelling.csv
```

```
Sentiment
Neutral     1358
Positive     477
Negative     207
Name: count, dtype: int64


Sentiment  Negative  Neutral  Positive
Topic
0               109      472       162
1                14      202       176
2                75      572       125
3                 9      112        14
```



VISUALIZATION

❖ **Sentiment Analysis:**

- **Bar Graphs:** Displayed sentiment distribution (Positive, Neutral, Negative) across platforms.
- **Count Plots:** Compared sentiment categories by source (YouTube, Reddit, Medium, Bing News).

❖ **Topic Modeling:**

- **Bar Graphs:** Showed the frequency of each discovered topic.
- **Heatmaps:** Mapped the relationship between topics and sentiment categories, providing insights into how certain topics align with user sentiments.

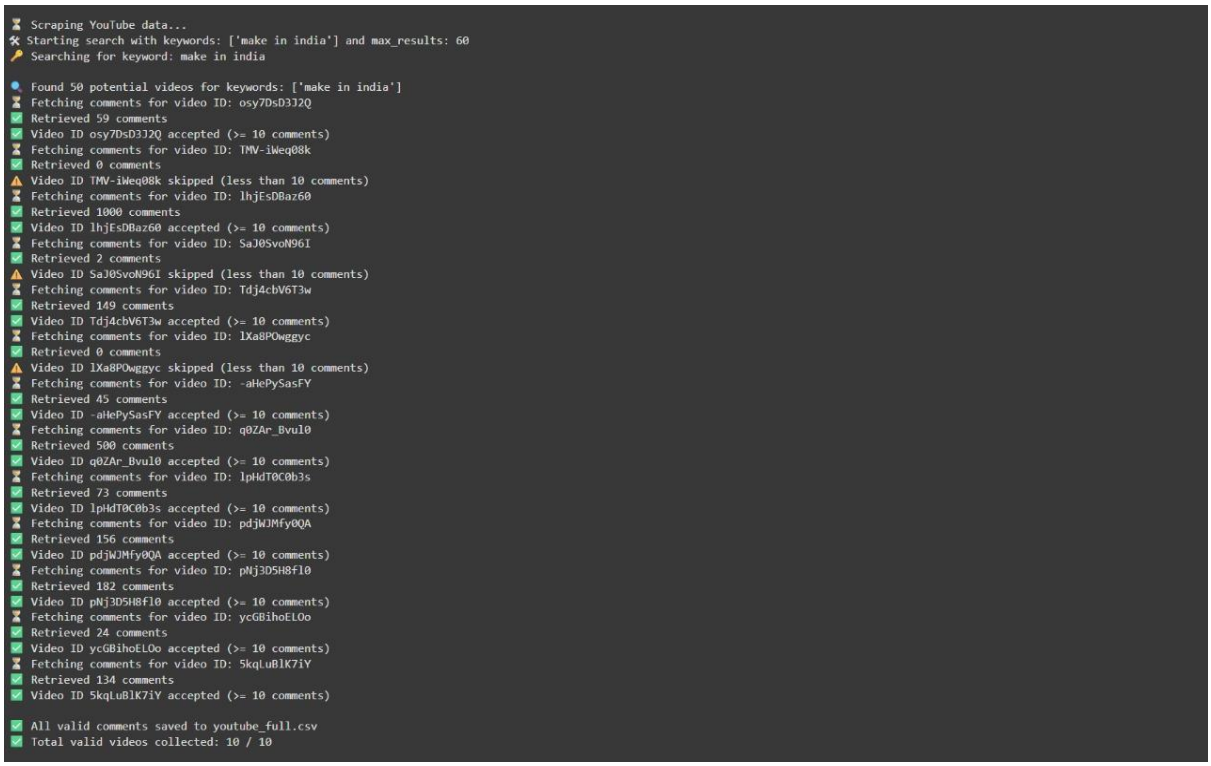
❖ **Tools Used:**

- **Matplotlib**
- **Seaborn**
- **TQDM** (Progress Visualization)

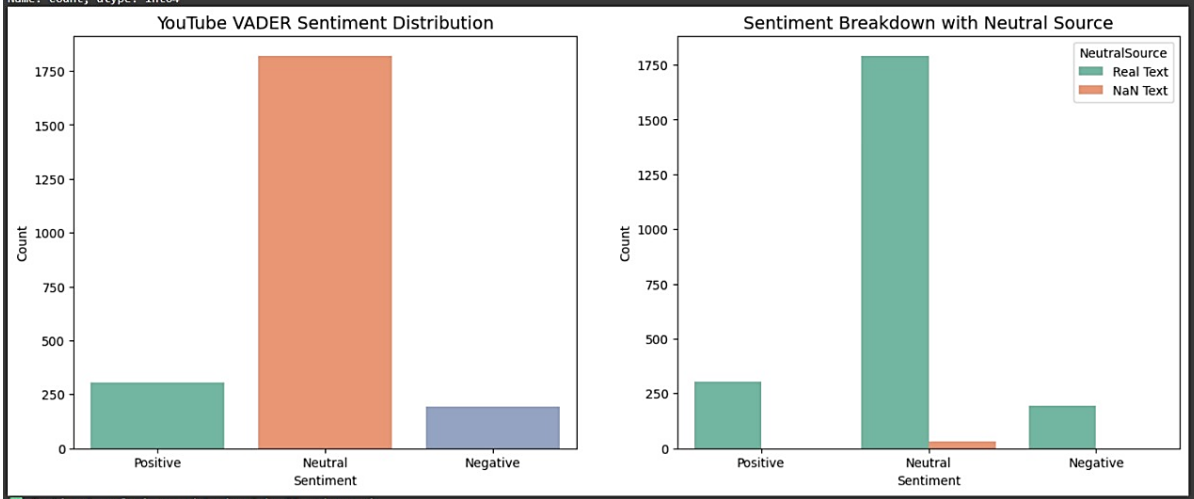
❖ **Purpose of Visualization:**

- Simplified complex analytical outcomes.
- Provided actionable insights through clear, engaging visuals.
- Enhanced interpretability for end users and stakeholders.

Sample Output – Full Pipeline run (YouTube)

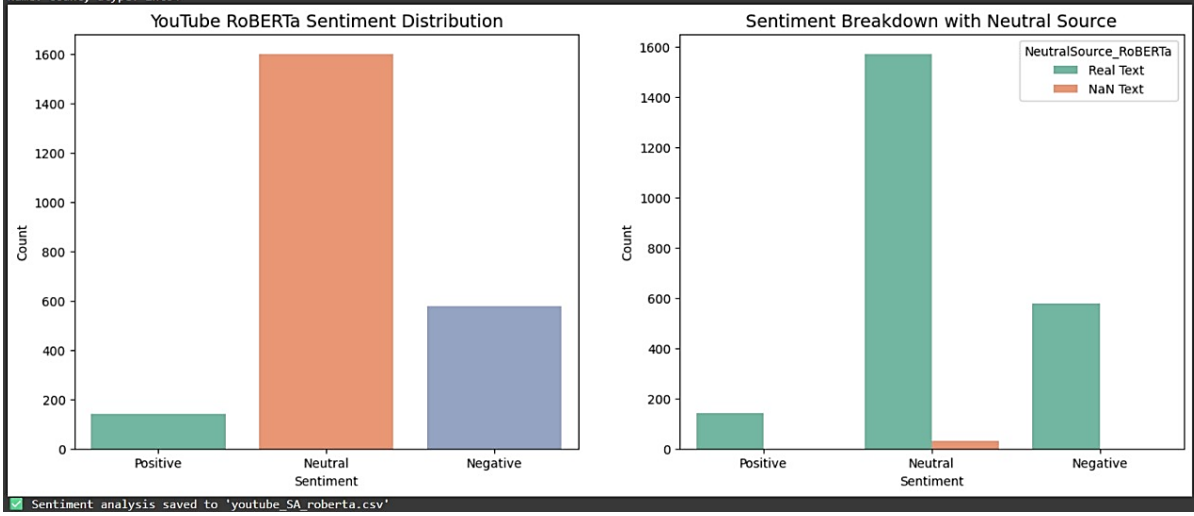


```
What would you like to do?
1 Sentiment Analysis
2 Topic Modeling
3 Sentiment vs Topics
4 Return to Home
5 Exit
Enter your choice: 1
Choose Sentiment Model (vader/roberta): vader

Sentiment
Positive     305
Negative     195
Neutral     1822
Name: count, dtype: int64
```



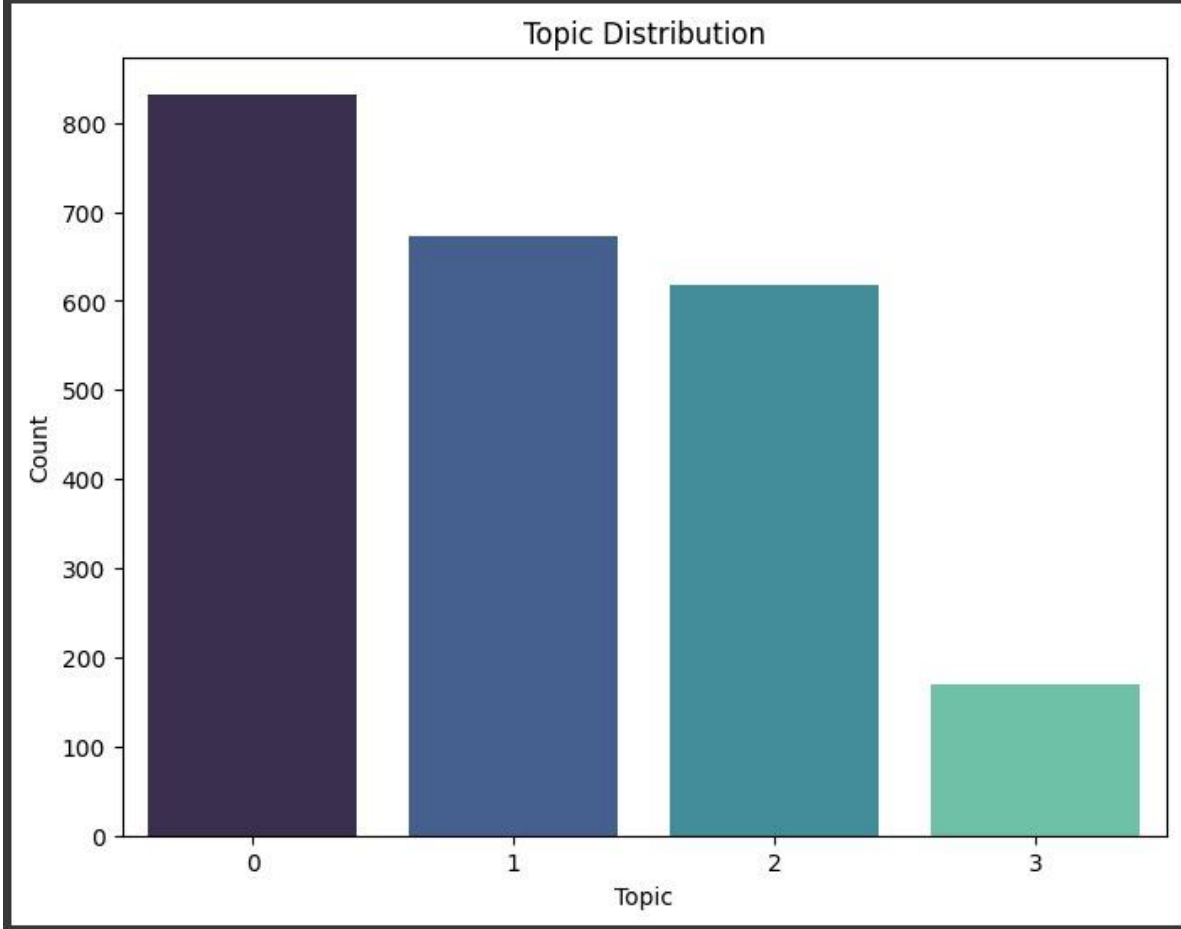

```
✅ Sentiment analysis saved to 'youtube_SA_vader.csv'
```

```
What would you like to do?
1 Sentiment Analysis
2 Topic Modeling
3 Sentiment vs Topics
4 Return to Home
5 Exit
Enter your choice: 1
Choose Sentiment Model (vader/roberta): roberta

Sentiment_RoBERTa
Positive     142
Negative     578
Neutral     1602
Name: count, dtype: int64
```




```
✅ Sentiment analysis saved to 'youtube_SA_roberta.csv'
```

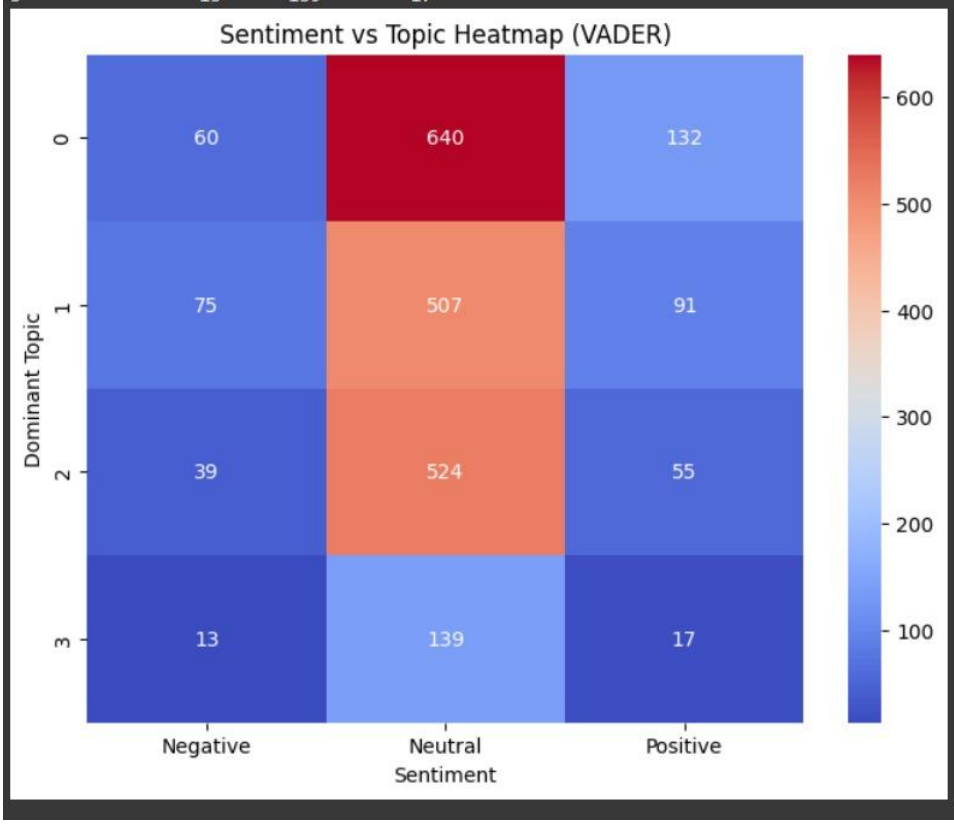
```
What would you like to do?
1️⃣ Sentiment Analysis
2️⃣ Topic Modeling
3️⃣ Sentiment vs Topics
4️⃣ Return to Home
5️⃣ Exit
Enter your choice: 2
Enter number of topics : 4


🔹 Topic 0:
india, make, china, dont, hygiene, tesla, world, manufacturing, cheap, people

🔹 Topic 1:
ice, dirty, cheap, water, floor, drink, dont, food, like, use

🔹 Topic 2:
hai, china, ki, ko, se, aur, ka, modi, bhi, ke

🔹 Topic 3:
rusty, steel, molds, copper, lol, thats, moulds, rust, arent, like
```




```
Data saved to topic_modelling.csv
```

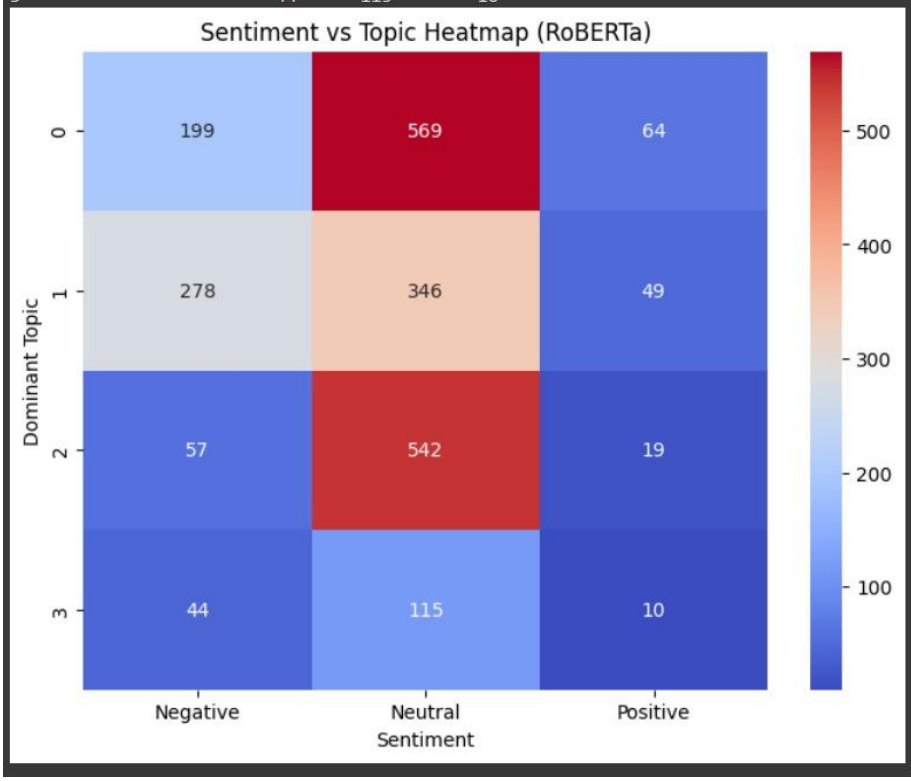
```
What would you like to do?
1 Sentiment Analysis
2 Topic Modeling
3 Sentiment vs Topics
4 Return to Home
5 Exit
Enter your choice: 3
Enter model type : vader
Enter number of topics : 4


🔹 Topic 0:
india, make, china, dont, hygiene, tesla, world, manufacturing, cheap, people

🔹 Topic 1:
ice, dirty, cheap, water, floor, drink, dont, food, like, use

🔹 Topic 2:
hai, china, ki, ko, se, aur, ka, modi, bhi, ke

🔹 Topic 3:
rusty, steel, molds, copper, lol, thats, moulds, rust, arent, like
```

```
Sentiment
Neutral     1810
Positive     295
Negative     187
Name: count, dtype: int64


Sentiment  Negative  Neutral  Positive
Topic
0                60      640       132
1                75      507        91
2                39      524        55
3                13      139        17
```



```
What would you like to do?
 1 Sentiment Analysis
 2 Topic Modeling
 3 Sentiment vs Topics
 4 Return to Home
 5 Exit
Enter your choice: 3
Enter model type : roberta
Enter number of topics : 4


 🔹 Topic 0:
india, make, china, dont, hygiene, tesla, world, manufacturing, cheap, people

 🔹 Topic 1:
ice, dirty, cheap, water, floor, drink, dont, food, like, use

 🔹 Topic 2:
hai, china, ki, ko, se, aur, ka, modi, bhi, ke

 🔹 Topic 3:
rusty, steel, molds, copper, lol, thats, moulds, rust, arent, like
```

```
Sentiment_RoBERTa  Negative  Neutral  Positive
Topic
0                         199      569        64
1                         278      346        49
2                          57      542        19
3                          44      115        10
```




```
What would you like to do?
 1 Sentiment Analysis
 2 Topic Modeling
 3 Sentiment vs Topics
 4 Return to Home
 5 Exit
Enter your choice: 5

 👋 Exiting the chatbot. Thank you!


**********************************************Session Terminated**********************************************
```

RESULTS AND OBSERVATIONS

Summary of Sentiment Analysis Results:

Sentiment analysis was performed on data from **YouTube, Reddit, Medium, and Bing News** using **VADER** and **RoBERTa** models. Both provided unique insights, highlighting the difference between lexicon-based and deep learning-based approaches.

Insights from VADER Analysis:

- **VADER (Valence Aware Dictionary and sEntiment Reasoner)** is a **lexicon and rule-based model** specifically attuned to sentiments expressed in social media and informal texts.
- **Key Observations:**
  - Tends to label **neutral or mildly positive tones more favorably**.
  - Recognizes emojis, slang, and common intensifiers (e.g., "very good", "super bad").
  - Sometimes misclassifies **long, complex sentences** or **sarcasm** as neutral or even positive.

Insights from RoBERTa Analysis:

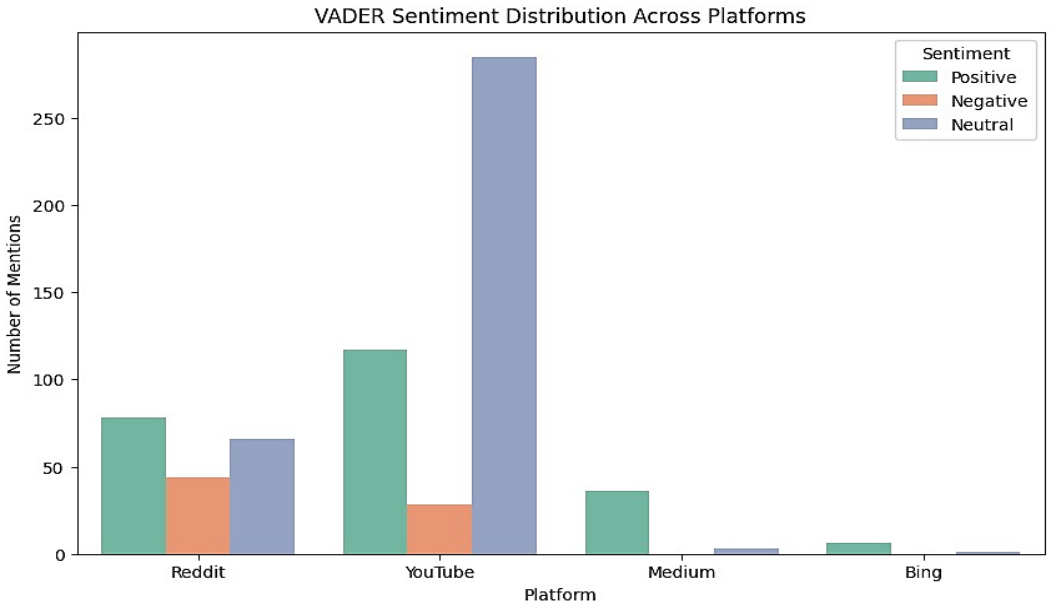
- **RoBERTa (Robustly Optimized BERT Pretraining Approach)** is a **deep learning transformer model** fine-tuned for sentiment tasks.
- **Key Observations:**
  - Evaluates sentiment based on **contextual depth**, detecting nuances and hidden tones.
  - More effective at picking up **subtle negativity or indirect sentiment expressions**.
  - Handles long texts, sarcasm, and varied sentence structures more robustly.

Aspect	VADER	RoBERTa
Strength	Fast, lightweight, social-media friendly	Deep understanding, context-aware
Weakness	Struggles with complex texts	Requires more computation, slower
Bias Tendency	Slightly optimistic	Balanced, realistic
Neutral Detection	High	Balanced

CONCLUSION

This project successfully developed a **robust and modular pipeline** for:

- ✓ Collecting data from multiple platforms (YouTube, Reddit, Medium, Bing News).
- ✓ Efficient data cleaning and structuring for accurate sentiment analysis.
- ✓ Performing **sentiment classification** using both **VADER** (lexicon-based) and **RoBERTa** (transformer-based) models for broader perspective.
- ✓ Extracting meaningful **topics** through **NMF topic modeling**.
  Providing **clear, visual insights** through well-crafted graphs and breakdowns.



FUTURE SCOPE

- ➢ Integrate **real-time dashboards** for continuous monitoring of sentiments and topics.
- ➢ **Deploy as a user-friendly web application** for wider accessibility.
- ➢ Expand analysis by incorporating **emotion detection** (joy, anger, fear) for deeper understanding of public opinion.
- ➢ Incorporate trend analysis over time to observe how sentiments evolve.
- ➢ Expand data sources to include forums, blogs, and news APIs for richer insights.

REFERENCES

- **YouTube Data API v3 Documentation**
  https://developers.google.com/youtube/v3
- **PRAW: The Python Reddit API Wrapper**
  https://praw.readthedocs.io/
- **Google Custom Search JSON API Documentation**
  https://developers.google.com/custom-search/v1/overview
- **Bing News Search**
  https://www.bing.com/news
- **BeautifulSoup: Documentation**
  https://www.crummy.com/software/BeautifulSoup/bs4/doc/
- **Transformers Library (HuggingFace)**
  https://huggingface.co/transformers/
- **Cardiff NLP Twitter RoBERTa Base Model**
  https://huggingface.co/cardiffnlp/twitter-roberta-base-sentiment
- **VADER Sentiment Analysis**
  Hutto, C.J. & Gilbert, E.E. (2014). VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. *Proceedings of the Eighth International Conference on Weblogs and Social Media (ICWSM-14)*.
- **Scikit-learn Documentation: NMF for Topic Modeling**
  https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.NMF.html