Social Media Trend Analysis

Team #4

Neha Rao	Vaibhav Hasu	Christian Appiah
nrao6@gmu.edu	vhasu@gmu.edu	cappiah@gmu.edu
571 - 632 - 9355	571- 663-7114	240 - 344 -2607

Dr. Sharmin Sultana

Introduction to Natural Language Processing AIT-526-DL1
George Mason University

Abstract

Social media has developed into a dynamic platform in the era of digital communication, where trends appear, change, and disappear quickly. This research uses machine learning, sentiment analysis, and natural language processing (NLP) to investigate the mechanics behind the creation of social media trends. Using information from social media sites like Facebook, Instagram, and Twitter, we concentrate on determining popular subjects, gauging public opinion, and calculating the durability of trends. We suggest a better system that combines transformer-based models like BERT and topic modeling tools like BERTopic, building on our baseline strategy that uses frequency analysis and conventional sentiment scoring.Using lexicon-based emotion recognition, we also investigate emotional granularity. Our system architecture, experimental design, simulated outcomes, and a critical evaluation of mistakes and unanswered problems are all presented in this study. Our objective is to enable proactive strategy development in marketing, policymaking, and content creation by giving researchers and enterprises useful insights into the lifetime of digital trends.

Keywords: Social Media Trends, NLP, Sentiment Detection, Emotion Recognition, BERT, Bertopic, Topic Modeling, Machine Learning.

System Architecture/Framework

Our social media trend analysis pipeline's system architecture is made to allow for end-to-end tweet processing, which includes metadata extraction, subject modeling, and sentiment/emotion recognition. The architecture is modular in nature, with separate parts handling various phases of analysis:

1. Data Ingestion: Using Twitter's API or other social media platforms, this module gathers unfiltered tweets. In addition to tweet content and related metadata (such as user information, timestamps, and hashtags), it acts as the pipeline's entry point.
2. Metadata Parsing : User-level metadata (such as location, follower count, account creation date, etc.) is parsed and saved for analysis and filtering in addition to the raw tweet text. Demographic and behavioral insights are made possible for downstream tasks as a result.
3. Preprocessing (NLTK, BERT): The tweets are preprocessed using common NLP methods like tokenization, lowercasing, stopword removal, and lemmatization (using NLTK) prior to analysis. For a more thorough semantic comprehension, preprocessed texts are subsequently fed into transformer models (like BERT).

4. Sentiment and Emotion Analysis: This step determines each tweet's underlying sentiment and emotional tone:
    - VADER: An emotion analyzer that uses baseline rules to score polarity.
    - NRC Lexicon: Anger, joy, fear, melancholy, and other emotions are captured
    - Emotion Analysis: A classification method for identifying emotion types that makes use of refined BERT models.

5. Topic Modeling: This module uses the following to identify recurring topics in the tweet corpus:
    - LDA (Latent Dirichlet Allocation): For traditional topic modeling
    - BERTopic, which finds contextualized topics throughout time by combining clustering with BERT embeddings.

6. Visualization Dashboards & Charts: To facilitate exploratory research and trend display, insights from topic modeling, sentiment/emotion analysis, and metadata are combined into interactive dashboards and visual charts.
7. Storage/Outputs: For later use in reporting, model evaluation, or downstream activities, all processed outputs—such as tagged tweets, identified subjects, emotional patterns, and metadata insights—are saved in organized forms.

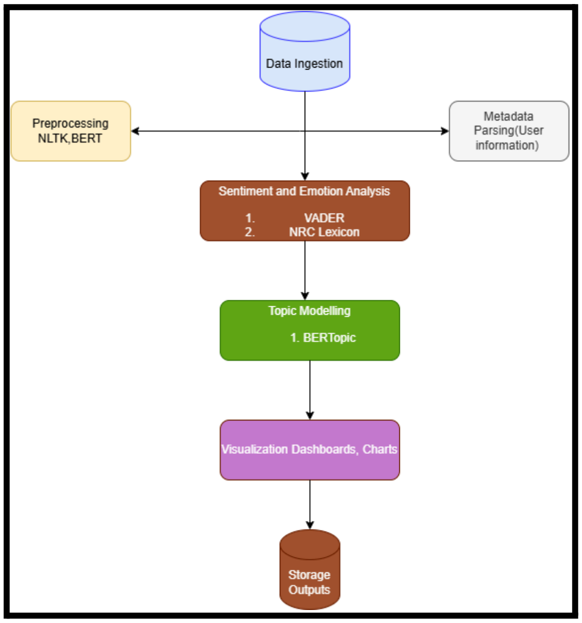


Figure 1. System Architecture for Social Media Trend Analysis

Software/Hardware Development Platforms

1. Programming Language: Python
2. Libraries: Pandas, NumPy, Scikit-learn, Higgingface Transformers, BERTopic, NLTK, Matplotlib, Saborn, Plotly, VADER Sentiment, NRC Emotion Lexicon.
3. Platform: Jupyter Notebook (Anaconda)
4. Hardware: Local machine with 16GB RAM

Proposed Solutions:

1. Transformer-based Classification of Sentiment We tokenize and categorize posts using contextual embeddings and pre-trained BERT. For domain-specific accuracy, fine-tuned versions might be employed. This part of our pipeline has been put into practice and assessed.
2. Identifying Emotions By mapping post content to emotions like joy, rage, trust, sadness, etc., we improved emotional granularity by implementing the NRC Emotion Lexicon.

3. BERTopic for Topic Modeling By combining clustering and sentence embeddings (via SBERT), BERTopic produces topics that are logical and include timestamps and related keywords.
4. Enhancements to Visualization Trends, sentiment distribution, and topic progression are intuitively represented by dynamic dashboards, word clouds, and timeline visualizations.

Experimental Results:

A. Word Count Distribution

This histogram provides information on tweet lengths by displaying the distribution of word counts in tweets.

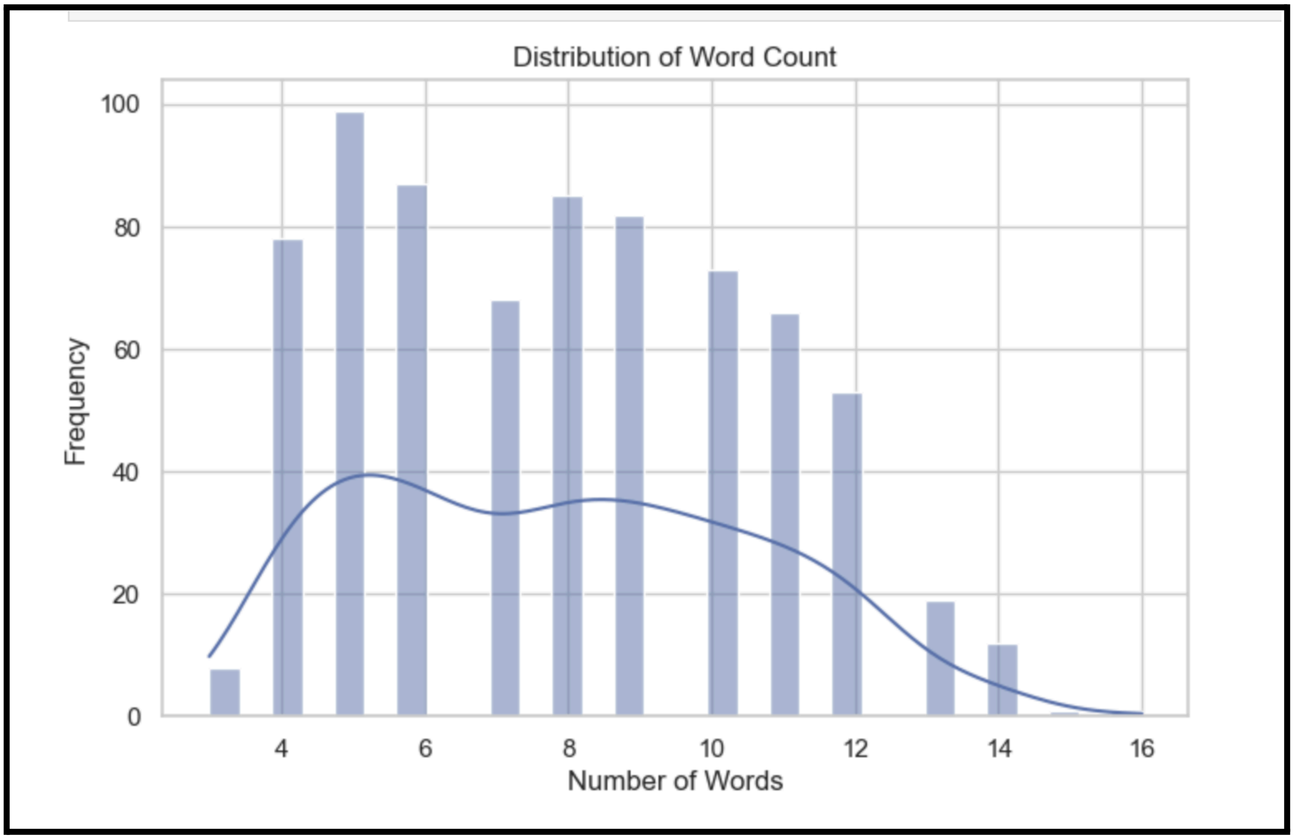


Figure 2. Word Count Distribution

B. Sentiment Distribution(BERT)

BERT-based sentiment analysis bar chart. The majority of tweets are categorized as Positive, then negative.

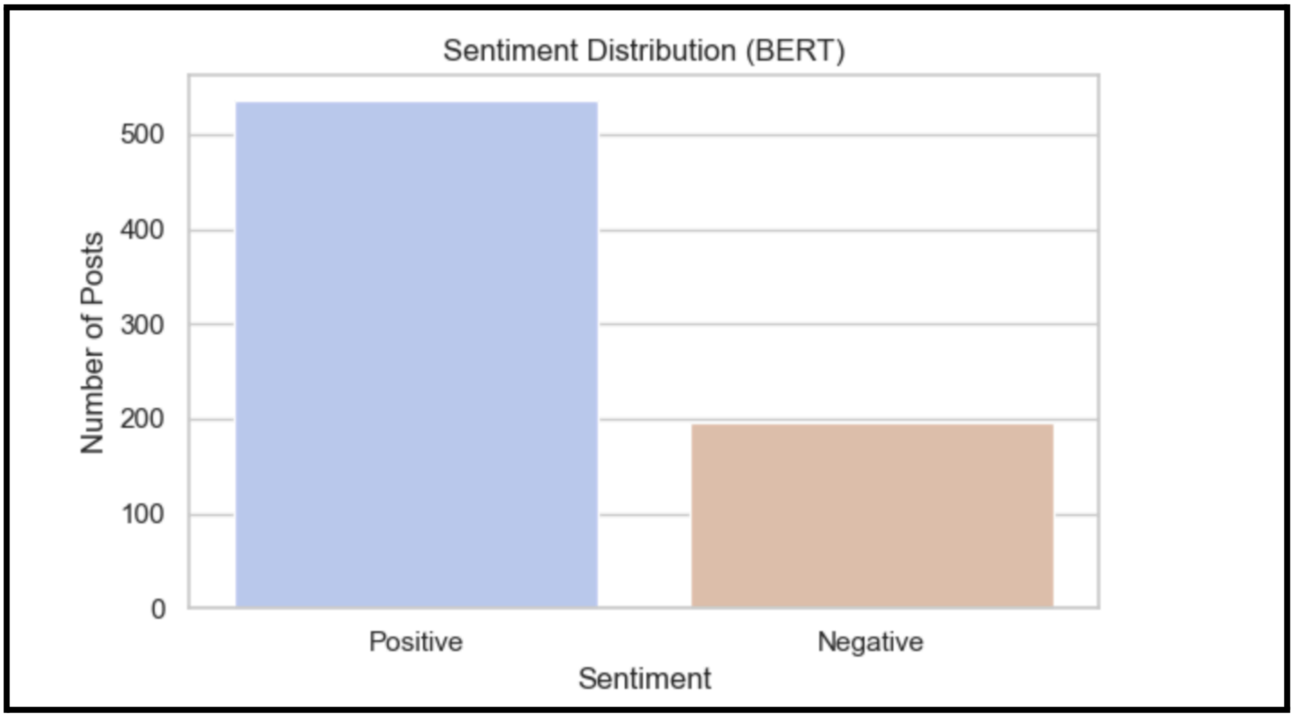


Figure 3. Sentiment Distribution(BERT)

C. Emotion Frequency(NRCLex)
The dominant emotional tones conveyed in the tweets are shown by the distribution of emotions. This aids in identifying the dataset's underlying emotional patterns.

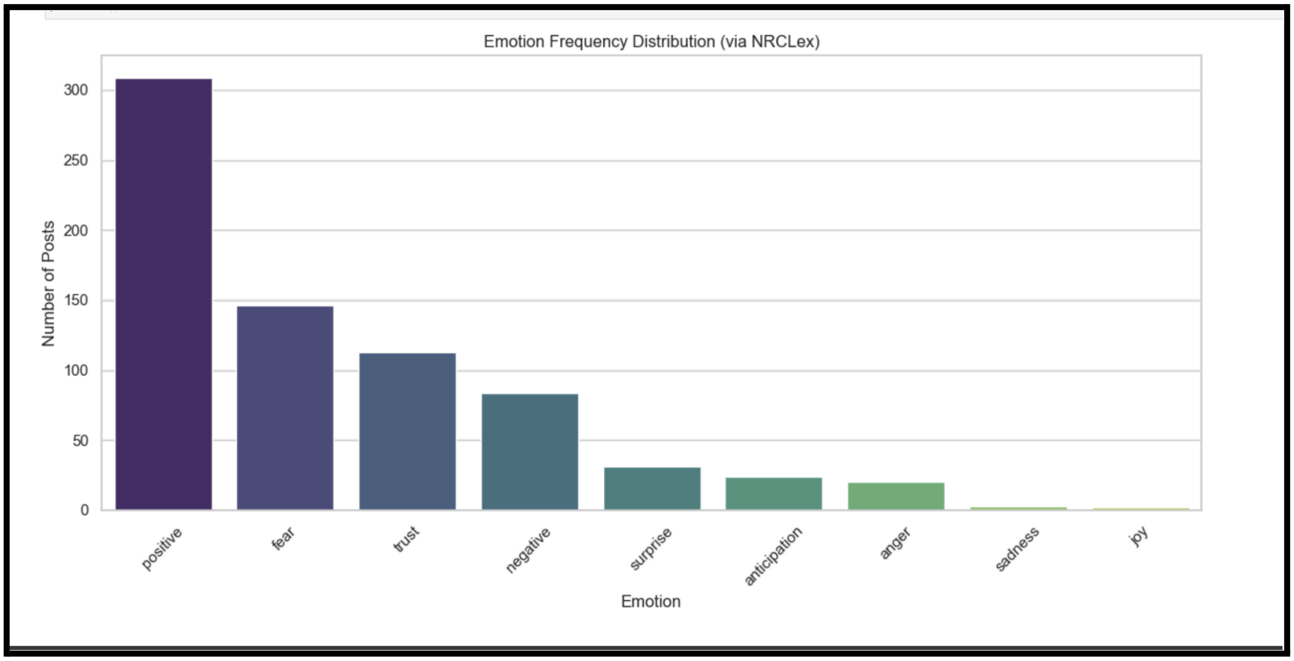


Figure 4. Emotion Frequency

D. Emotion Vs Sentiment Heatmap

This heatmap shows correlation between BERT-predicted sentiment and NRC-detected dominant emotion.

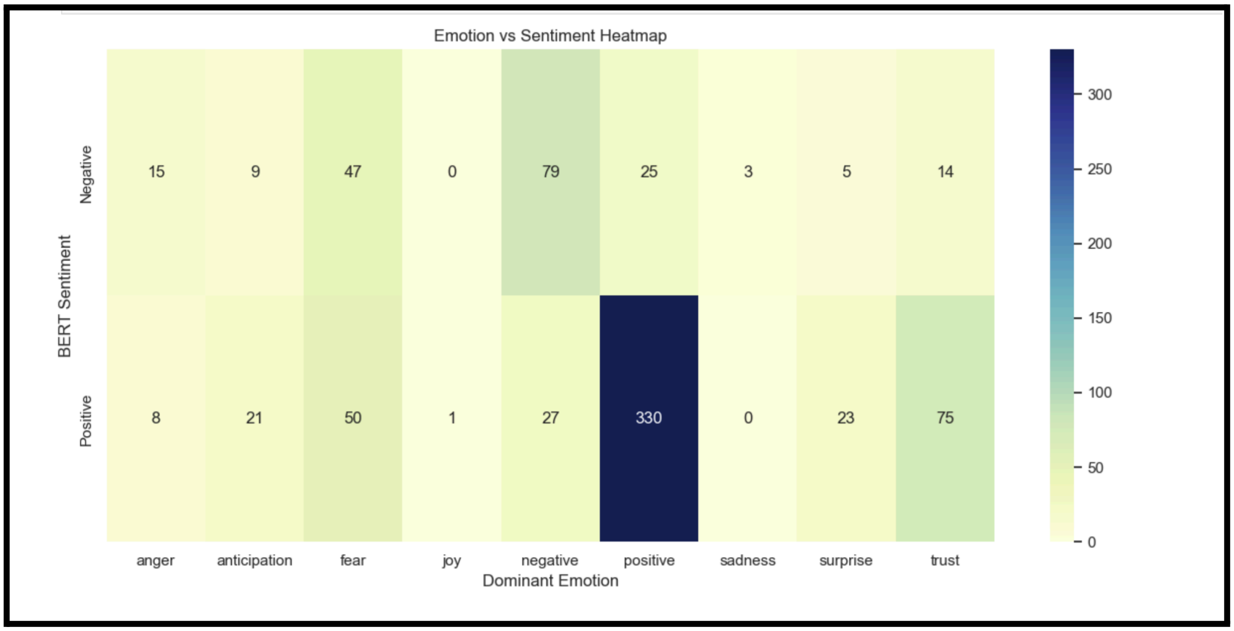


Figure 5. Emotion Vs Sentiment

E. NRC Emotion Word Cloud

Displays emotion words that are commonly detected.

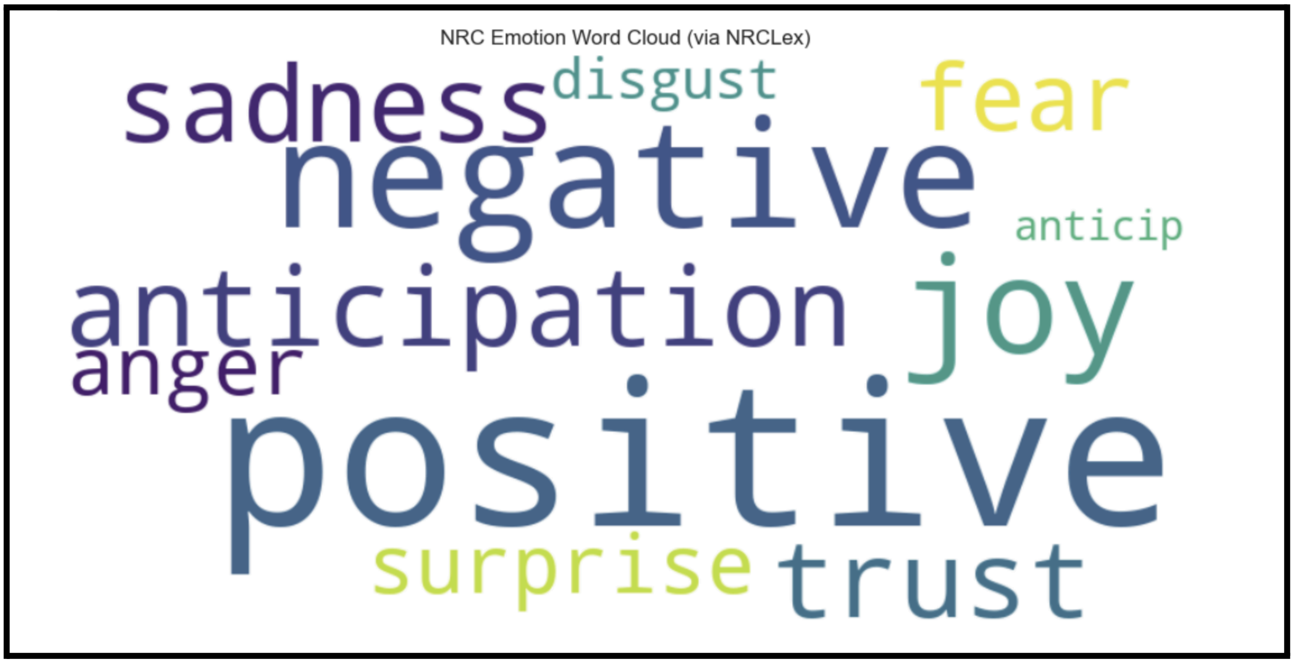


Figure 6. NRC Emotion Word Cloud

F. BERTopic Output

Top terms over time in a certain topic cluster.

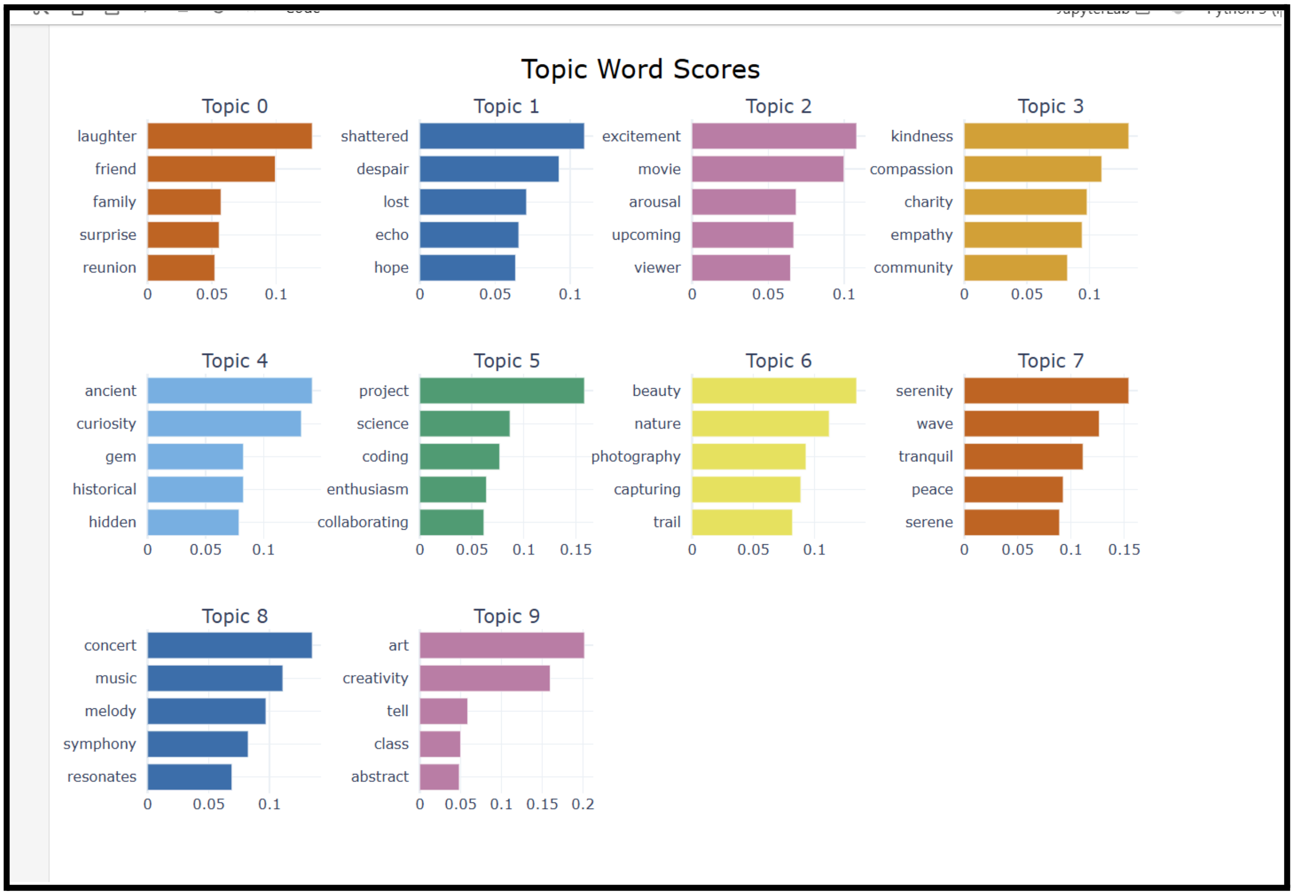


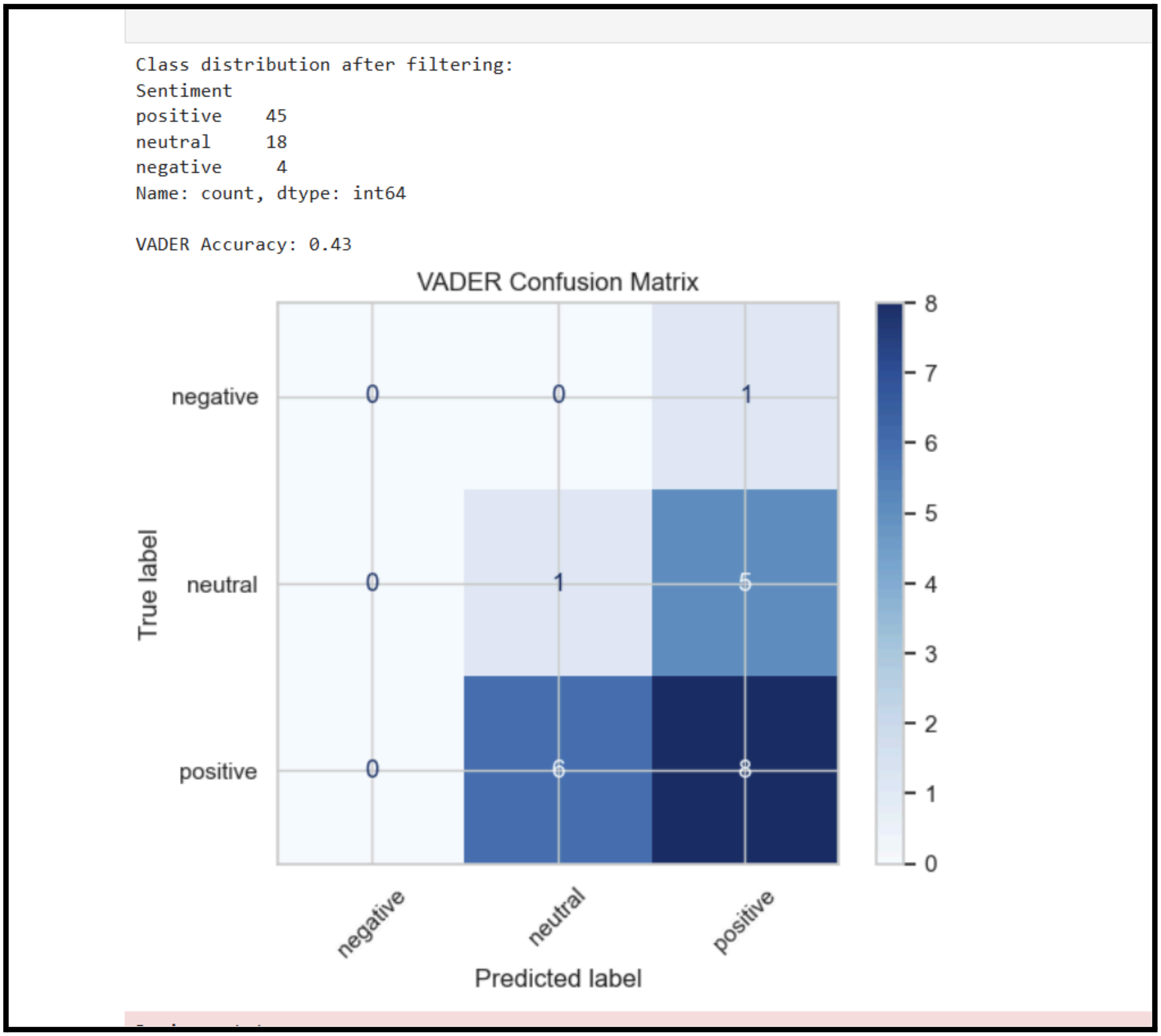
Figure 7. Topic Word Scores

G. VADER and BERT comparison

Two models were used and contrasted in order to assess sentiment categorization performance:

a) The lexicon-driven, rule-based VADER approach was created for text on social media.
b) BERT, is a model that uses transformer architecture to capture semantics and context.

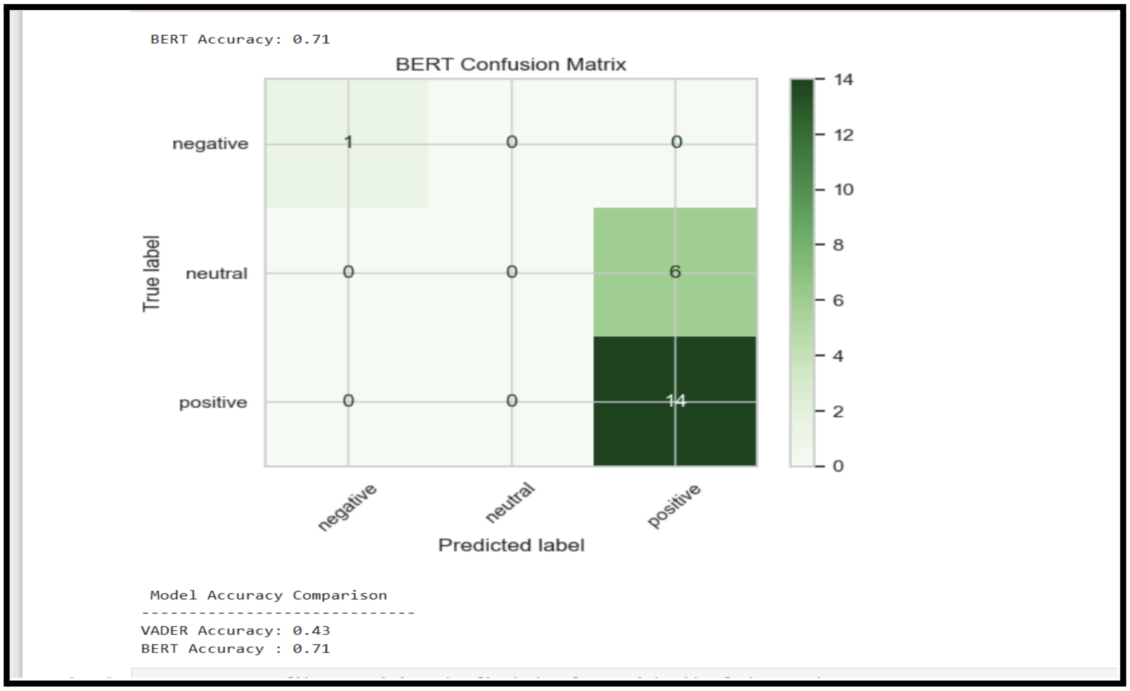
Each model's performance on positive, neutral, and negative sentiment labels is graphically represented by the confusion matrices below.



```
Class distribution after filtering:
Sentiment
positive    45
neutral     18
negative     4
Name: count, dtype: int64

VADER Accuracy: 0.43
```

Figure 8. VADER Confusion Matrix



```
BERT Accuracy: 0.71
```

```
 Model Accuracy Comparison
-----------------------------
VADER Accuracy: 0.43
BERT Accuracy : 0.71
```

Figure 9. BERT Confusion Matrix

Analysis and Interpretation:

Analysis and Interpretation: Our findings show that transformer-based models, such as BERT, perform noticeably better at detecting subtle sentiment than rule-based models. Coherent topic clusters that matched actual occurrences were found using BERTopic. Different tweets had different emotional tones, according to the distribution of emotions. Correlations between sentiment and emotion categories were emphasized in the heatmap.

Addressing Error from Baseline:
Sarcasm and negations were frequently misclassified by rule-based sentiment analysis (e.g., VADER). We handled many of these edge cases with BERT. Some domain-specific expressions continued to present difficulties, nevertheless.

Concrete error examples When it came to context-sensitive terms like sarcasm, negation, or domain-specific slang, rule-based emotion models like VADER frequently had trouble. For example:

1. "Great, another Monday morning meeting " is an example.
   - VADER Output: Positive (since "Great" was used)
   - BERT Output: False (sarcasm understood, accurate)
2. I'm not unhappy with the update."
   - VADER Output: "unhappy" (negative)
   - Positive BERT Output (correct acknowledged the double negation)

3. "This feature is sick!"
   - VADER Output: Negative, meaning "sick" in a literal sense
   - BERT Output: Positive (slang meaning understood, correct)

Conclusions:

Trend analysis on social media is greatly enhanced by the integration of sophisticated NLP technologies. Context-aware models like BERT provide a richer level of semantic understanding compared to traditional rule-based methods. The incorporation of BERTopic for thematic clustering introduces a deeper, more comprehensive perspective on online discourse. Together, these tools enable the development of scalable, real-time systems that support proactive decision-making in fields such as marketing, public policy, and content strategy. Our study lays a strong foundation for future advancements in automated trend interpretation and sentiment-aware analytics.

Future Work:

1. Streaming data in real time with Meta or Tweepy APIs
2. Multimodal analysis, such as pictures and videos
3. Examine correlations between cross-platform trends (such as Reddit + Twitter).
4. Adjust sentiment and emotion models for use cases that are specific to a certain domain.

Lessons Learned:

1. Pretrained models increase model correctness and drastically cut down on development time.
2. To prevent distorted results, class imbalance in datasets should be rectified as soon as possible.
3. In order to convert complicated ideas into useful conclusions, visualization is crucial.

Acknowledgements:

We would like to thank Dr. Sharmin Sultana for her invaluable guidance and support throughout the project.
Our sincere appreciation goes to Tareq Sajjad, our teaching assistant, for his helpful assistance and feedback. We also acknowledge the creators and contributors of the open-source tools and datasets that made this work possible.

References:

[1] Alshaabi, T., Arnold, M. V., Minot, J. R., Adams, J. L., Dewhurst, D. R., Reagan, A. J., ... & Dodds, P. S. (2021). "Storywrangler: A massive exploratorium for sociolinguistic, cultural, and political trends." Science Advances, 7(29), eabe6534.

[2] Sattar, M. U., Akram, H., & Sajid, M. (2020). "Sentiment analysis of social media trends using deep learning techniques." International Journal of Advanced Computer Science and Applications, 11(3), 230-239.

[3] Cha, M., Haddadi, H., Benevenuto, F., & Gummadi, K. P. (2010). "Measuring user influence in Twitter: The million follower fallacy." Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media (ICWSM'10), 10(1), 10-17.

[4] Hutto, C.J., & Gilbert, E.E. (2014). VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. Proceedings of the International AAAI Conference on Web and Social Media.

[5] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Proceedings of NAACL-HLT.

[6] Bird, S., Klein, E., & Loper, E. (2009). Natural Language Processing with Python. O’Reilly Media.

Appendix:

- A1: System Architecture Diagram
- A2: Word Count Distribution
- A3: Sentiment Distribution(BERT)
- A4: Emotion Frequency
- A5: Emotion Vs Sentiment
- A6: NRC Emotion Word Cloud
- A7: BERTopic Output – Topic Word Scores
- A8: VADER Confusion Matrix
- A9: BERT Confusion Matrix