Tweet Evaluation

Harsh Sharma, MS Data Science, First year
Harshika Santoshi, MS Data Science, Second year
Swathi Arasu, MS Data Science, First year

Introduction

Importance of Twitter Data:

1. Real-time insights into public sentiment and trends.
2. Essential for tracking brand perception and customer feedback.
3. Supports crisis management by identifying negative sentiments early.

Business Relevance:

1. Helps businesses respond promptly to customer grievances.
2. Enables data-driven decision-making with quantifiable metrics.

Business Problem

Challenges:

1. Manual analysis is time-consuming and inconsistent.
2. Need for automated models to classify sentiment efficiently.

Data Characteristics:

1. Handle unique Twitter features like colloquial language, emojis, and sarcasm.
2. Address class imbalance and missing information issues.

Objectives:

1. Develop robust models to process large volumes of unstructured data efficiently.
2. Extract meaningful insights such as sentiment, major topics and engagement patterns.

Data Overview

Dataset Source:

1. Two datasets from Kaggle with tweets labeled as positive, neutral, or negative.

Data Components:

1. Includes tweet text, no. of likes, comments and retweets a post received, hashtags, mentions, etc.

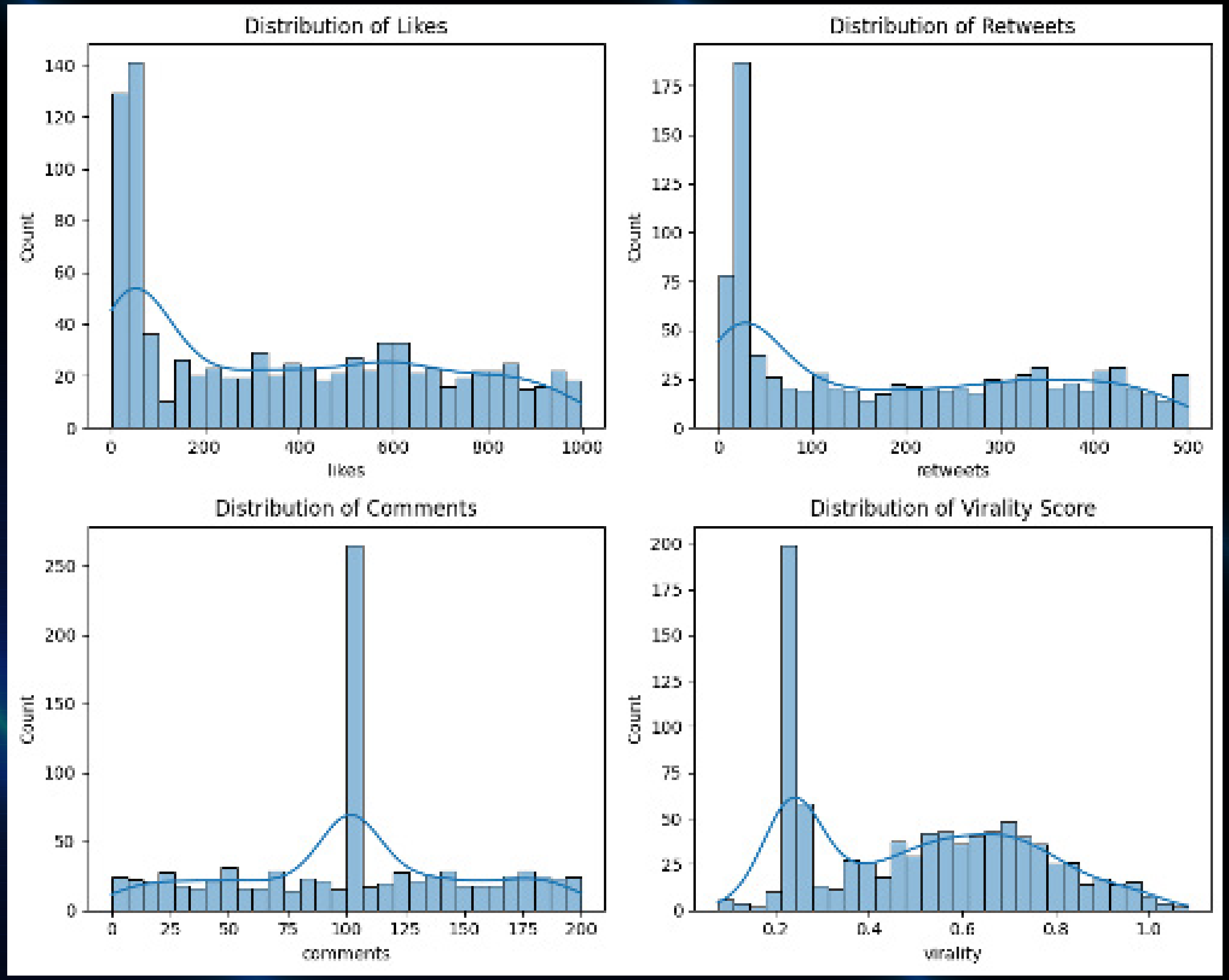
Challenges:

1. Class Imbalance: More neutral posts than positive or negative.
2. Language Complexity: Colloquial language and non-standard expressions.
3. Sarcasm Detection: Difficulty in interpreting sarcastic tones.

Data Analysis

Feature Discovery:

1. **User Mentions:** Analyze tweets for user interaction patterns.
2. **Domain Classification:** Categorize tweets by topics or industries to identify key themes.
3. **Key Themes/Topics:** Extract major topics from tweet content to enhance sentiment analysis.



Data Processing

- **Platform Selection: Focused on Twitter data for sentiment analysis.**
- **Data Cleaning: Removed unnecessary spaces to improve text quality.**
- **Sentiment Categories: Standardized sentiment labels to positive, negative, and neutral.**
- **Column Refinement: Selected relevant columns and renamed them for clarity.**
- **Dataset Enhancements:**

- **Addressed missing data by adding timestamps to both datasets.**
- **Conducted feature extraction for hashtags instead of disregarding them.**

- **Data Integration: Concatenated datasets; replaced null values with mean values.**
- **Target Variables:**

- **Defined and calculated target variable Virality as a metric of the weighted number of likes, comments, and retweets.**

Learning Algorithms Used

1. **XGBoost (eXtreme Gradient Boosting):**

- **Reason for Selection:**

Handles varied data effectively, making it suitable for diverse Twitter datasets.
Known for its efficiency and accuracy in processing structured data.
Offers efficient and scalable training due to parallel processing capabilities.

- **Advantages:**

Provides robust performance through gradient boosting, reducing overfitting.
Capable of handling missing data and capturing complex patterns.

- **Disadvantages:**

Requires careful tuning of hyperparameters, which can be computationally expensive and time-consuming.

Learning Algorithms Used

2. LSTM (Long Short-Term Memory):

- **Reason for Selection:**

Ideal for sequential text data like tweets due to its ability to capture temporal dependencies.
Excels in understanding the context and long-term dependencies in text.

- **Advantages:**

Retains information over long sequences, crucial for sentiment analysis in text.
Effective in capturing nuances and emotional tone in tweets.

- **Disadvantages:**

Requires large amounts of data to learn effectively, which may not always be available. Prone to overfitting.

Related Works

1. VADER (Valence Aware Dictionary for sEntiment Reasoning)

- VADER is a rule-based sentiment analysis tool specifically designed for social media text, such as tweets.
- It combines a sentiment lexicon with five heuristic rules that account for grammatical and syntactical conventions used in expressing sentiment intensity.

Significance and Contributions:

- VADER provides a robust solution for analyzing sentiment in microblogging contexts, which are characterized by short, informal text.
- It outperforms several traditional sentiment analysis benchmarks, including human raters, in accurately classifying tweets into positive, neutral, or negative sentiments.

Why It Makes Sense:

- Social media platforms like Twitter present unique challenges due to the brevity and informal nature of posts. VADER addresses these by incorporating lexical features like emoticons and slang.
- Its rule-based approach allows it to generalize well across different contexts without the need for extensive training data.

Corroborated Findings:

- Empirical validation shows that VADER achieves high classification accuracy, outperforming individual human raters.
- The system's ability to handle sentiment intensity makes it particularly effective in social media environments compared to other lexicons like LIWC or machine learning models that require large datasets and extensive training.

Related Works

2. TWEETEVAL: Unified benchmark and comparative evaluation for tweet classification.

- TweetEval is a unified benchmark designed for Twitter-specific classification tasks, addressing the fragmented landscape in social media NLP research.
- It includes 7 heterogeneous tasks: sentiment analysis, emotion recognition, offensive language detection, hate speech detection, stance prediction, emoji prediction, and irony detection.

Significance and Contributions:

- The framework standardizes evaluation protocols across these tasks, providing a strong set of baselines and facilitating comparative evaluation.
- It leverages existing pre-trained language models like RoBERTa, which are further trained on Twitter corpora to enhance performance on social media text.

Why It Makes Sense:

- Social media text presents unique challenges due to its conversational and idiosyncratic nature. TweetEval addresses these by offering a comprehensive suite of tasks that reflect real-world applications.
- The benchmark facilitates the development of robust NLP models capable of handling the unique features of Twitter data, such as short text length and informal language

Corroborated Findings:

- Demonstrated that pre-trained models fine-tuned on Twitter data enhance task performance, supporting the efficacy of transfer learning in NLP.
- Established competitive baselines that serve as benchmarks for future research in tweet classification.
- Showed that models like RoBERTa, when retrained on domain-specific (Twitter) data, can outperform generic pre-trained models in several tasks.

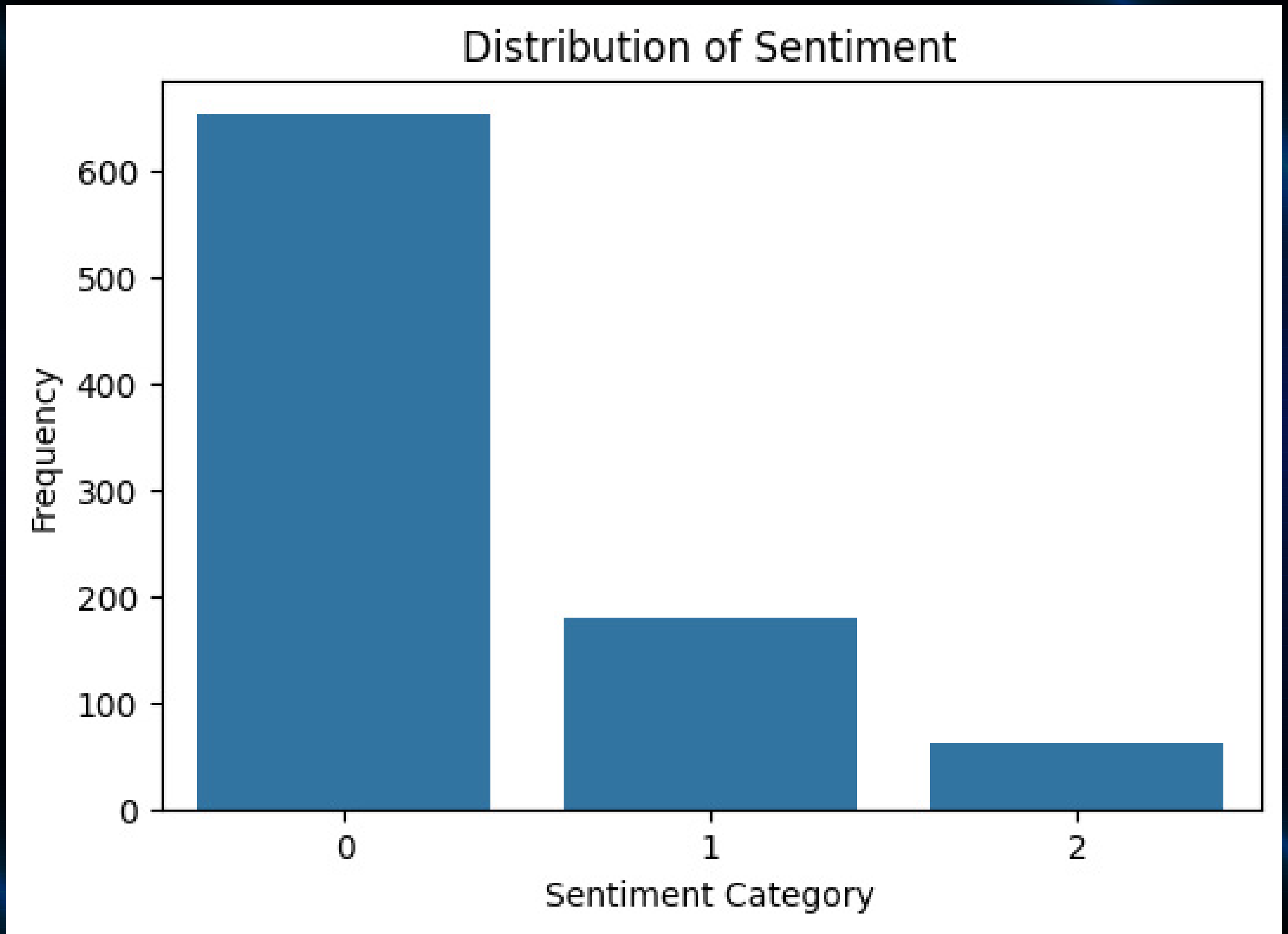
Learnings from Dataset

- **Dataset Limitations: The dataset was not large enough and lacked some essential features for efficient classification. Presence of too many sentiment classes led to overfitting and incorrect sentiment predictions, affecting overall accuracy.**

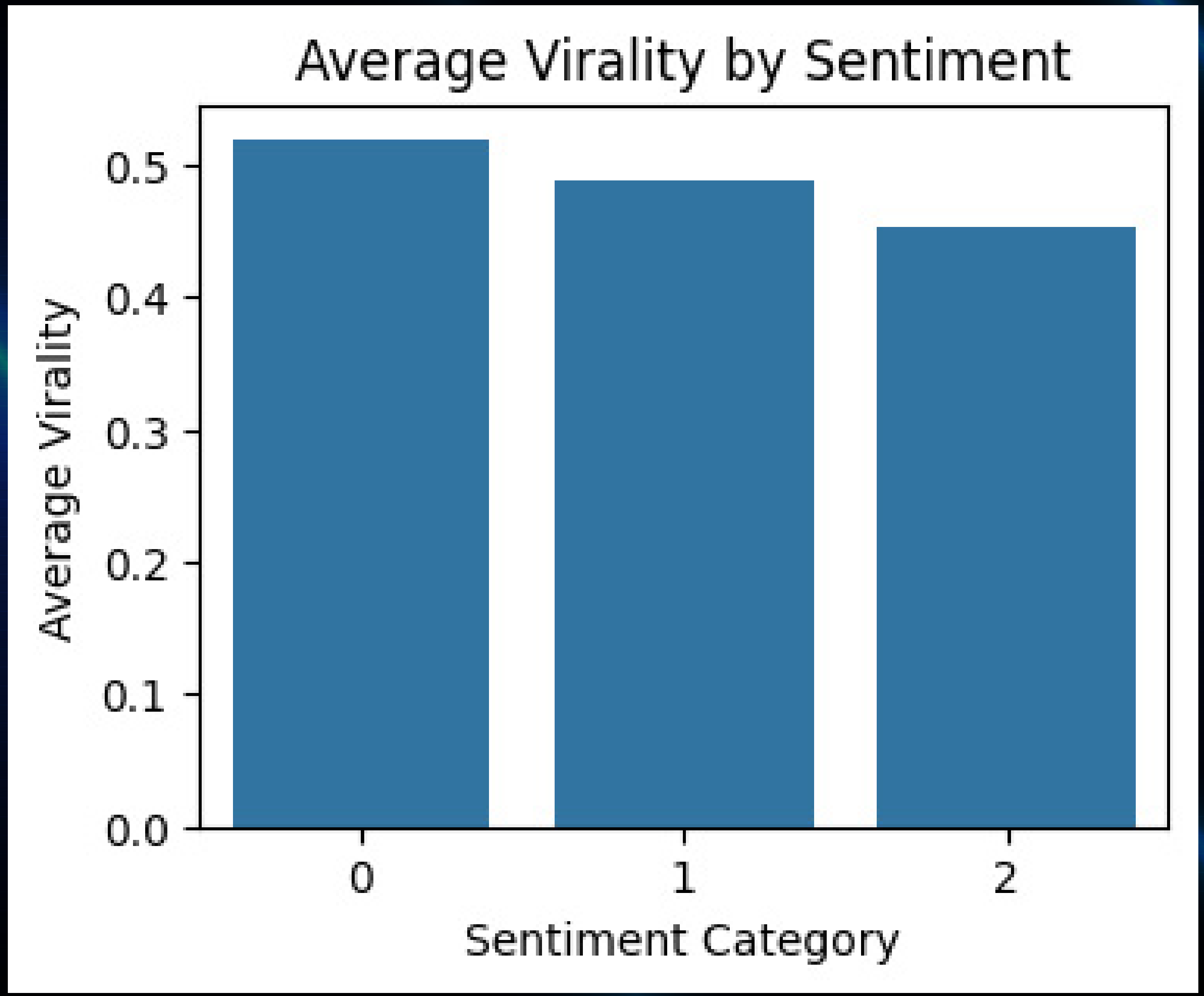
- **Data Challenges: Incomplete features like missing comments, follower counts, and timestamps in datasets required additional preprocessing. The need to standardize sentiment categories to avoid class imbalance.**

- **Solution Development: Simplified sentiment categories to reduce complexity and improve model generalization. Enhanced feature extraction techniques to capture key themes and improve classification accuracy.**

Visualizations:



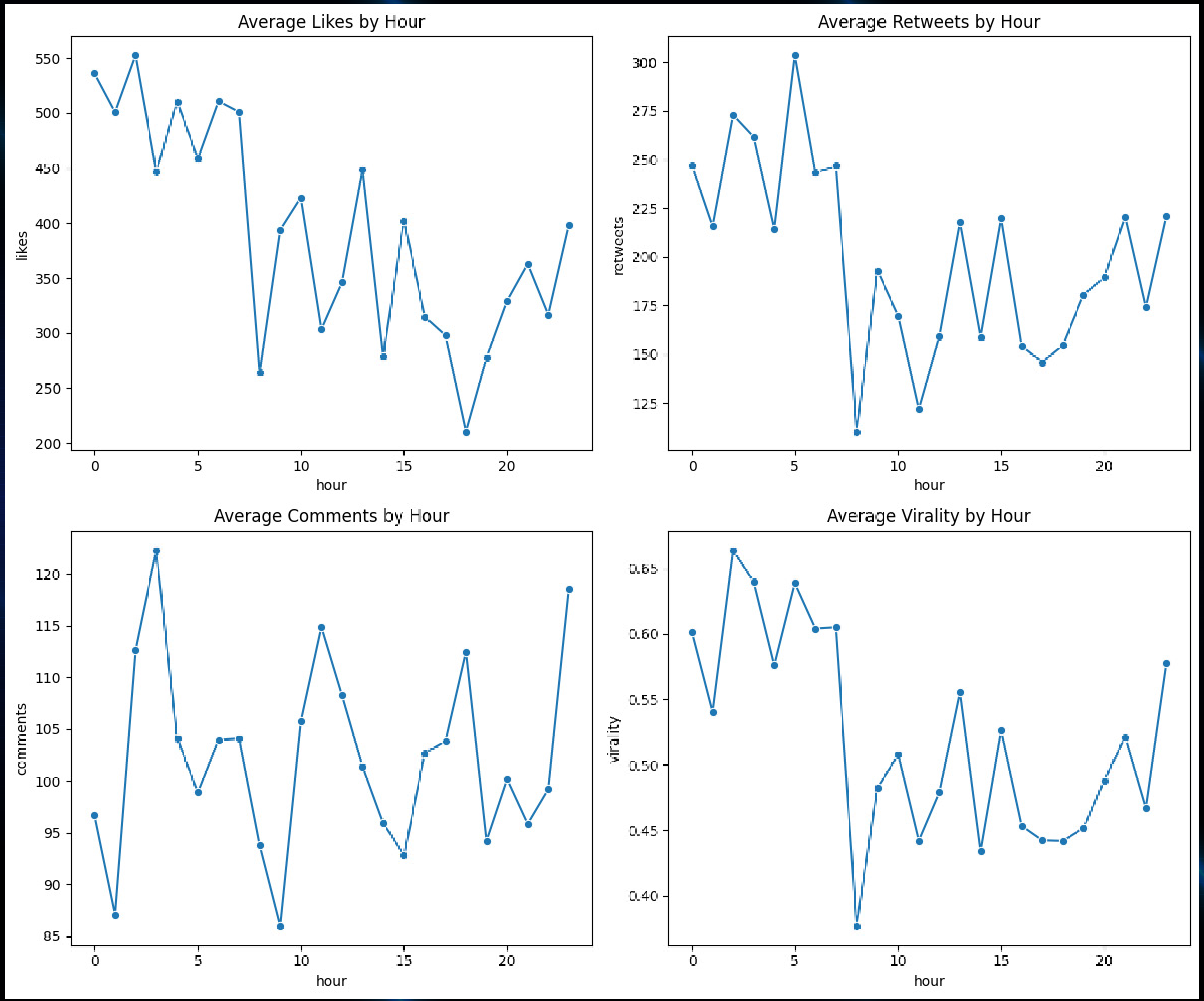
Outlines the general mood or tone prevalent in the dataset. A predominance of positive sentiment indicates general satisfaction or happiness among the user base.



After calculating and analyzing average virality. we can determine that neither of the positive, negative, or neutral sentiments is more likely to go viral.

AVERAGE PLOTS

These insights are useful for determining optimal posting times. For instance, it can be seen that posts made during the early hours i.e. early morning or late night tend to go more viral, this can guide content scheduling strategies to maximize reach and interaction.



Conclusion

1. **Recommended Models:**
   - **For Sentiment Classification: VADER is recommended due to its high accuracy with smaller datasets and its effectiveness in handling unbalanced data proportions.**
   - **For Virality Prediction: XGBoost is suitable for regression tasks, though there is room for improvement in R^2 scores through feature engineering or hyperparameter tuning.**

2. **Scalability Considerations:**
   - **Logistic Regression and Random Forest offer computational efficiency, making them suitable for scaling beyond local settings with limited resources.**
   - **XGBoost provides better performance but requires more computational resources.**

TABLE I. PERFORMANCE STATISTICS OF MODELS

Model	Accuracy (%)	F1 Score	Mean Squared Error (MSE)	R^2 Score
Random Forest	68.71	0.67	0.409	0.111
Logistic Regression	72.08	0.71	0.304	0.339
XGBoost	86.03	0.78	0.00339	0.040
LSTM	53.07	0.37	0.0536	-0.54