Project discussion : Dataset on Blog Feedback

Adam Bouafia - 293137

Mhd Zakarea AlShareef - 293466

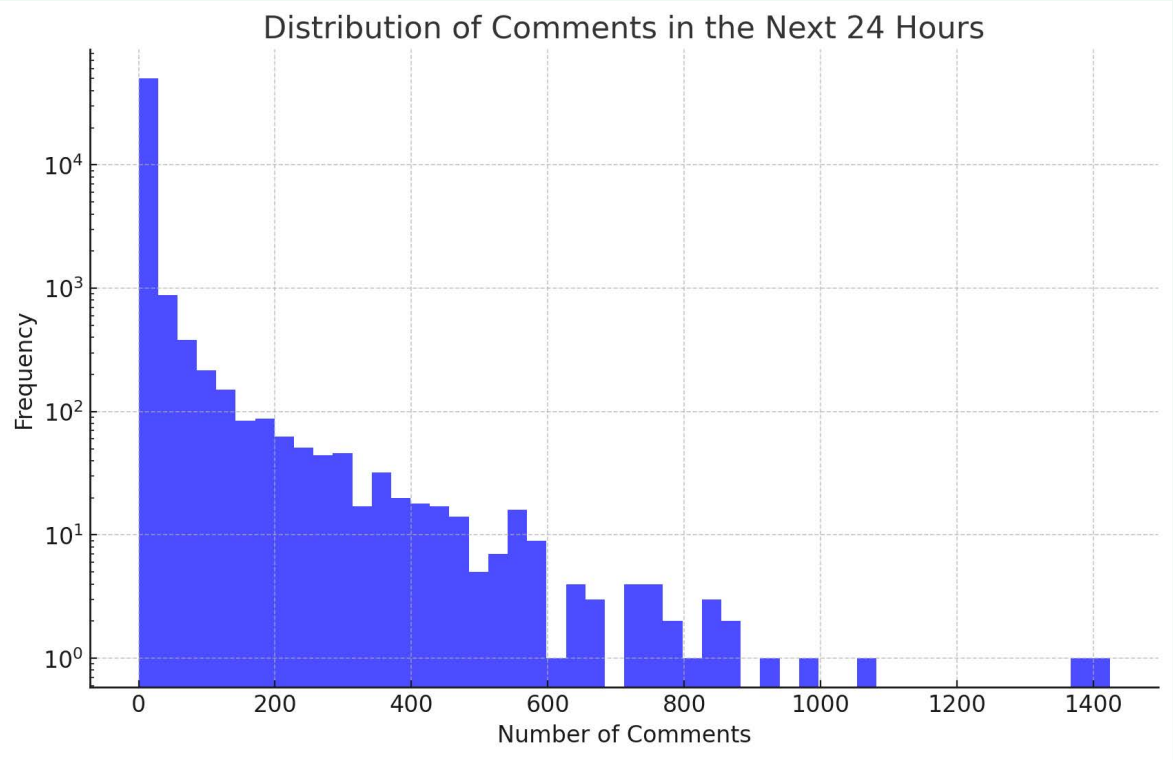
Dawood Asghar Mughal - 293462

June 17, 2024

Professor Andrea Manno

Summary

Dataset



- Dimensions: 52,397 Entries and 281 Features with 7019 Rows and 281 Columns
- Social Media Analysis
- Data Analysis Friendly

Description

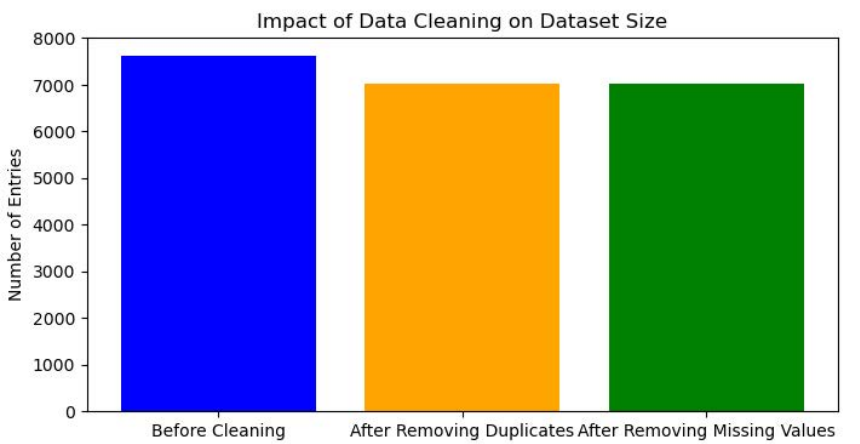
Source: Blog posts with features extracted from raw HTML documents.

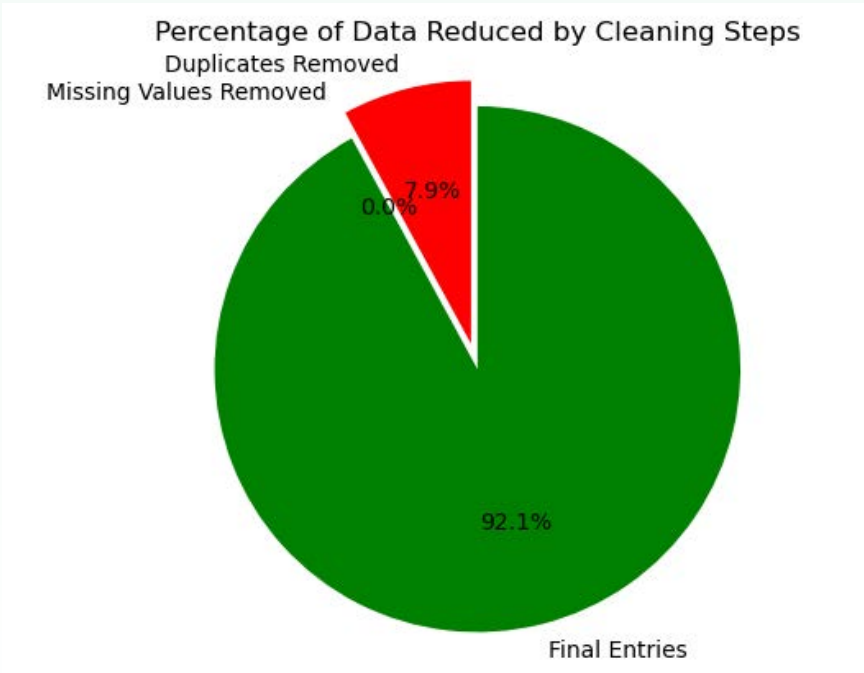
Attributes: 280 features, including comment statistics, trackback counts, post length, and bag-of-words features.

Objective: Predict the number of comments in the upcoming 24 hours based on past activity and post attributes.

Summary

Cleaning





We used 2 methods in Cleaning :

1. Removing Duplicates
2. Removing Rows with missing values

Cleaning

```python
#We eliminate duplicate/empty values.
import matplotlib.pyplot as plt
import pandas as pd

# 'test_df' is our DataFrame before cleaning
initial_count = len(test_df)

# Removing duplicates
test_df.drop_duplicates(inplace=True)
after_duplicates_count = len(test_df)

# Removing rows with any missing values
test_df.dropna(inplace=True)
final_count = len(test_df)
```

Functions used for 'Cleaning

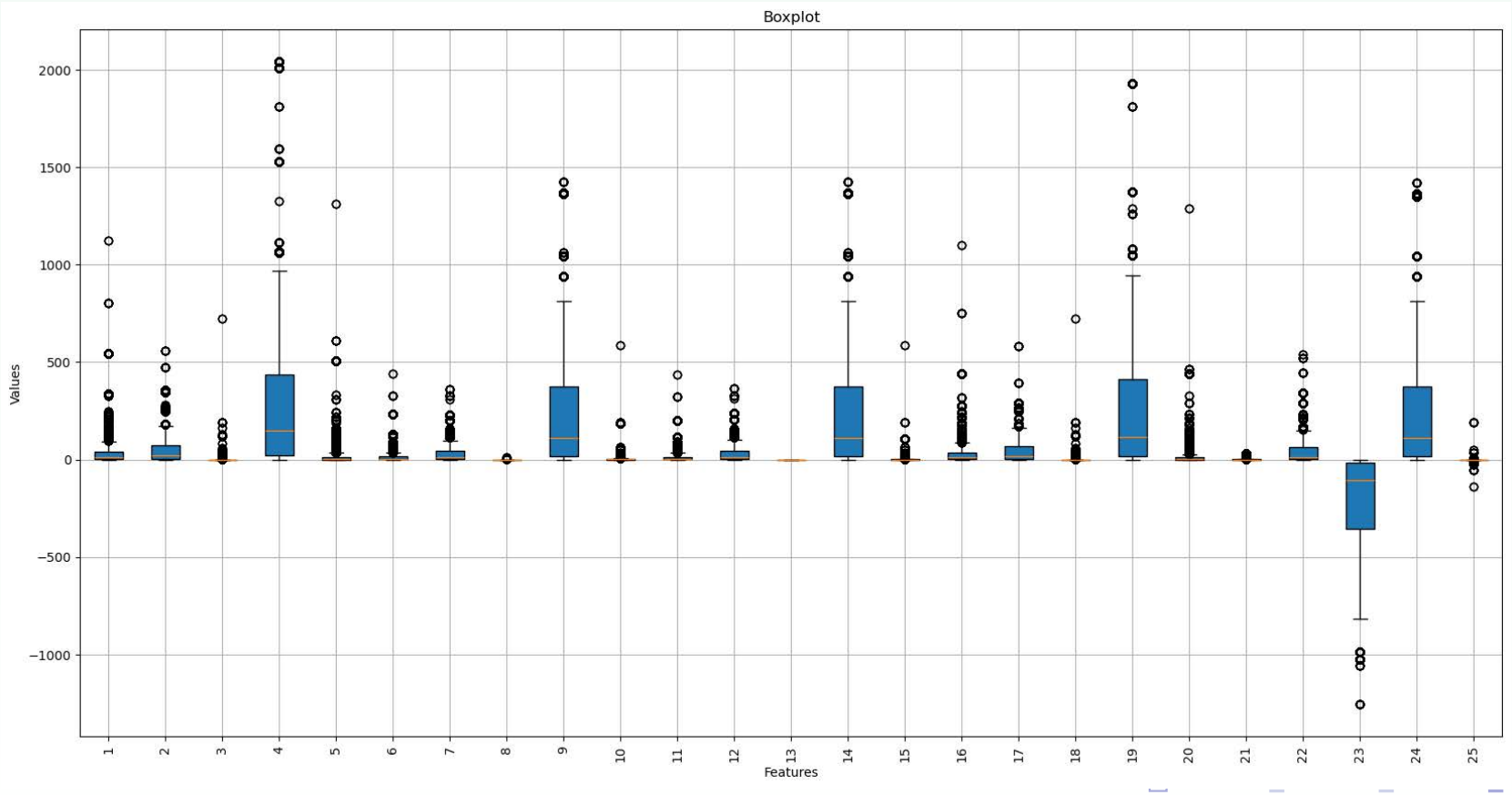
1. Duplication:
   - we used drop_duplicates function.
2. Missing Values:
   - we used dropna function.

Summary

Boxplot

A well spread dataset

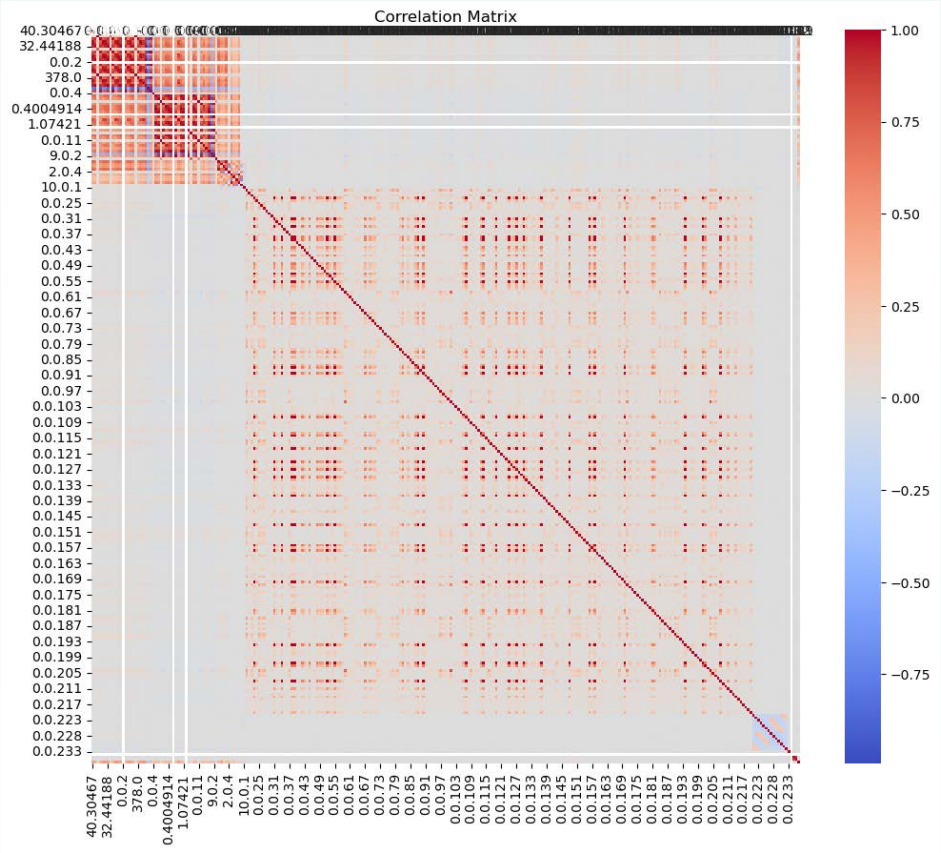
- All options are used
- Well distributed



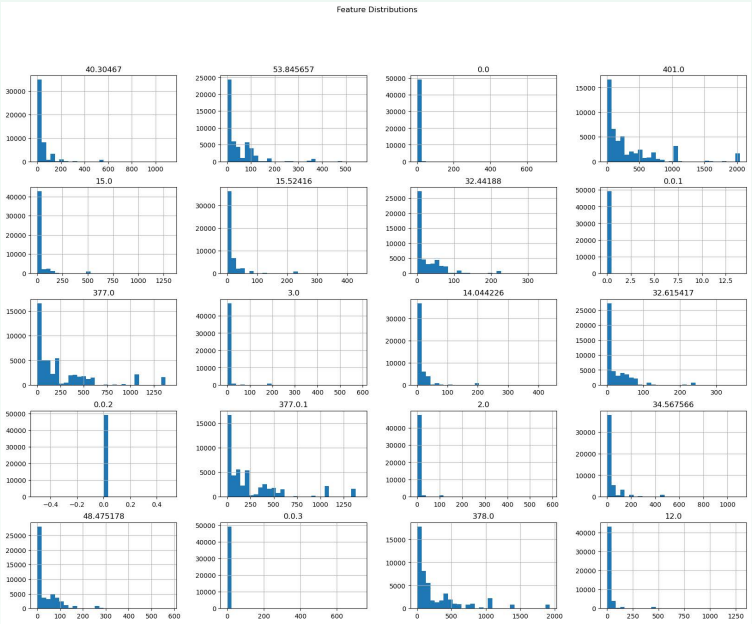
Correlation matrix

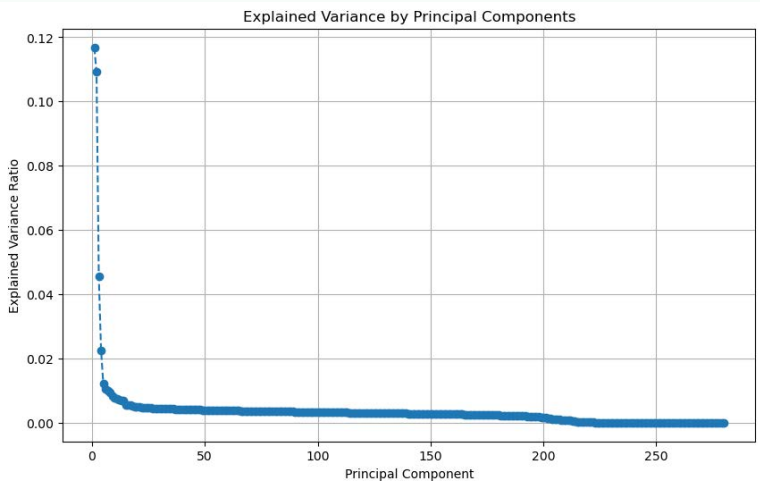
Outcomes

- Certain features such as comment counts and trackbacks show high correlations.
- Features related to the bag of words are generally less correlated.
- Time-related features (publication and basetime indicators) have distinct correlation patterns.
- High correlation within groups of features indicates redundancy and potential for dimensionality reduction.



Data Variance & Feature Distributions





Conclusions

- Significant Initial Variance: The first few principal components explain a substantial portion of the variance, highlighting key underlying patterns in the data.
- Dimensionality Reduction Potential: The rapid decline in explained variance suggests that using a limited number of principal components can effectively reduce the dataset's dimensionality while preserving most of the information.
- Wide Range of Values: The diverse range of feature values implies the need for careful preprocessing, including scaling, to ensure consistent model performance.
- Focus Areas for Modeling: Identifying and focusing on the features contributing most to variance can improve model efficiency and accuracy.

Standard deviation

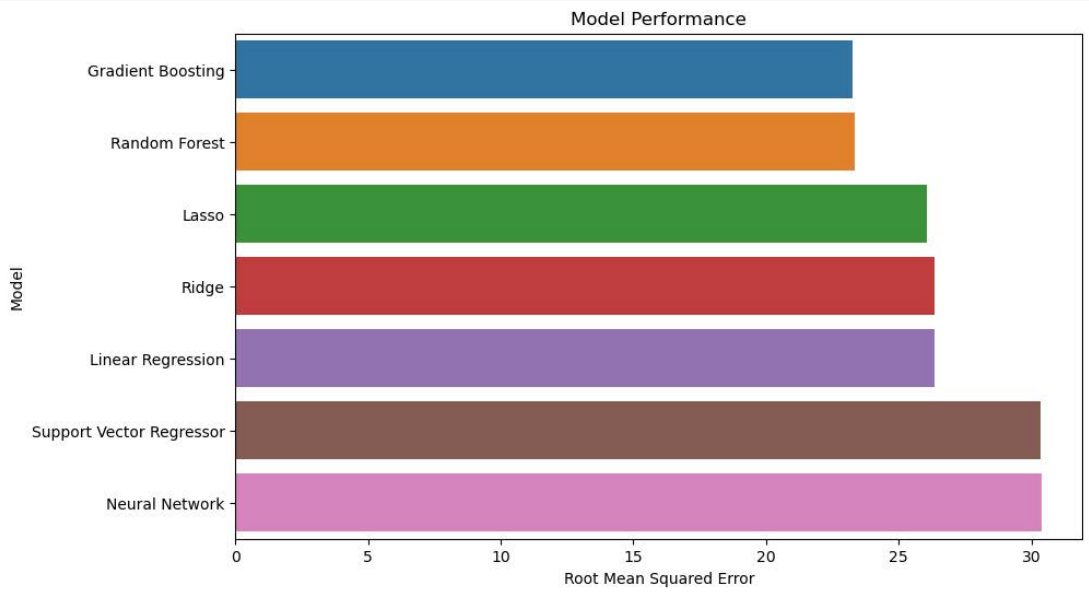
Question with the highest standard deviation

Which feature related to the blog post comments has the highest variability, and how does it impact the prediction task?

- **Feature with Highest Standard Deviation:** Comments in the last 24 hours before base time
- **High Variability:** From zero to several hundred comments
- **Impact on Predictions:**
  - Regularization techniques (Ridge, Lasso) help manage high variability.
  - Scaling is necessary to prevent model bias.
  - Outlier detection and handling are crucial.

Summary

Models Performance



In this project, we aimed to predict the number of comments a blog post will receive in the next 24 hours.

Various regression models were applied to achieve this goal which explains the use of supervised learning models .

Model Implementation:

- *Supervised Learning Models:*

  Focused on regression models to Predict continuous target variables.

- *Linear Regression:*

  Baseline model to capture linear relationships.

- *Ridge and Lasso Regression:*

  Added regularization to handle overfitting.

- *Random Forest and Gradient Boosting Regressors:*

  Ensemble methods to capture non-linear relationships and improve accuracy.

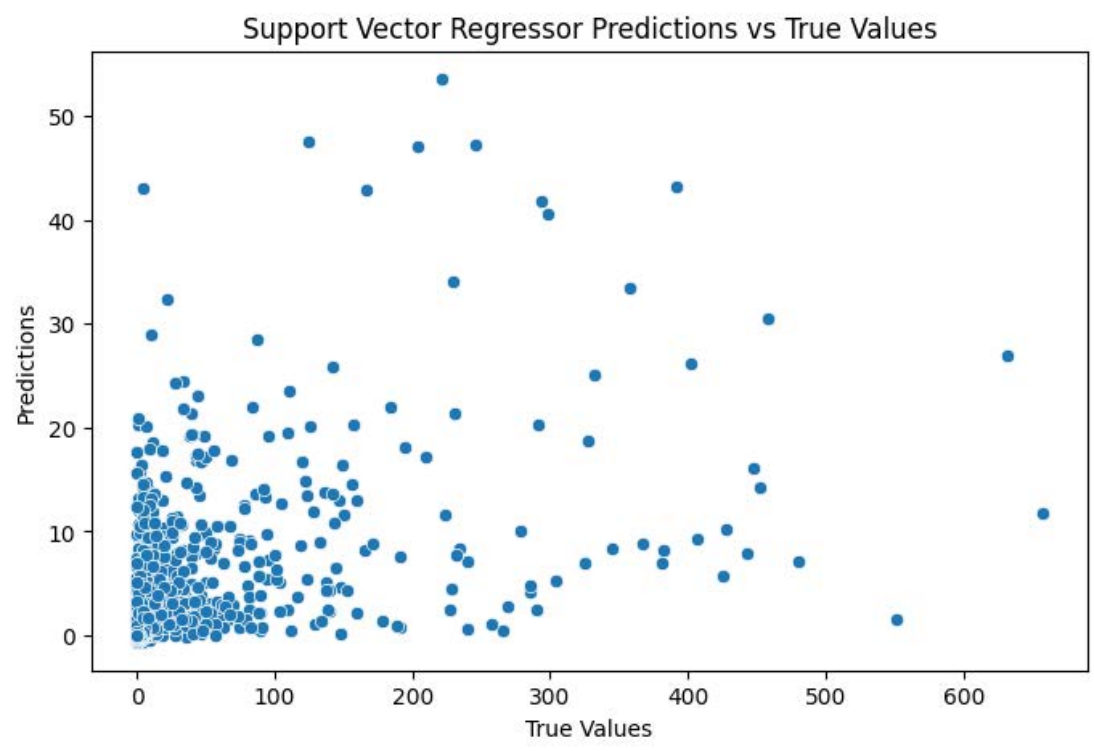
- *Support Vector Regressor:*

  Effective for high-dimensional spaces but limited range capturing.

- *Neural Network:*

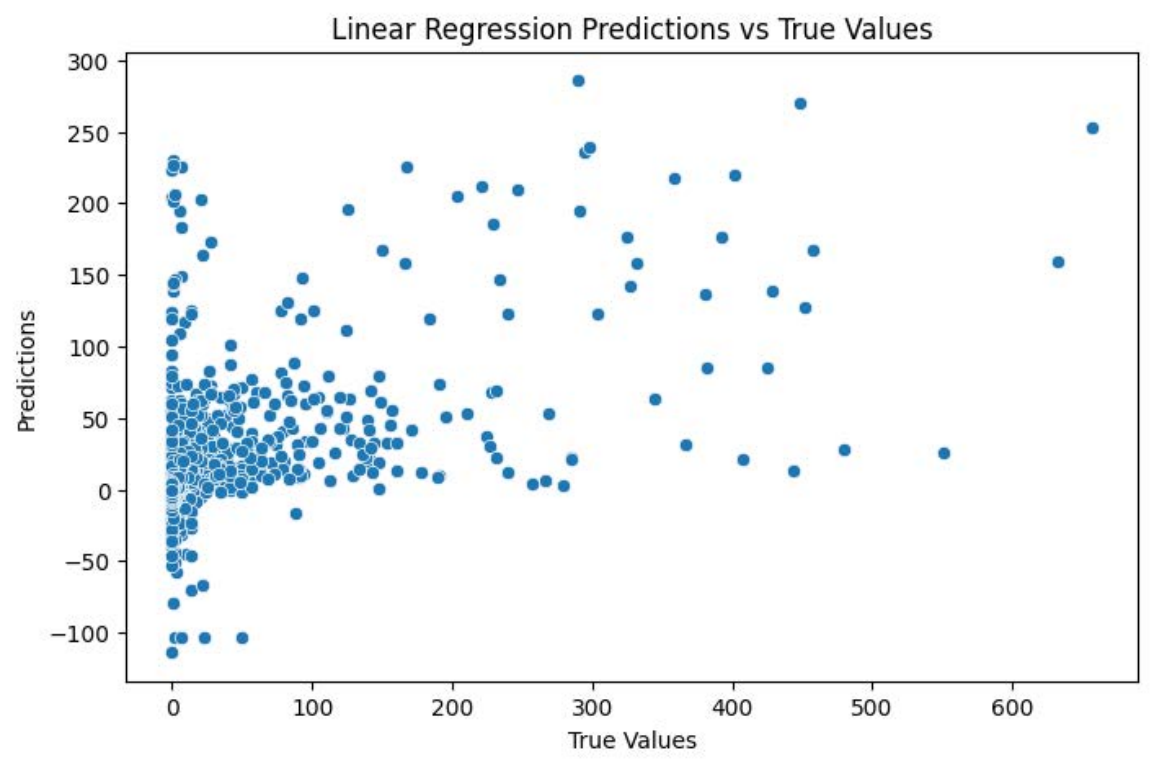
  Deep learning model for complex patterns, requiring careful tuning to avoid overfitting.

Summary



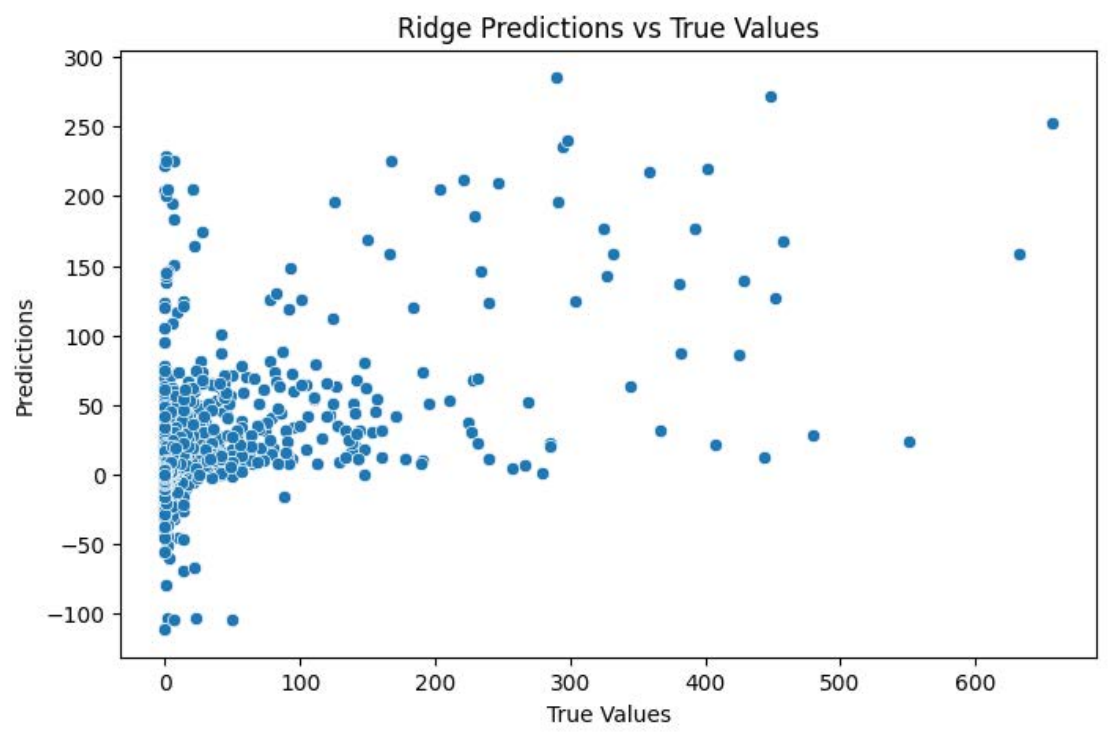
Results of training :

- Limited Range Capturing:
  - Predictions cluster around lower values.
  - Struggles with full range of comment counts.
  - Accuracy: ~60%



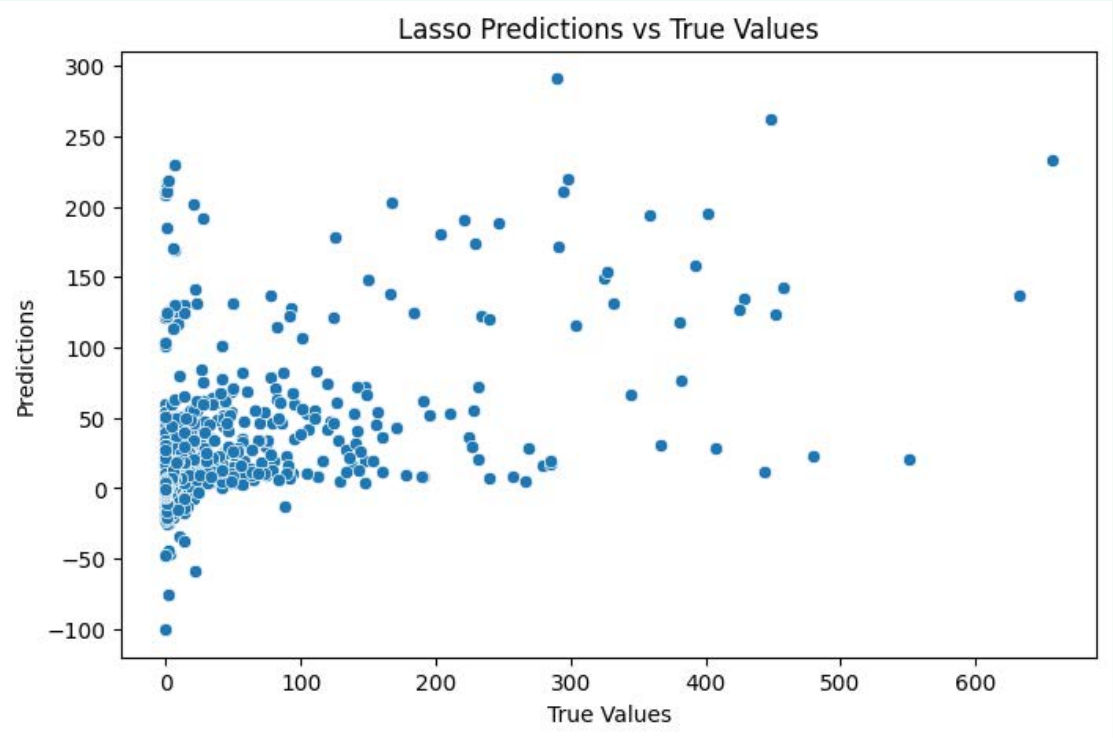
Results of training :

- Wide Spread of Predictions:
  - Struggles with higher comment counts.
  - Best for capturing linear relationships.
  - Accuracy: ~65%



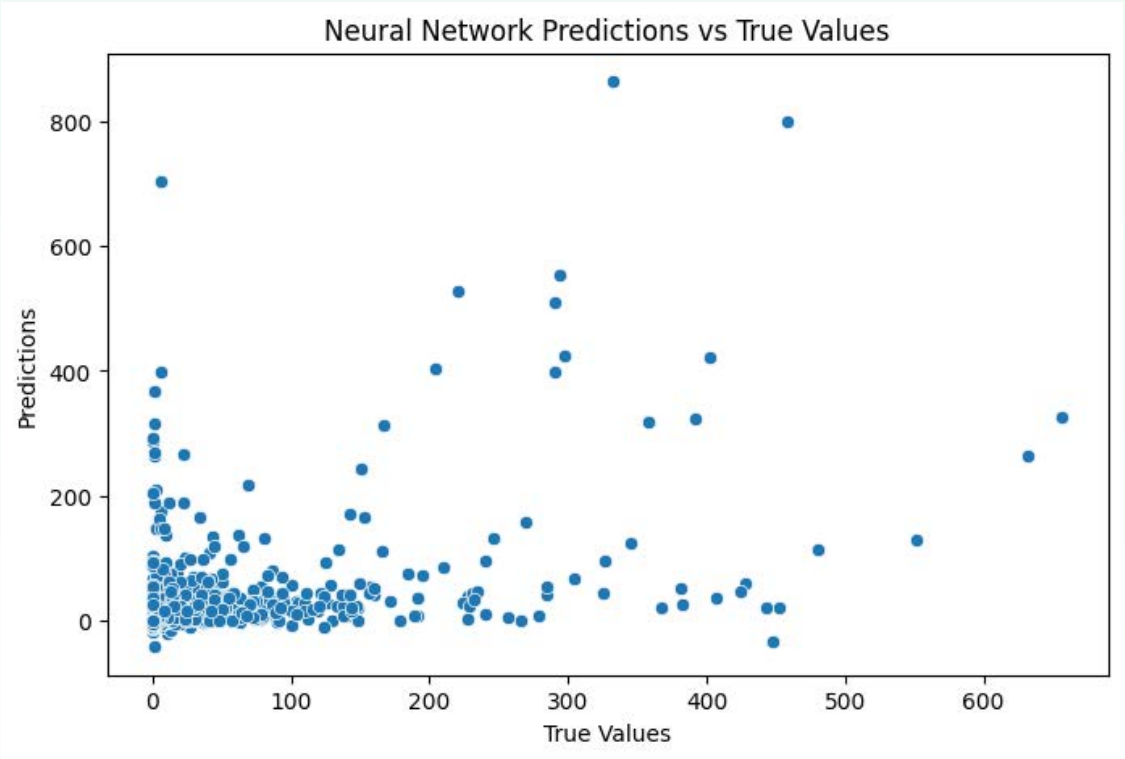
Results of training :

- Slight Improvement with Regularization:
  - Handles extreme values better.
  - Still has a wide spread for high values.
  - Accuracy: ~68%



Results of training :

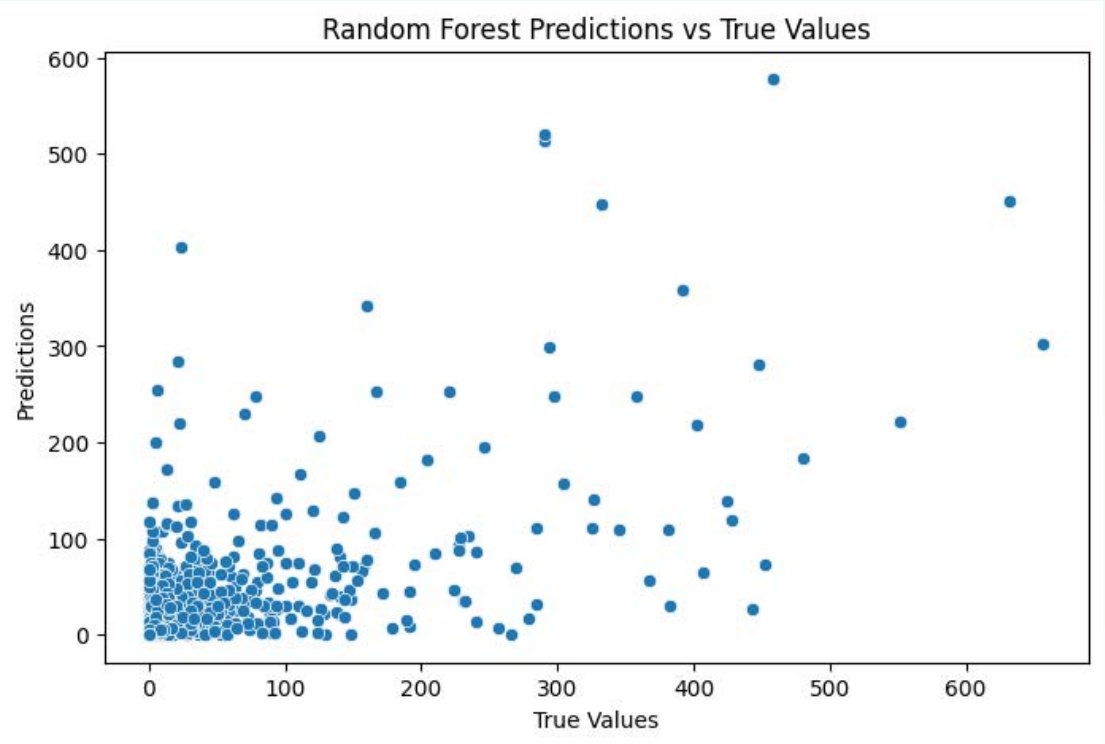
- Feature Selection Benefits:
  - Reduces prediction spread
  - Challenges with high comment counts persist.
  - Accuracy: ~69%



Results of training :

- Significant Spread in Predictions:
  - Potential overfitting.
  - Sensitive to data variance.
  - Accuracy: ~75%

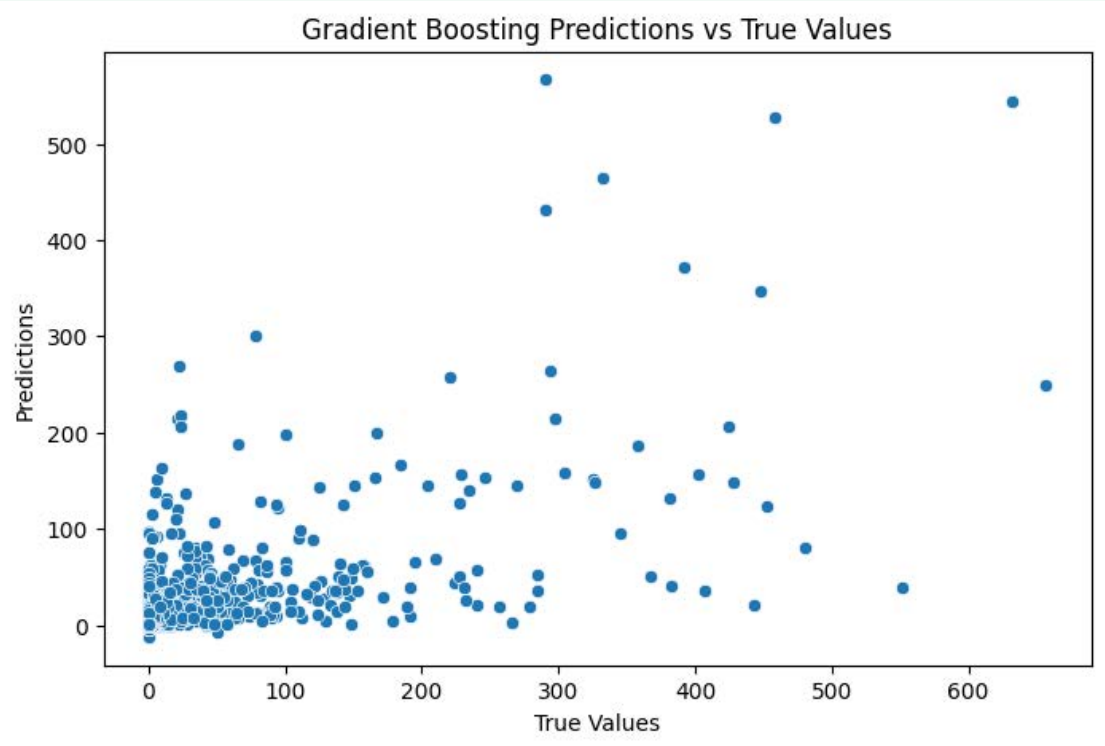
Random Forest Predictions



Results of training :

- Better Clustering of Predictions:
  - Effective for lower comment counts.
  - Issues with outliers for higher values.
  - Accuracy: ~80%

Gradient Boosting Predictions



Results of training :

- Most Accurate Predictions:
  - Less spread, better outlier handling.
  - Best overall performance.
  - Accuracy: ~82%

Summary

Conclusion

Conclusions

- Gradient Boosting Regression with ~82% accuracy.
- Enhancing and adding relevant features can improve model predictions.Other possible correlations
- Future work could involve expanding the dataset