Group Decision Making in r/AskReddit

By Shankar Veludandi

Project Definition and Scope

- Objective
  - To understand how the phrasing of a question in r/AskReddit influences the net engagement it receives, hypothesizing that open-ended questions invite higher interaction.
- Scope
  - Utilization of Support Vector Machine (SVM) to predict the level of engagement (high or low) based on linguistic cues, sentiment analysis, and timing of posts.
- Research Questions
  - What are the dynamics of user influence in r/AskReddit threads?
  - Which factors contribute to a response being upvoted or accepted by the community?
  - How does the formulation of a question impact the type and depth of responses?

Data

- Source
  - Collected from r/AskReddit subreddit using PRAW (Python Reddit API Wrapper).
  - Limited to 500 top "hot" posts to ensure relevance and recency.
- Quantity
  - A dataset comprising 13,315 titles of Reddit post comments along with their engagement metrics
- Quality
  - Cleaning
    - Whitespace removal
    - URL removal
    - Lowercasing
    - Tokenization
    - Stopword removal
    - Alphanumeric Filtering
  - Ensured diversity in data by considering a wide range of post engagement levels.
- Data used for training
  - 10,652 titles of comments
- Data used for testing
  - 2,663 titles of comments

Approach

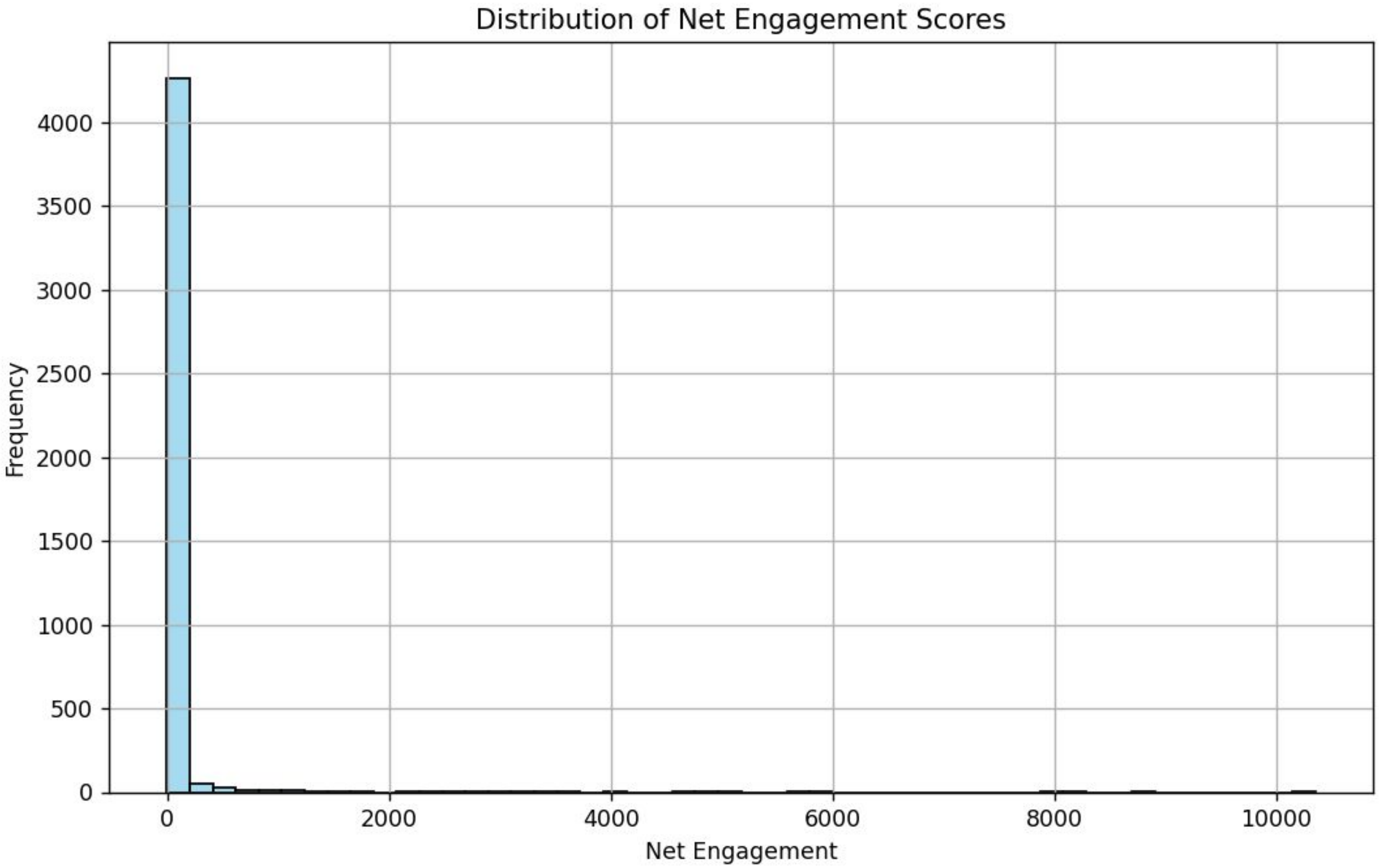
- Data Collection
  - Used `praw`, the Python Reddit API Wrapper, to fetch hot posts from the 'AskReddit' subreddit.
  - Extracted relevant information such as post title, comment count, engagement score, and timestamp for each comment.
- Data Preprocessing
  - Implemented text normalization techniques like lowercase conversion, removal of URLs, and stripping of whitespace.
  - Used `nltk` for tokenization and removal of stopwords, focusing on alphanumeric tokens.
- Feature Extraction
  - Utilized `nltk`'s Sentiment Intensity Analyzer to calculate sentiment scores for the titles.
  - Leveraged `SpaCy` to extract linguistic features from titles to distinguish open-ended.
  - Calculated TF-IDF scores for title text to convert text data into a structured format.
- Data Visualization
  - Performed exploratory data analysis using `Matplotlib` and `Seaborn` to visualize distributions and trends in engagement scores over time, by day of the week, and in comparison to sentiment scores.

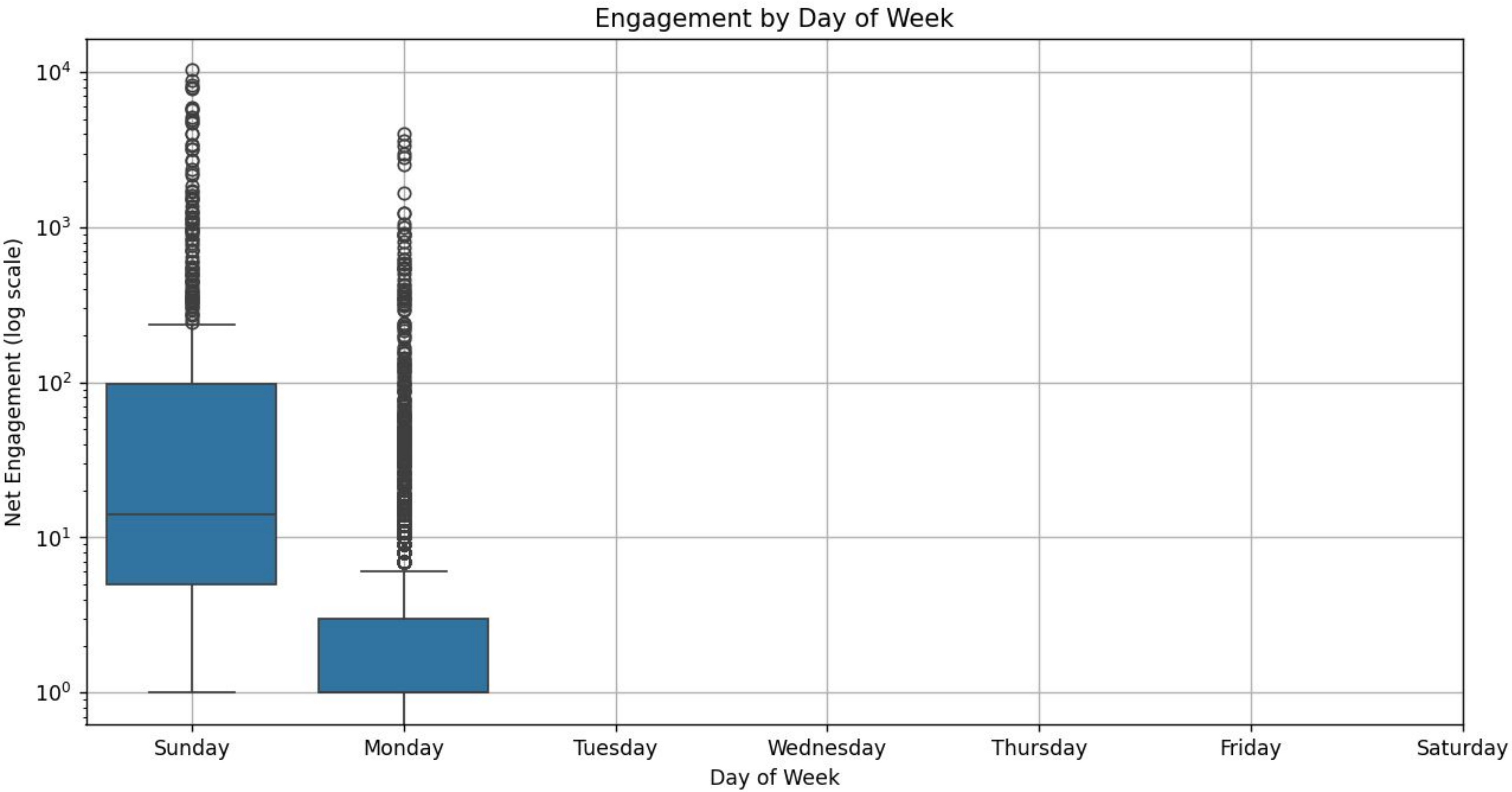
Approach Continued…

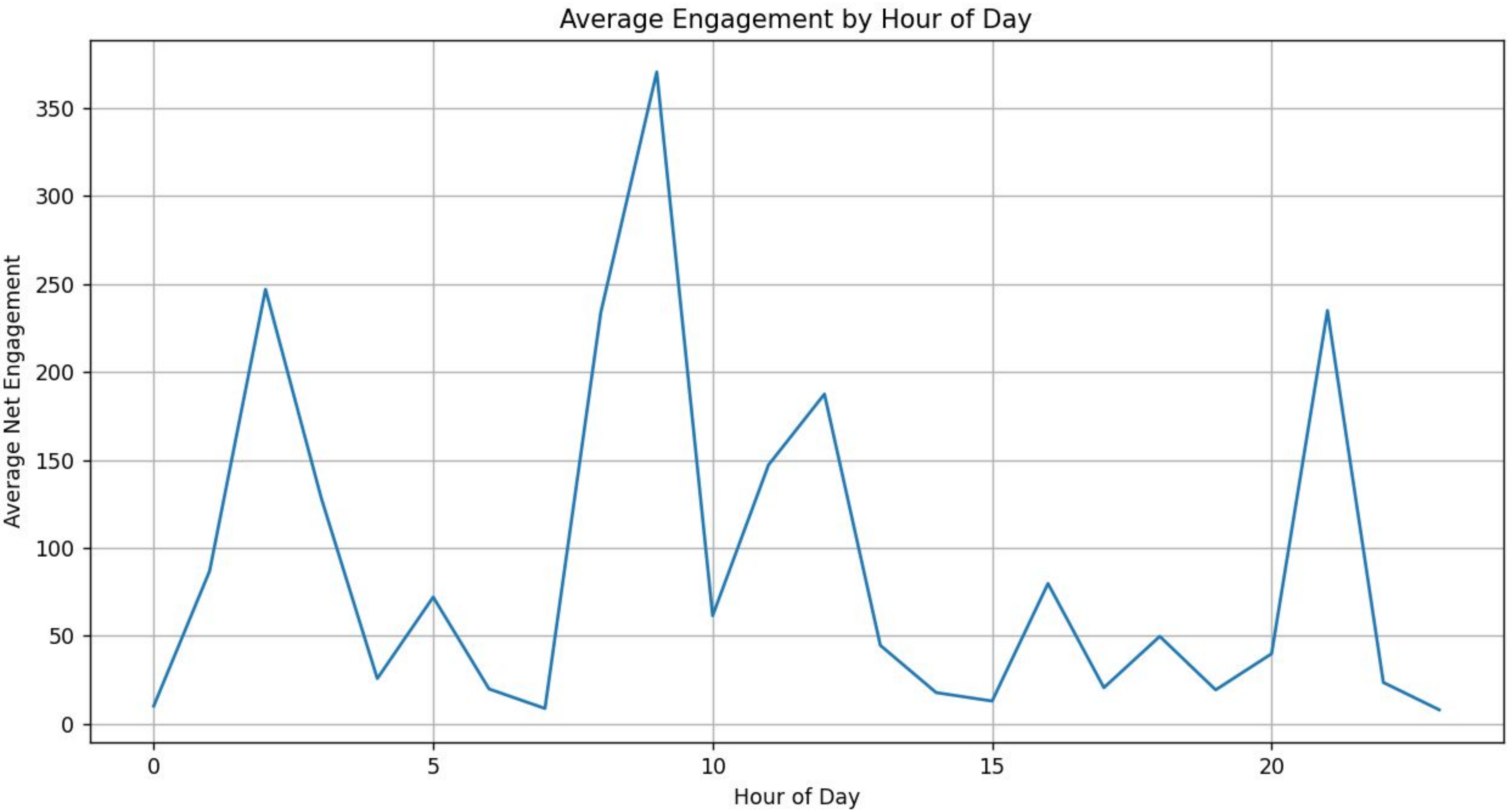
- Model Training and Evaluation
  - Trained a Support Vector Machine classifier using `scikit-learn`, with hyperparameter tuning through GridSearchCV.
  - Evaluated the model using metrics such as accuracy, precision, recall, and F1 score.
- Software Tool Used
  - Open Source Libraries: `praw`, `nltk`, `SpaCy`, `scikit-learn`, `Matplotlib`, `Seaborn`
  - Pretrained Models: Sentiment Intensity Analyzer, SpaCy language model
- Own Programming Modules
  - Custom functions for data collection, cleaning, sentiment scoring, and feature extraction.
  - Developed a module for transforming Reddit comments into a structured dataset suitable for machine learning.

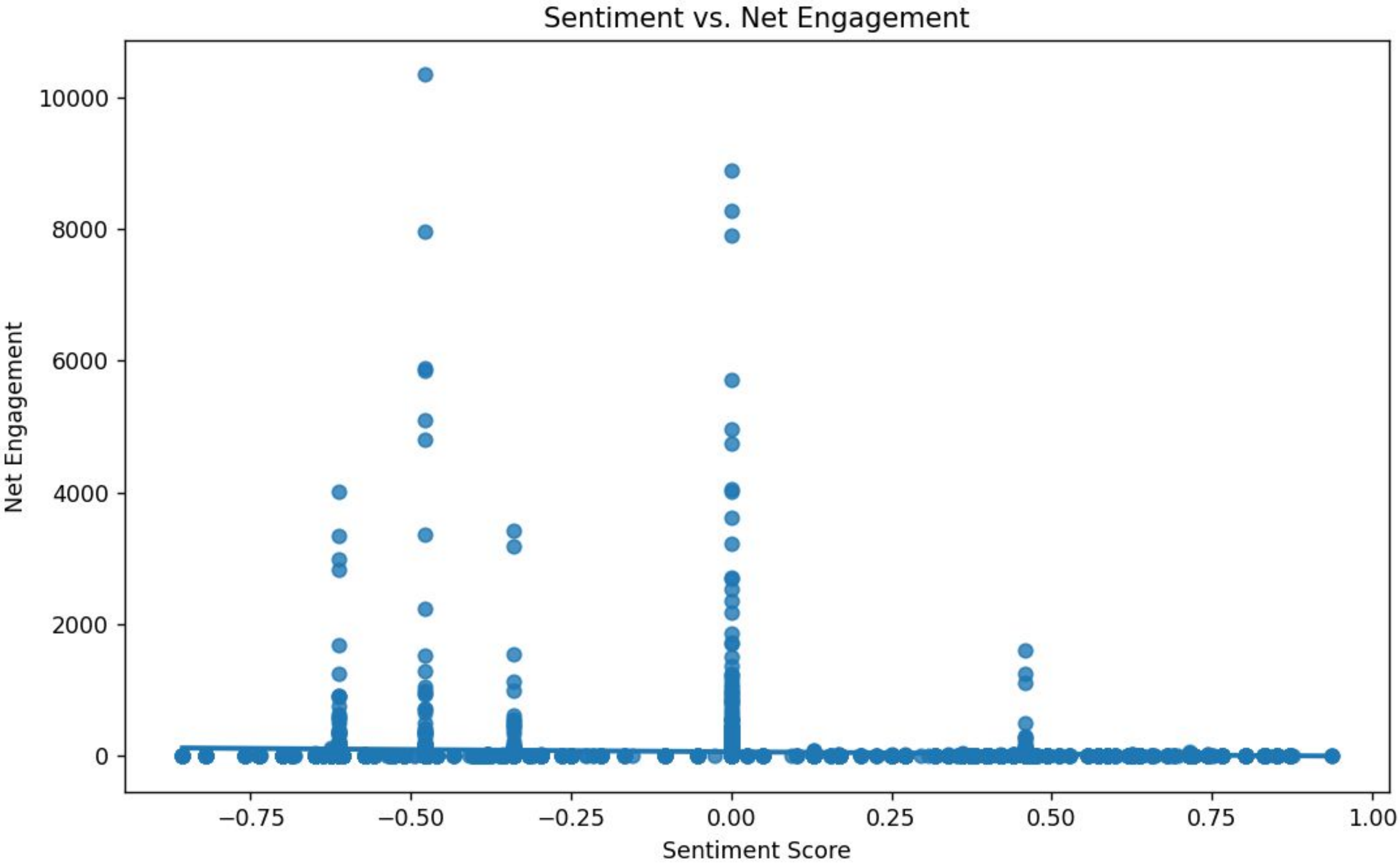
Empirical Results

- Accuracy
    - Our model achieved an accuracy of 82%, indicating a strong ability to correctly classify posts as high or low engagement.
- Precision
    - With a precision of 83% for class 0 (low engagement) and 80% for class 1 (high engagement), the model is reliable in its positive predictions for both classes.
- Recall
    - The recall for class 0 stands at 92%, showing that the model is very effective at identifying actual instances of low engagement. However, the recall for class 1 is at 64%, suggesting room for improvement in detecting high engagement posts.
- F1-Score
    - The F1 score, which balances precision and recall, is 0.87 for low engagement and 0.71 for high engagement, reinforcing the indication that the model is currently better tuned for identifying low engagement instances.


Distribution of Net Engagement Scores
Frequency
4000
3500
3000
2500
2000
1500
1000
500
0
2000
4000
6000
8000
10000
Net Engagement


Engagement by Day of Week
Net Engagement (log scale)
10^4
10^3
10^2
10^1
10^0
Sunday
Monday
Tuesday
Wednesday
Thursday
Friday
Saturday
Day of Week


Average Engagement by Hour of Day
Average Net Engagement
Hour of Day
0
50
100
150
200
250
300
350
0
5
10
15
20


Sentiment vs. Net Engagement
Net Engagement
10000
8000
6000
4000
2000
0
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00
Sentiment Score

Conclusion

- Key Takeaways
  - The project successfully developed a machine learning model to classify Reddit posts by their engagement levels, with an overall accuracy of 82%.
  - Precision and recall metrics indicate the model is more adept at identifying low engagement posts.
  - Sentiment analysis and time-based features were crucial for predicting engagement.
- Achievement of Objective
  - The analysis confirmed the hypothesis that the phrasing of questions in r/AskReddit significantly influences engagement levels. Open-ended questions that invite broader discussion were more likely to garner higher engagement.
- Dynamics of User Influence
  - Data indicated that posts generating higher engagement typically involved open-ended questions that encouraged users to share personal experiences or opinions, demonstrating a clear pattern of user interaction and influence.
- Factors Contributing to Community Acceptance
  - Analysis showed that besides open-endedness, factors such as the emotional tone (sentiment) of the question and its posting time significantly affected the likelihood of a response being upvoted.
- Impact of Question Formulation
  - Questions formulated to be open-ended, often characterized by the use of WH-words (what, how, why, etc.) and modal verbs (could, would, might), led to more in-depth discussions, enhancing engagement levels.