Wise Shopper: A Paradigm of Customer Review Analysis

Adarsh Mahor†
College of Engineering & Applied Sciences
University of Colorado
Boulder, Colorado, US
Adarsh.Mahor@colorado.edu

Mayuresh Dongare†
College of Engineering & Applied Sciences
University of Colorado
Boulder, Colorado, US
Mayuresh.Dongare@colorado.edu

Marshall Pauley
College of Engineering & Applied Sciences
University of Colorado
Boulder, Colorado, US
Marshall.Pauley@colorado.edu

ABSTRACT

These days, online evaluations are essential for improving consumer communications worldwide and affecting consumer purchasing behaviour. E-commerce behemoths such as Amazon, and others give customers a forum to vent about their experiences and give prospective customers accurate information about how well a product works. To extract meaningful insights from a huge set of reviews, reviews must be classified as either positive or negative. Sentiment analysis is computational research that uses the text to extract subjective information. Beyond sentiment analysis, we also investigate GPUs and CPUs using a multi-pronged analytical strategy in the field of data mining. We use advanced modelling techniques, regression analysis, and data visualization to answer ten important questions and reveal complex patterns and insights in the data. These analyses involve estimating GPU clock speed determinants, categorizing GPUs according to performance, and estimating GPU memory capacity from technical parameters. We also look at how product specs differ depending on the vendor and how GPU development trends can be tracked over time using machine learning models. Further research focuses on Principle Component Analysis for numerical attributes and examines the impact of post scheduling on social media exposure and interaction. Additionally, we use advanced techniques such as Latent Dirichlet Allocation (LDA) and Non-negative Matrix Factorization (NMF) in Natural Language Processing (NLP) to identify trends in the textual content of submissions and comments. This thorough analysis not only improves our knowledge of product sentiment but also helps forecast consumer behavior and technology trends, offering useful information to both consumers and businesses. This work represents a comprehensive strategy for comprehending and utilizing user opinions, technical assessments, and industry developments to support well-informed decision-making in the quickly changing technology sector.

KEYWORDS

Data Mining, Sentiment Analysis, KNN, SVM

1 Introduction

In the present era of online shopping, the credibility of user reviews has become a crucial factor in the decision-making process of potential buyers. However, the increasing use of fake reviews generated by online bots has made it challenging to distinguish between genuine and fake reviews. It is imperative to ensure that user reviews are honest and authentic, as they play a significant role in shaping the reputation of the product and impacting the purchasing decision of the buyer. Therefore, it is

†Authors hold equal amount of contribution.

Wise Shopper'30, April 2024, Boulder, Colorado USA

https://doi.org/xxx/xxxxxxx

essential for online platforms to implement measures to prevent the use of fake reviews and ensure that the reviews are from genuine users who have actually used the product.

An increasing number of consumers prefer to purchase on different e-commerce platforms due to the quick development and widespread use of e-commerce technologies. In contrast to traditional offline shopping, which requires customers to wait until the weekend to go shopping, online shopping allows consumers to purchase whenever and wherever they choose, saving time and effort. Additionally, e-commerce platforms offer various products in various designs, allowing customers to purchase the items they want without leaving their homes. While consumers find online shopping convenient, there are several issues with the products sold on these platforms due to their virtual nature. These issues include inconsistent descriptions of products and actual goods, subpar quality, flawed after-sales care, and more. As a result, sentiment analysis of the commodity rating of the acquired goods on e-commerce platforms is quite important.

PROBLEM STATEMENT

World trade is shifting on various e-commerce websites; evaluating things before purchasing is typical. Additionally, buyers are more likely to rely on product reviews while purchasing. Thus, one of the most important fields currently is analyzing the data from those customer evaluations to make the data more valuable to new buyers. While the e-commerce website gives buyers the option for product reviews and rating the product, in this world, several sellers on the e-commerce market provide fake ratings and product reviews. This is required to cross verify the product review on another website before buying it. The Jhaveri, D (2015) gives an example of this. Another issue buyers face is multiple products from different sellers/brands. This creates massive confusion among customers about buying a product. So, the biggest problem addressed in this research is making purchasing easier for customers. Summarizing and being more specific, the following problems will be addressed in this research: 1. Cross-verifying and giving the factual review/sentiment to the customer 2. Analyzing product data to make purchases easier for the customer 3. Finding more constraints which could affect the sale of the products.

OBJECTIVES

As explained in the problem statement, the research articulates two objectives, which are described as follows:

Objective 1: Sentiment Analysis on Products Purchased by Real-Life People

➢ Textual Data Collection - Gathering posts and comments about e-commerce product discussions by using the Reddit API or web scraping techniques and capturing relevant metadata like post title, content, author, and timestamp.

➢ Text Preprocessing - Improve the text data by removing unnecessary elements like stop words, noise, and special characters, and to ensure standardized word forms, use lemmatization and stemming techniques in addition to tokenization.

➢ Result Aggregation - Combining the sentiment score for each product to obtain an overall sentiment analysis.

➢ Visualizations and insights - Making use of graphs or charts to illustrate the sentiment analysis results. Examine the sentiment trends over time or between various products to learn more about the preferences and viewpoints of your customers.

Objective 2: Analyzing Product Data - Prices, Products, Customer Ratings and Reviews

➢ Data Collection - Scraping data from e-commerce, content rating platforms, and forum social networks or utilizing their APIs to collect information about products, prices, customer ratings, reviews, and other attributes of interest.

➢ Data Integration and Cleaning - Consolidate the gathered information into an organized format. Maintain uniformity in the titles of products, standardize costs, client ratings, and reviews, and deal with anomalies as well as missing data.

➢ Exploratory Data Analysis (EDA) - Understanding how product prices, ratings, and other attributes are distributed using EDA and Utilizing statistical

measures and visualizations to examine relationships between various variables.

➢ Statistical Analysis - Finding relationships between features of a product using statistical techniques and looking for any noteworthy trends or patterns.

➢ Predictive Modelling - Building predictive models to forecast product prices or customer ratings based on different features. Implementing methods such as machine learning algorithms or regression analysis on acquired data.

➢ Visualization and Report: Make use of Python libraries to showcase findings using graphs and charts and prepare a report outlining the conclusions and revelations made during the analysis.

2 Data Exploration

When it comes to exploring data from Reddit, tapping into the platform's Application Programming Interface (API) offers a robust and efficient solution. Reddit provides developers with access to its vast repository of posts, comments, and other user-generated content through its API, enabling seamless data retrieval and analysis. Reddit's API offers an organized approach to accessing data, eliminating the need for complex web scraping techniques. Reddit encourages API usage to ensure compliance with legal and ethical guidelines, avoiding the risk of legal repercussions associated with unauthorized web scraping activities. PRAW simplifies interactions with Reddit's API, providing developers with a high-level interface for accessing posts, comments, user profiles, and more. PRAW is a powerful tool that enables developers to retrieve data from Reddit's API in a streamlined manner. With PRAW's intuitive methods and classes, developers can easily specify search queries, filter results, and retrieve data in real-time or through historical archives. PRAW also provides flexible querying options, catering to the specific requirements of developers. There are versatile querying capabilities available, allowing developers to analyze posts from specific subreddits, keep track of user interactions, or even monitor trending topics. Integrating PRAW into Python-based data exploration workflows is easy. The tool has comprehensive documentation and a user-friendly interface, making it effortless to integrate with popular data analysis tools and frameworks. This enables developers to leverage Reddit's data in conjunction with other data sources. Following code block shows the code which is used to implement PRAW:

```
import praw
import csv
from datetime import datetime

# Initialize PRAW with your Reddit app credentials
reddit = praw.Reddit(
    client_id="arr32sSYSTUkkgugEdShVQ",
    client_secret="moz5DTxGYrDBXlO496uJmZKH2bB80w",
    user_agent="Wise Shopper",
)
```

Fig 1. Implementation of PRAW in python

The data which is web-scraped from reddit is analyzed using some plots for better understanding and further analysis.

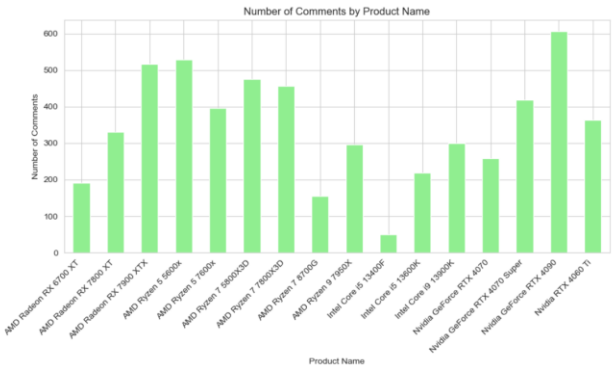

Fig 2. The following Bar Plot displays the overall number of comments on various Reddit posts about specific products. As we examine the plot, we can clearly see that the Nvidia GeForce RTX 4090 is the most talked-about product, while the Intel Core i5 13400F is the least discussed. This information is valuable in understanding the popularity and engagement of different products within the online community.

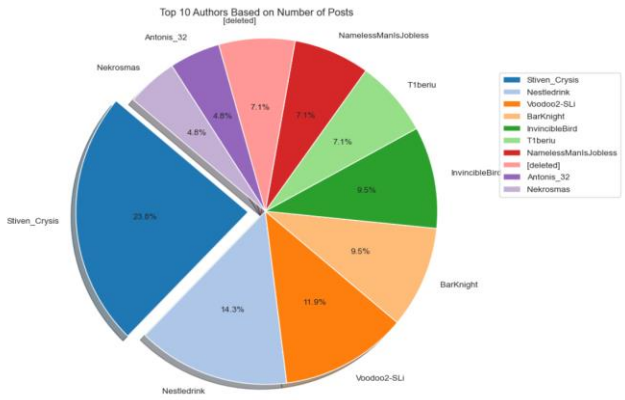

Fig 3. The pie chart presented above provides a visual representation of the distribution of posts made on the popular social media platform Reddit. The chart is based on data obtained from a specific data set, and it offers valuable insights into the contribution made by various users.

As per the chart, Stiven_Crysis stands out as the most active user with 23.8% of posts attributed to them. This indicates that they have been consistently engaged with the platform and have contributed significantly to the Reddit community. Furthermore, the chart also highlights that 7.1% of the authors or users who made the posts have either deleted their accounts or changed their usernames. This could be due to various reasons like privacy concerns, account deletion, or a change in the user's identity. Overall, the pie chart provides a clear and informative overview of the distribution of posts on Reddit, and it helps to identify the most active users and the impact of such users on the platform.

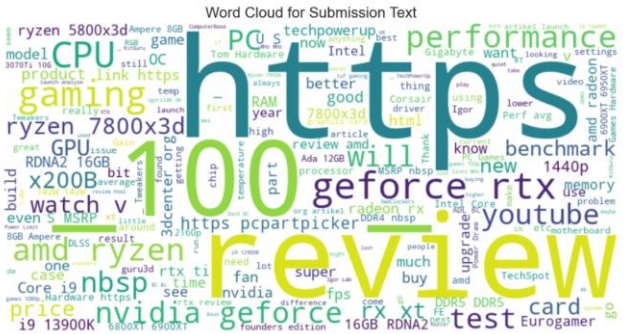

Fig 4. Word Cloud for the submissions

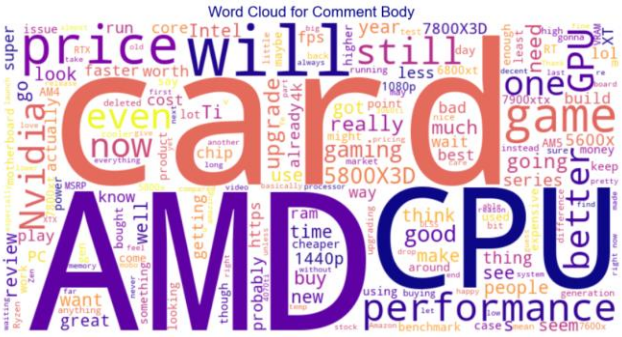

Fig 5. Word Could for the Comments

From the above Fig 4 and Fig 5, we can observe the word cloud for two columns of data. While A word cloud is a visual representation of text data where the size of each word is proportional to its frequency in the text. It provides a quick overview of the most common words in a corpus and helps identify key themes or topics. It is noticeable that the most commonly used words in comments on Reddit posts are AMD, CPU, and card. The most frequently used words in Reddit posts are https, review, and 100.

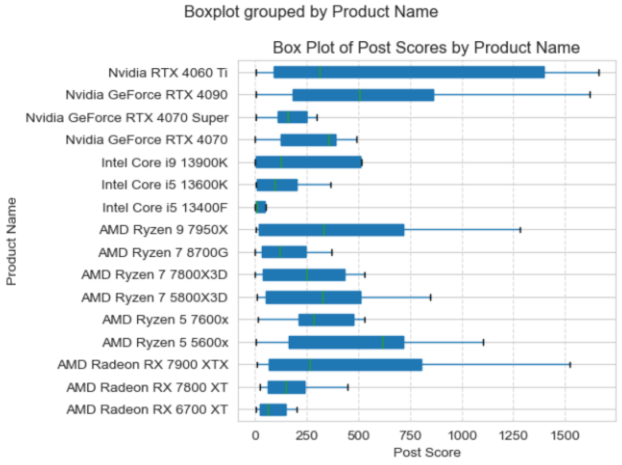

Fig 6. Post score refers to the numerical value that determines the difference between the total number of upvotes and downvotes on a certain post.

From Fig 6, we can understand the central tendency of the data, which includes the median post scores, interquartile range (IQR), and any outliers present in our dataset. This data comprises various processors and graphics processing units, providing a comprehensive view of the performance of these components.

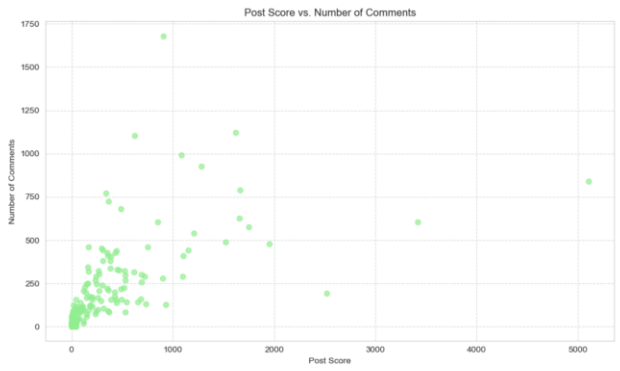

Fig 7. Scatter plot to Understand variation of number of comments as per post score

Fig 7. plot provides a visual representation of the dataset and highlights the presence of a few outliers. These outliers have significantly high post scores but relatively low numbers of comments, whereas some others have low post scores but high numbers of comments. The plot helps to identify these unusual data points and better understand the distribution of the dataset.

3 Models Implemented and their Observations

In this section, various different models used are discussed and implemented on the dataset. The section is defined by stating a question or hypothesis and trying to solve it; it is discussed in the following bullets:

Question 1: What Do Reddit Users Really Think? An Insightful Sentiment Analysis of User

Reviews

Removal of explicit language, such as "F***" and "F******", to prevent sentiment distortion caused by varying interpretations of such words. Our `clean_text` function strips away punctuation and stops words, refining the dataset for more accurate sentiment evaluation. We introduced the `get_sentiment` function to quantify the emotional tone of the text, transforming subjective sentiments into objective, numerical scores. Score dynamic assesses words for positive or negative connotations, generating a sentiment score that reflects the text's emotional impact. These scores facilitate a structured analysis of user sentiments, enabling the identification of trends and comparisons across different comments for deeper insights into public perceptions

AMD's Ryzen processors, particularly the Ryzen 5 5600x, seem to have a higher sentiment score, suggesting more positive reviews or comments. Nvidia's RTX 4090, a high-end GPU, has a relatively low sentiment score compared to other products, which could be due to various factors such as cost, or expectations not being met. This contrasts with the RTX 4070 Super, which scores better. The overall sentiment score across all comments is approximately 0.127, indicating a generally neutral to slightly positive sentiment.

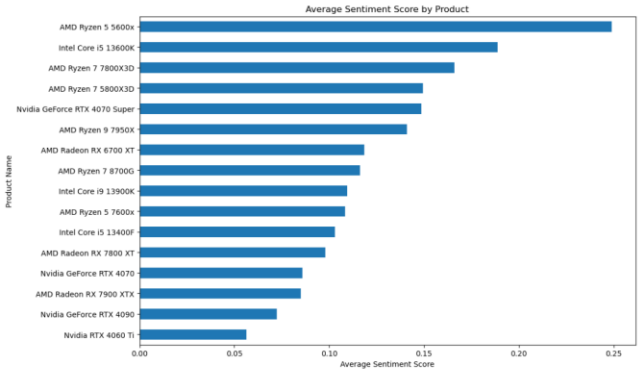

Fig 8. A sample of average sentiment scores on the products

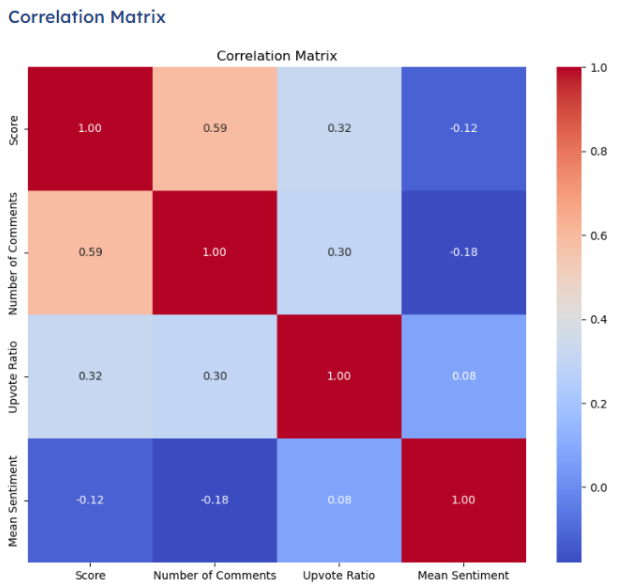

Fig 9. Correlation Matrix to understand the sentiments and scores relation.

Score vs. Number of Comments (0.593690): Posts with higher scores generally receive more comments, suggesting that engaging content drives discussion.

Score vs. Upvote Ratio (0.322291): Posts with higher scores often have better upvote ratios, indicating that well-received content tends to be more favorably rated.

Score vs. Mean Sentiment (-0.119986): There is a slight negative correlation, indicating that highly scored posts aren't necessarily more positively viewed, hinting that content that sparks strong reactions may garner higher engagement.

Number of Comments vs. Upvote Ratio (0.302917): More commented posts usually have higher upvote ratios, likely due to more engaged discussions being viewed positively.

Number of Comments vs. Mean Sentiment (-0.179709): More comments correlate with slightly more negative sentiments, possibly pointing to controversy or varied opinions fueling discussions.

Upvote Ratio vs. Mean Sentiment (0.075435): A very weak positive correlation suggests a minimal influence of sentiment on the likelihood of receiving upvotes.

Final Observation: The strong correlation between Score and Number of Comments highlights that more engaging content not only scores higher but also prompts more discussion. Meanwhile, the weak correlation between Mean Sentiment and metrics like Score and Number of Comments indicates that sentiment does not significantly straightforwardly influence engagement or visibility.

Question 2: Can we predict the memory size of a GPU based on its clock speed and other specifications?

Here, we will be using Linear Regression Model to define the Baseline and then would proceed to try complex models like Random Forests to improve the accuracy.

The following are the findings after implementation:

Linear Regression resulted in a slightly higher MSE of 11.2299, reflecting less precise predictions. Random Forest resulted in an improved MSE of 11.1901, indicating more accurate predictions compared to the linear model. While, in terms of R^2; both models showcased strong predictive capabilities of around 0.794, meaning they could explain about 79.4% of the variability in GPU memory sizes from the given features.

Final Observation: The analysis demonstrated that GPU memory size can be effectively predicted from technical specifications like clock speed. Random Forest marginally outperformed linear regression, suggesting a slight edge in handling complex relationships within the data.

Question 3: What are the major factors that determine the GPU clock speed?

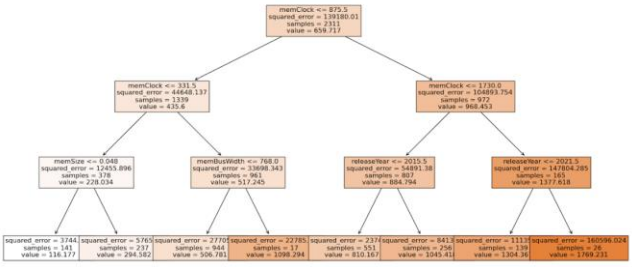

Fig 10. Decision Tree for GPU Clock Speed Predictors

Key Determinants:

memClock: With an overwhelming importance score of approximately 90.35%, memory clock speed is by far the most influential factor in determining the GPU clock speed. This makes sense as the speeds of the GPU core and memory are often correlated in the design of the GPU architecture.

releaseYear: The release year of the GPU accounts for about 6.03% of the importance, suggesting that more recent GPUs tend to have higher clock speeds due to advancements in technology and manufacturing processes.

memBusWidth: This feature has an importance score of approximately 2.45%, indicating a lesser but still relevant impact on the GPU clock speed. This could be related to the overall bandwidth capabilities of the GPU.

memSize: Surprisingly, memory size seems to have a very low influence on the GPU clock speed, with an importance score of about 1.18%. It appears that the amount of memory a GPU has does not significantly dictate its operating speed.

Model Performance:

Mean Squared Error (MSE): At 31539.807, the high MSE suggests that the model, while indicative of general trends, may not always align closely with actual clock speeds.

R^2 Score: With a score of 0.7812, the model explains about 78.12% of the variance in GPU clock speeds, demonstrating good predictive strength but also room for improvement by possibly incorporating additional features or using a more complex modeling approach.

Question 4: Which authors generate the most engaging content in terms of scores and comments?

Implementing K-means clustering to find which authors generate the most engaging content in terms of scores and comments:

Cluster 0: Moderately Engaging Authors

Contains the largest number of authors (108). Authors in this group have a moderate level of engagement, with scores and comments suggesting consistent but not exceptionally high interaction. An example author from this cluster is '5v73', with a score of 452 and 331 comments.

Cluster 1: Highly Engaging Authors

The smallest group with just 2 authors, 'Nestledrink' and 'Stiven_Crysis', stand out significantly in terms of engagement metrics. 'Stiven_Crysis' has an impressive score of 6309 and 3227 comments, indicating a strong ability to generate interest and discussion.

Cluster 2: Actively Engaging Authors

This cluster comprises 13 authors who are quite successful in engaging their audience. They exhibit high but not extreme levels of engagement compared to Cluster 1. For instance, 'East_Personality_771' has a score of 5103 with 841 comments.

Implementing K-means clustering :

```
Authors in cluster 0:
                  Author  Score  Number of Comments
0                    5v73    452                 331
1            AalbatrossGuy     44                 118
2    Adventurous_Time_227      4                   7
3            Agender_Azul      4                  54
4         Alarmed-Bad7994     17                  66
..                    ...    ...                 ...
116    sips_white_monster    694                 303
117           skyline385     47                  16
118         theacclaimed    120                 207
120          unixbhaskar     46                   3
121            wookmania    297                 452

[108 rows x 3 columns]
Authors in cluster 1:
          Author  Score  Number of Comments
49    Nestledrink   4054                2866
61  Stiven_Crysis   6309                3227
Authors in cluster 2:
                  Author  Score  Number of Comments
23   East_Personality_771   5103                 841
47   NamesTeddy_TeddyBear   3419                 604
48              Nekrosmas    900                1409
53         Old_Miner_Jack    831                1451
55         PapaBePreachin   1657                 628
56            Progenitor3   1281                 927
66        TheEternalGazed   1088                 990
77            Voodoo2-SLi   2594                1376
87          anotherwave1   1750                 576
92             chrisdh79   2325                 563
95              f0xpant5    907                1680
119     thebelsnickle1991   2520                 193
122             zer0_c0ol   1209                 539
```

Fig 11. Summary of K-mean clustering model

Final Observation: The analysis effectively highlights that while most authors generate a moderate level of engagement (Cluster 0), a small number of authors (Clusters 1 and 2) are particularly adept at creating content that significantly resonates with the audience, as evidenced by the high scores and number of comments they receive.

Question 5: How well can we classify GPUs based on their performance into categories such as low, medium, and high performance?

Here, SVM classifier is used to classify GPUs based on their performance into categories such as low, medium, and high performance?

The SVM Classifier

```
              precision    recall  f1-score   support

        high       0.61      0.73      0.67        15
         low       0.90      0.92      0.91       225
      medium       0.93      0.91      0.92       338

    accuracy                           0.91       578
   macro avg       0.82      0.86      0.83       578
weighted avg       0.91      0.91      0.91       578

[[ 11   0   4]
 [  0 207  18]
 [  7  22 309]]
```

Fig 12. Implementing SVM Classifier

Confusion Matrix:

High Performance: Out of 15 GPUs, 11 were correctly classified, and 4 were misclassified as medium.

Low Performance: Out of 225 GPUs, 207 were correctly identified, while 18 were misclassified as medium.

Medium Performance: Out of 338 GPUs, 309 were correctly classified, 22 were misclassified as high, and 7 as low.

Model Performance Summary:

High Performance GPUs:

Precision: 61% - This implies that when the model predicts a GPU as high performance, it is correct about 61% of the time.
Recall: 73% - This indicates that the model successfully identifies 73% of all actual high-performance GPUs.
F1-Score: 67% - The F1-score is a balance between precision and recall, providing a comprehensive measure of the model's accuracy for the high category.

Low-Performance GPUs:

Precision: 90% - Indicates very high accuracy in predicting low-performance GPUs.
Recall: 92% - The model captures 92% of all actual low-performance GPUs.
F1-Score: 91% - Demonstrates excellent model performance for the low category.

Medium Performance GPUs:

Precision: 93% - Suggests that the model is very reliable when it classifies a GPU as medium performance.
Recall: 91% - Reflects the model's ability to identify 91% of all true medium-performance GPUs.
F1-Score: 92% - A very high F1-score indicates strong model performance for the medium category.

Final Observation: The model can effectively classify GPUs into these performance categories with high accuracy. The results demonstrate that the selected features and the SVM classifier are well-suited for this task.

Question 6: Can we predict the release year of a GPU based on its technical specifications?

The Gradient Boosting Regressor has demonstrated excellent capability in accurately predicting the release years of GPUs based on their specifications.

Model Performance Metrics:

Mean Squared Error (MSE): An MSE of 2.866 suggests that, on average, the model's predictions deviate from the actual years by roughly the square root of 2.866, which is approximately 1.69 years. This is a relatively low error, indicating high prediction accuracy.

R^2 Score: An R^2 score of 0.927 means that approximately 92.7% of the variance in the GPU release years is explained by the model. This is an excellent score, suggesting that the model has a strong predictive power.

Final Observation: The analysis demonstrates that it is feasible to predict the release year of GPUs with high accuracy using their technical specifications. The Gradient Boosting model effectively utilized the given GPU specifications to forecast release years, as evidenced by the low MSE and high R^2 values.

Question 7: How do the specifications of GPUs vary by manufacturer?

memSize by Manufacturer:

NVIDIA and Intel show a broad range of memory sizes, with NVIDIA having some GPUs with significantly high memory.
AMD shows a less spread-out range but still considerable variation, with a few outliers indicating some GPUs with very high memory sizes.
Other manufacturers like ATI, Sony, Matrox, XGI, and 3dfx have GPUs with generally lower memory sizes.

gpuClock by Manufacturer:

NVIDIA and AMD GPUs exhibit a wide range of GPU clock speeds, with NVIDIA having a higher upper range, suggesting they have models with higher clock speeds.
Intel GPUs seem to have a more concentrated range of clock speeds, not reaching as high as NVIDIA or AMD.
ATI and other manufacturers tend to have lower clock speeds in comparison.

memClock by Manufacturer:

AMD GPUs have a relatively high memory clock, also with a wide range.
NVIDIA also shows a broad range but with many models having a lower memory clock than AMD.
Intel, ATI, and other manufacturers show less variation and generally lower memory clock speeds.

unifiedShader by Manufacturer:

NVIDIA stands out with the highest range of unified shader counts, including GPUs with extremely high counts.
AMD has a significant number of GPUs with a high count of unified shaders but not as high as NVIDIA.
Intel and ATI have lower counts of unified shaders, and other manufacturers have very few or constant values which may indicate older or less complex GPUs.

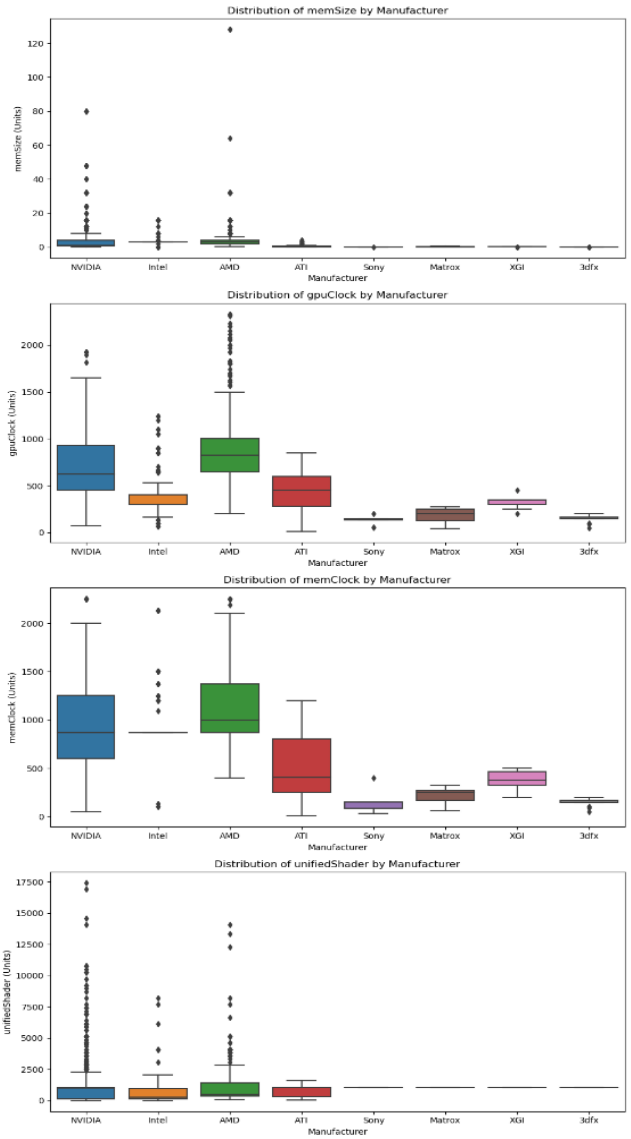

Fig 11. Variation of GPU specification according to Manufacturer Final Observation: NVIDIA and AMD are the leaders in high-spec GPUs with NVIDIA taking the edge in memory size and shader count, while AMD leads in memory clock speed.

Intel GPUs are generally lower spec compared to NVIDIA and AMD, with lower clock speeds and shader counts. ATI, Sony, Matrox, XGI, and 3dfx are much less varied and tend to have lower specifications across the board.

Question 8: Can machine learning models identify trends in GPU development over the years?

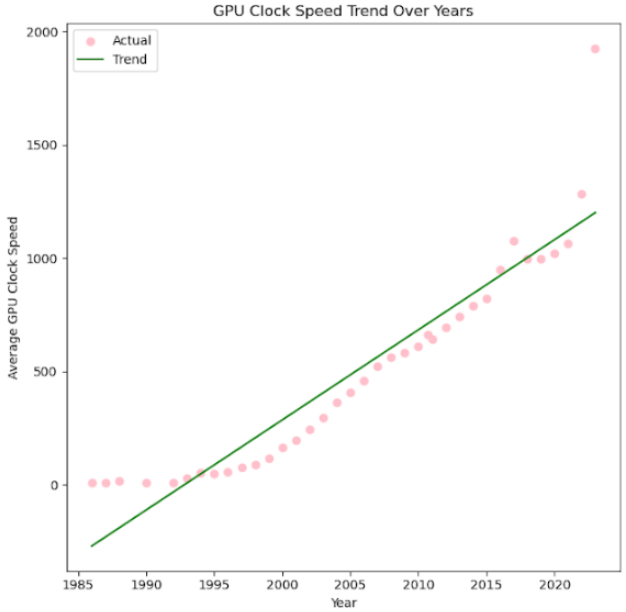

Fig 12. Trend of Average GPU clock speed with time

GPU Clock Speed Trend Over Years:

The trend line for GPU clock speeds indicates a clear upward trajectory from 1985 to the present. The actual data points largely follow this trend, with occasional years where the average clock speeds dip below the trend line. This upward trend suggests that manufacturers have been consistently improving the clock speed of GPUs over time, which is likely due to advancements in technology and manufacturing processes.

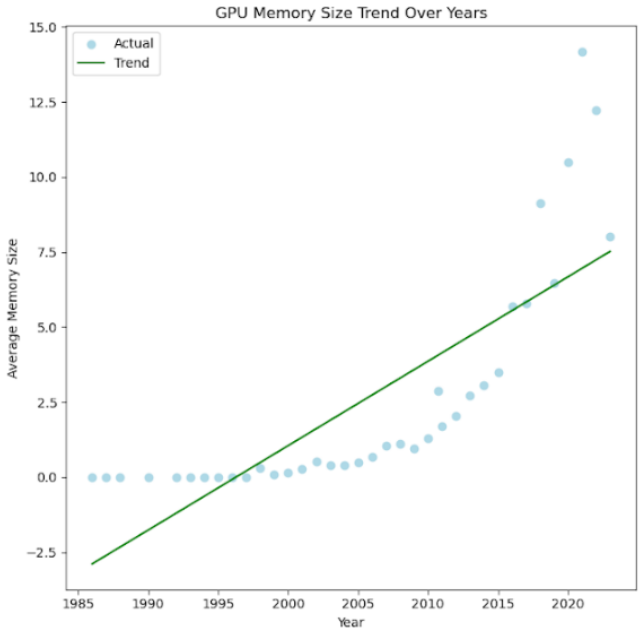

Fig 13. Trend of Average Memory size with time

GPU Memory Size Trend Over Years:

The memory size trend also shows a positive slope, indicating an increase in average GPU memory size over the years. The actual data points show that in recent years, there has been a significant increase in memory size, which could be due to the growing demand for more graphics-intensive applications and games that require larger amounts of memory. Some outliers are well above the trend line, likely representing high-end GPUs that are outliers to the general progression of GPU memory sizes.

Final Observation: A simple Linear Regression has identified clear trends in the development of GPUs over the years. Both GPU clock speeds and memory sizes have been increasing. This suggests that as technology has advanced, manufacturers have been able to produce GPUs with better performance specifications. For clock speeds, the steady increase aligns with improvements in chip design and manufacturing that allow for faster processing without overheating. For memory sizes, the demand for higher resolutions and more detailed textures in digital graphics likely drives the need for more memory.

Question 9: Principal Component Analysis on Numerical Attributes

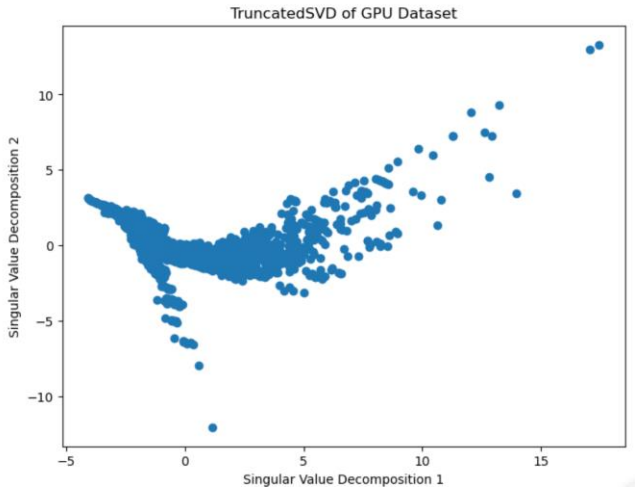

Fig 14. Truncated SVD of GPU Dataset

Final Observation on PCA:

The first two singular value components account for approximately 42.21% and 15.24% of the variance in the data, respectively. This suggests that while a fair amount of the variance is captured by these components, there are likely more dimensions (features) that contribute significantly to the dataset's variability since the two components together do not account for the majority of the variance. The scatter plot showcases the distribution of the GPUs in this reduced dimensionality space, revealing patterns and potential clusters. The distribution along the first component particularly seems to spread the data out well, indicating it captures a significant pattern in the dataset.

Question 10: Does the time of day or day of the week when a post is made affect its visibility and engagement level?

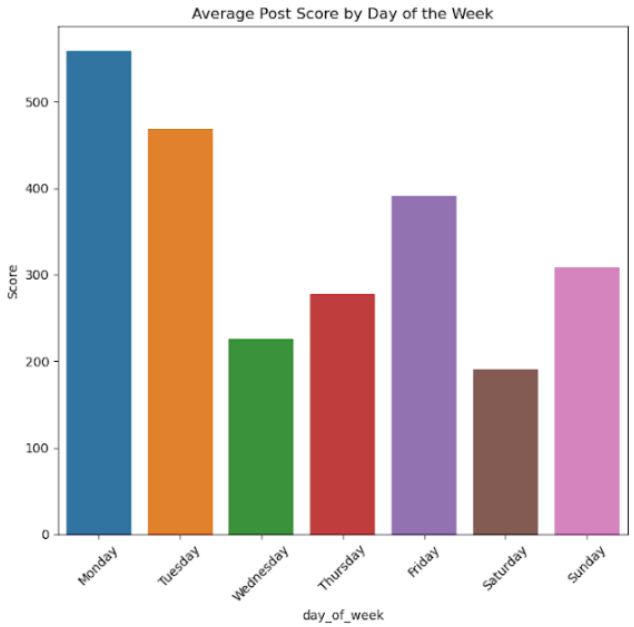

Fig 15 a

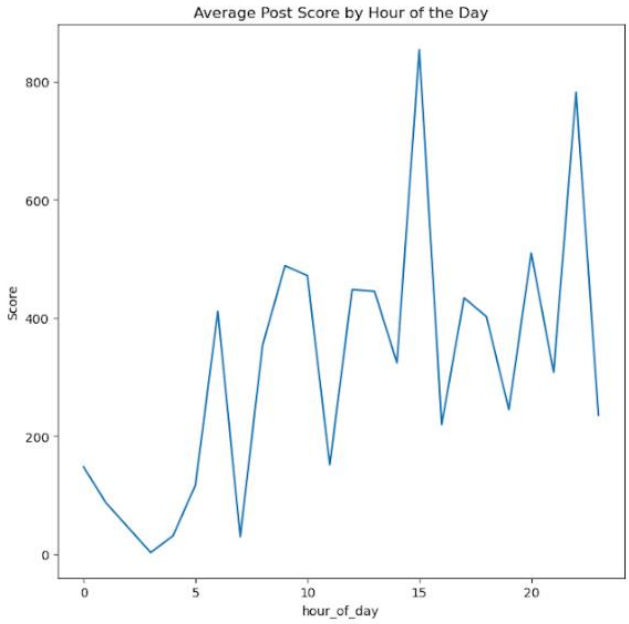

Fig 15 c

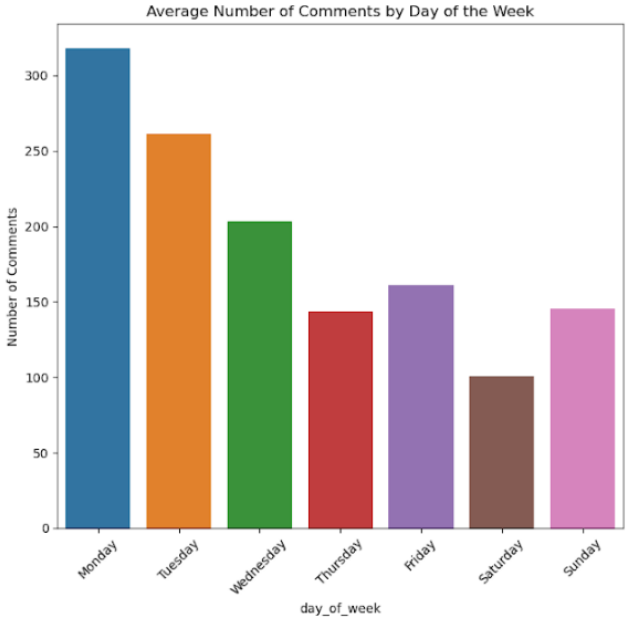

Fig 15 b

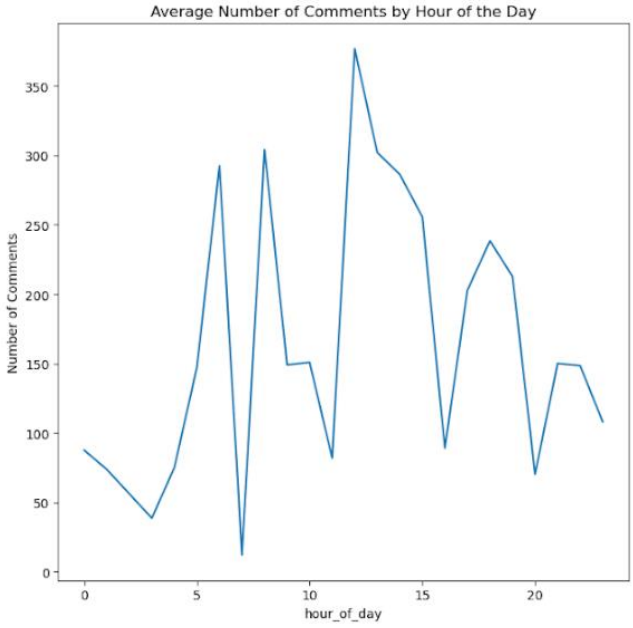

Fig 15 d

Fig 15. Variation of variables with respect to time as well as days in the weeks

Final Observations:

The charts comparing the average post score and number of comments by the day of the week suggest that there are certain days when posts receive more attention. Similarly, the graphs showing the average post score and number of comments by the hour of the day indicate that posts made during certain hours tend to garner more engagement. The data suggests that posts made on Mondays and Tuesdays receive a higher average score and more comments compared to other days of the week, indicating there might be optimal days for posting. In terms of time, posts made in the early morning hours appear to receive less engagement, while those posted in the late evening receive higher scores and comments, pointing towards more active user engagement during these hours. Content creators or social media managers could leverage this information to schedule their posts when the audience engagement is typically higher, thereby increasing the chances of their content being seen and interacted with. This strategic timing could be a valuable part of an effective social media strategy.

Question 11: Discovering common topics or themes within the Submission Text or Comment Body using NLP techniques such as Latent Dirichlet Allocation (LDA) or Non-negative Matrix Factorization (NMF).

```
Top words for topic #0:
['reviews', 'rx', 'xt', 'youtube', 'watch', 'amd', 'review', 'com', 'www', 'https']

Top words for topic #1:
['78', '80', 'www', 'ampere', 'raster', 'https', 'geforce', 'rtx', '4070', '100']

Top words for topic #2:
['games', 'performance', '8c', '4090', 'ryzen', 'www', 'https', 'nbsp', '_100', '100']

Top words for topic #3:
['ryzen', 'cpu', 'one', 'pc', 'fps', 'card', 'ddr4', '4080', 'like', '5800x3d']

Top words for topic #4:
['product', 'also', '4070', 'rtx', 'x200b', 'pcpartpicker', 'cpu', 'gaming', 'com', 'https']
```

Fig 16. Top word according to topics

Topic #0 - Reviews and Media Content: This topic focuses on reviews and content consumption, highlighting words such as "reviews," "YouTube," "watch," and "review." It suggests a strong interest in video reviews and information sourced from websites, as indicated by the frequent mentions of URLs.

Topic #1 - New GPU Technologies: This topic is centered around new technologies, particularly Nvidia's GPUs, with terms like "ampere" (Nvidia's architecture), "GeForce," "RTX," "4070," and performance metrics like "raster." It indicates discussions focused on the latest hardware specifications and performances.

Topic #2 - High-Performance Computing and Gaming: Featuring terms like "games," "performance," "4090," and "Ryzen," this topic appears to cover high-performance applications, including gaming and possibly discussions around comparisons of top-tier GPUs and CPUs.

Topic #3 - Hardware Configurations and User Experiences: With terms such as "ryzen," "cpu," "pc," "ddr4," "4080," and "like," this topic deals with personal computing setups, hardware configurations, and user experiences, particularly concerning gaming and system builds.

Topic #4 - Product Information and Recommendations: This topic includes references to specific products ("4070," "rtx") and resources ("pcpartpicker") as well as general terms about gaming and hardware ("gaming," "cpu"). It suggests a focus on product recommendations and building gaming setups.

4 Conclusion and Results

We will be listing our Results in the following bullet points:

- The strong correlation between Score and Number of Comments highlights that more engaging content not only scores higher but also prompts more discussion. Meanwhile, the weak correlation between Mean Sentiment and metrics like Score and Number of Comments indicates that sentiment does not significantly straightforwardly influence engagement or visibility.

- Effective Prediction: The analysis demonstrated that GPU memory size can be effectively predicted from technical specifications like clock speed.

- Model Preference: Random Forest marginally outperformed linear regression, suggesting a slight edge in handling complex relationships within the data.

- The analysis effectively highlights that while most authors generate a moderate level of engagement (Cluster 0), a small number of authors (Clusters 1 and 2) are particularly adept at creating content that significantly resonates with the audience, as evidenced by the high scores and number of comments they receive.

- The model can effectively classify GPUs into these performance categories with high accuracy. The results demonstrate that the selected features and the SVM classifier are well-suited for this task

- The analysis demonstrates that it is feasible to predict the release year of GPUs with high accuracy using their technical specifications. The Gradient Boosting model effectively utilized the given GPU specifications to forecast release years, as evidenced by the low MSE and high R^2 values.

- NVIDIA and AMD are the leaders in high-spec GPUs with NVIDIA taking the edge in memory size and shader count, while AMD leads in memory clock speed.

- Intel GPUs are generally lower spec compared to NVIDIA and AMD, with lower clock speeds and shader counts. ATI, Sony, Matrox, XGI, and 3dfx are much less varied and tend to have lower specifications across the board.

- For clock speeds, the steady increase aligns with improvements in chip design and manufacturing that allow for faster processing without overheating.

- For memory sizes, the demand for higher resolutions and more detailed textures in digital graphics likely drives the need for more memory.

- The first two singular value components account for approximately 42.21% and 15.24% of the variance in the data, respectively. This suggests that while a fair amount of the variance is captured by these components, there are likely more dimensions (features) that contribute significantly to the dataset's variability since the two components together do not account for the majority of the variance.

- The scatter plot showcases the distribution of the GPUs in this reduced dimensionality space, revealing patterns and potential clusters. The distribution along the first component particularly seems to spread the data out well, indicating it captures a significant pattern in the dataset.

In conclusion, we offer a comprehensive summary of the insights and outcomes derived from our detailed analysis of GPUs and processors through the lens of advanced data mining techniques. Our project tapped into a rich dataset sourced from Reddit, applying a variety of analytical methods to decode complex data into understandable trends and actionable insights.

The core of our analysis involved sophisticated data mining techniques including sentiment analysis, linear regression, Random Forests, Decision Trees, Principle Component Analysis (PCA), and various Natural Language Processing (NLP) methods such as Latent Dirichlet Allocation (LDA) and Non-negative Matrix Factorization (NMF). These techniques allowed us to tackle complex questions ranging from predicting GPU memory sizes based on technical specifications

to understanding consumer sentiments expressed in online reviews.

Key insights from our project include the identification of sentiment trends among various GPU models, where we could observe clear consumer preferences and satisfaction levels. For instance, our sentiment analysis revealed a generally positive sentiment toward AMD's Ryzen processors, whereas some high-end GPUs from other manufacturers showed varied sentiment scores, potentially reflecting users' expectations and the perceived value of these products.

Moreover, our predictive models successfully estimated GPU attributes such as memory size and clock speed, demonstrating how well certain technical specifications can forecast product characteristics. This not only highlights the predictive power of machine learning models but also provides manufacturers with valuable feedback on what technical features are most influential in user experiences.

Our analysis also extended to understanding the dynamics of content engagement on Reddit, where we explored how the timing of posts affects visibility and interaction. This aspect of the study offers useful insights for marketers and content creators who aim to optimize their engagement strategies based on user activity patterns.

In summary, this project exemplifies the potent application of data science techniques in navigating through complex, unstructured datasets to extract meaningful information. It provides a bridge between technical data and consumer insights, thereby enhancing decision-making processes for both users and manufacturers in the tech industry. Our conclusions not only reinforce the importance of analytical rigor in tech studies but also highlight the practical benefits of diverse data analysis methods in real-world applications.

From the thorough conclusion above, many more analyses could be done in the future. Here are several areas where future work could enhance and expand upon the initial project:

1. Extended Sentiment Analysis: Future projects could incorporate more complex sentiment analysis models that consider context and the subtleties of language used in reviews and comments. This would provide a deeper understanding of user emotions and could identify specific features or issues that influence consumer sentiment more significantly.

2. Real-Time Data Analysis: Implementing real-time data mining and analysis could track trends as they develop, offering more timely insights into consumer opinions and market dynamics. This could be particularly useful for identifying the impact of new product releases or marketing campaigns on public perception and engagement.

3. Cross-Platform Analysis: Expanding the dataset to include discussions from other social media platforms or online forums could offer a more comprehensive view of consumer opinions and broaden the scope of engagement analytics. This would enhance the understanding of how different platforms cater to different demographics and how that affects sentiment and engagement.

4. Predictive Model Enhancement: Future work could explore more advanced machine learning models or deep learning approaches to improve the accuracy of predictions regarding GPU performance and technical specifications. Incorporating additional predictors or using ensemble methods might also enhance model robustness.

5. Technological Advancements Impact: Tracking how advancements in technology, like AI accelerators or new manufacturing processes, impact GPU and processor performance could be another exciting avenue. This could involve predictive analytics to forecast future technology trends based on past data.

6. Personalized Recommendations: Building on the classification and clustering work, future projects could develop systems for personalized product recommendations based on user preferences and sentiment. This could be particularly useful for retailers or manufacturers in tailoring their offers and marketing strategies.

7. Economic and Market Analysis: Integrating economic data such as pricing, sales volume, and market share could provide insights into how these factors correlate with technological advancements and consumer sentiment. This would be useful for strategic planning and market positioning.

8. User Engagement Optimization: Further analysis on the optimal times for posting and engagement could be refined by experimenting with different content types and strategies. Additionally, A/B testing on post characteristics could scientifically measure the impact of different variables on user engagement.

These potential future insights and project directions not only extend the research but also offer practical applications that could benefit various stakeholders in the tech industry, from consumers and content creators to manufacturers and marketers.

ACKNOWLEDGMENTS

I want to express my sincere gratitude to the University of Colorado Boulder for giving us the chance to present our research. I am grateful to my Data Mining Professor, Alphonso Bastias, for his invaluable guidance and support throughout this project. Finally, I would like to thank my classmates for their collaboration and insights which have greatly enriched this research experience.

REFERENCES

[1] Jhaveri, D., Chaudhari, A. and Kurup, L., 2015. Twitter sentiment analysis on e-commerce websites in India. International Journal of Computer Applications, 127(18), pp.14-18.

[2] "Understanding Support Vector Machine algorithm from examples (along with code)." [Online]. Available: https://www.analyticsvidhya.com/blog/2017/09/understaingsupport-vector-machine-example-code/. [Accessed: 01-Dec2018].

[3] "Sentiment Analysis | Lexalytics." [Online]. Available: https://www.lexalytics.com/technology/sentiment. [Accessed: 24- Nov-2018].