Reddit Sentiment Analysis and Readability

Chenzheng Li
School of Computing Science
Simon Fraser University
8888 University Dr, Burnaby, BC
cla429@sfu.ca

Eric Chan
School of Computing Science
Simon Fraser University
8888 University Dr, Burnaby, BC
eca104@sfu.ca

Ziying Peng
School of Computing Science
Simon Fraser University
8888 University Dr, Burnaby, BC
ziyingp@sfu.ca

Abstract

This study serves as the final project for the Computational Data Science (CMPT353) course during the Summer 2023 term at Simon Fraser University's (SFU) School of Computing Science, under the instruction of Professor Greg Baker. Our project involves the exploration of key factors influencing the perceived quality of Reddit submissions. We hypothesize that elements such as publication timing, sentiment analysis, and readability scores can play a significant role. By leveraging data analysis and machine learning techniques on a large Reddit dataset, we aim to gain insights into these factors and how they influence user engagement, measured by likes. Not only does this study aspire to enhance understanding of user-generated content, but it also seeks to provide constructive insights for platform developers and content creators aiming to enhance user experiences. Preliminary results suggest a possible correlation between these factors and post perception, promising further exploration and application in enhancing content engagement.

1 Introduction

In today's digital era, platforms facilitating user-generated content have proliferated. These platforms offer a space for users to express their creativity and engage with a global audience. However, the quality of these submissions varies widely, prompting the question: What factors contribute to a high-quality submission?

In our study, we aim to delve into this issue, positing that elements such as the timing of the submission, the sentiment expressed, and its readability score could potentially impact its perceived quality. While likes, comments, and shares often serve as indicators of a post's popularity, they might not fully capture the perception of its quality.

Our project's goal is not just to examine this issue through various data analysis techniques, but also to offer valuable insights to platform developers and content creators looking to enhance their engagement and user experiences. Hence, our refined problem statement is:

We endeavor to discern the influence of factors such as publication time, sentiment, and readability on user perception of a post. Perception is to be assessed via tangible indicators like likes."

We provide all the code files for this project for reference, and the process and instructions for running the code in its entirety are available on github[1].

School of Computing Science in Simon Fraser University (2023).

Figure 1 show the whole workflow in this project.

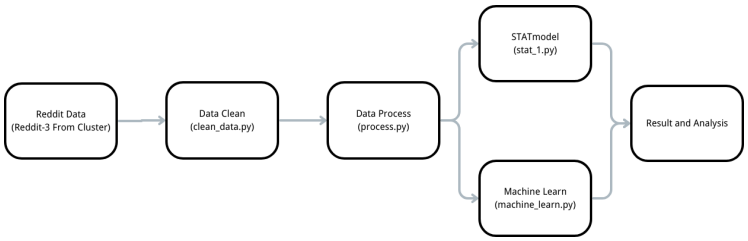

Figure 1: Project Work Flow

2 Data Acquisition and Preprocessing

2.1 Data Acquisition

The 'reddit3' dataset is a rich corpus of text data derived from Reddit, an extensively used online forum permitting users to engage in posts, comments, and voting on an array of topics. This specialized dataset, sourced from a JSON file stored in Simon Fraser University's (SFU) cluster, is concentrated explicitly on five medium-volume subreddits for the year 2016 and encompasses a total size of 790MB. The dataset furnishes diverse information encompassing comment content (body), author, creation time (created_utc), score, subreddit affiliation, upvotes (ups), and additional facets, offering a comprehensive view of the interactions and engagements on these particular Reddit threads.

2.2 Data Set

For different replies in different reddits we can get these different labels.

Data Preprocessing The data preprocessing primarily involves two parts: data cleaning (handled by the clean_data.py script) and data processing (handled by the process.py script). I will now detail these scripts individually.

2.3 Data Cleaning (clean_data.py)

The aim of data cleaning is to eliminate irrelevant information and invalid data. In our project, we conducted the following data cleaning procedures:

Removal of Deleted Comments: We eliminated comments from the dataset where the author or the comment body was marked as [deleted]. This is because such comments do not offer substantial content and are not helpful to our analysis.

Removal of Edited Comments: We also eliminated comments that were marked as edited. This is because these comments may have been modified by their authors and may no longer reflect the authors' initial sentiments or views.

Timestamp Addition: We transformed the created_utc from the raw data into a timestamp. The PySpark from_unixtime function was used for this transformation, which transforms Unix timestamps into human-readable datetime formats.

Day Type Addition: We added a new column, daytype, to indicate whether a comment was posted on a weekend or a weekday. This was accomplished using PySpark's date_format and when functions.

Day of the Week Addition: We also added a new column, day_of_week, to specify the day of the week the comment was posted. Here 1 represents Monday and 7 represents Sunday. This was again accomplished using PySpark's date_format and when functions.

2.4 Data Processing (process.py)

The goal of data processing is to add new features to support data analysis and modeling. In our project, we conducted the following data processing steps:

Attribute	Attributes Description
1	author
2	body
3	timestamp
4	score
5	subreddit
6	ups
7	daytype
8	day_of_week

Table 1: List of Attributes in the dataset after clean

Sentiment Score Calculation: We used the VADER model from the NLTK library to calculate a sentiment score for each comment. The VADER model is designed for text sentiment analysis and can handle complex textual contexts such as negation, intensification, emoticons, slang, etc. The detailed calculation can be referred to in the VADER model's original paper[1].

Readability Score Calculation: We used the textstat library to compute a readability score for each comment. We chose the Flesch-Kincaid Grade Level as our readability metric, which is a commonly used readability rating assessing text complexity. The specific calculation can be referred to on the Wikipedia page of the Flesch-Kincaid Grade Level[2].

Comment Quality Calculation: We added a new column quality to determine the quality of comments. The quality of a comment was determined based on its score percentile within its subreddit. Comments with scores above the 90th percentile were labeled as "good", comments below the 10th percentile and those with scores less than -5 were labeled as "bad", and the rest were labeled as "normal"[Figure 2].

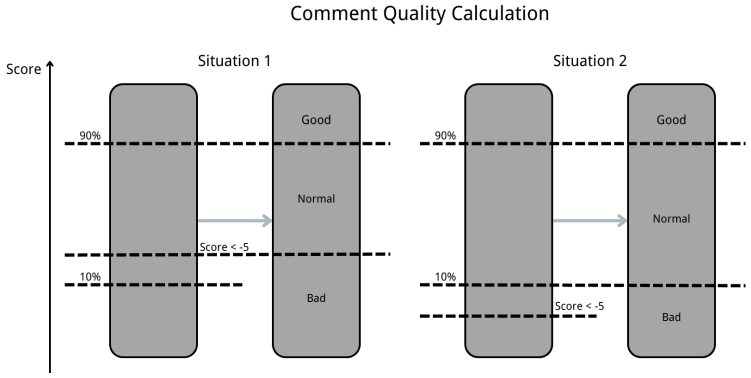

Figure 2: Comment Quality Calculation

With these steps, we ensured that the raw data was cleaned and transformed effectively to perform subsequent analysis and generate meaningful insights.

3 Methodology

In our research, we utilized a blend of qualitative and quantitative methodologies, primarily using Python's NLTK (Natural Language Toolkit) library and textstat package to carry out sentiment analysis and readability assessment, respectively. Both these steps are essential for our data analysis and feed into our overall goal of understanding the elements influencing user perception of a Reddit post.

3.1 Sentiment Analysis with NLTK VADER

VADER (Valence Aware Dictionary and sEntiment Reasoner) is a model incorporated within the NLTK library. It is lexicon and rule-based, specifically built to analyze sentiments in social media text, making it a perfect fit for our Reddit data.

Given the size of our dataset (multiple terabytes of comments), performing sentiment analysis on local machines was infeasible due to computational and memory constraints. Thus, we leveraged Simon Fraser University's (SFU) high-performance computing cluster to run our analysis. The distributed nature of this cluster allows the processing of large-scale data in a time-efficient manner.

The VADER model outputs a compound score for each comment ranging from -1 (most negative) to +1 (most positive). This score reflects the overall sentiment embodied within the comment. This analysis will help us determine whether the sentiment expressed in a comment has any bearing on how it is perceived by the Reddit community.

3.2 Readability Assessment with Textstat

To analyze the readability of Reddit comments, we employed the textstat library. More specifically, we used the Flesch-Kincaid Grade Level metric. This readability test measures the complexity of an English text. The output is a number that corresponds to a U.S. school grade level, indicating the level of education needed to understand the text.

Again, given the volume of our dataset, running this readability assessment on a local machine would be impractical. By utilizing SFU's cluster, we could effectively handle the processing load and carry out this analysis in a feasible time frame.

The output of this readability score can provide insights into whether the complexity of the comment's text influences its reception by users. For example, a comment that is easier to read (lower grade level) might be more appreciated by the general Reddit user base and thus, could potentially have a higher score.

To summarize, our methodology included sentiment analysis using NLTK's VADER model and readability assessment using textstat's Flesch-Kincaid Grade Level metric. Running these analyses on SFU's cluster enabled us to handle our vast dataset and yield meaningful insights that can contribute to understanding the elements affecting user perception of Reddit posts.

Consider the following two Reddit comments:

Comment A: "This subject is of profound importance to our understanding of space and time."

Comment B: "Space and time are like super cool, dude."

If we use the Flesch-Kincaid Grade Level metric of Textstat to analyze these sentences, we might get the following results:

Comment A: Flesch-Kincaid Grade Level = 12 Comment B: Flesch-Kincaid Grade Level = 3 The Flesch-Kincaid Grade Level corresponds to the US educational grade level that would comprehend the sentence. In this case, Comment A, with its complex vocabulary ("profound", "importance"), requires a 12th-grade reading level for comprehension, whereas Comment B, with its informal and simple language, only requires a 3rd-grade reading level.

This illustrates how the readability score can vary based on the complexity of language used in the text. These scores can then be analyzed in relation to the reception of the comments by Reddit users. A hypothesis might be that comments with a lower grade level (easier to read) receive more upvotes because they are more accessible to a broad range of users. However, it's also plausible that more complex comments could be appreciated in certain subreddit communities. These are the types of patterns and insights that our analysis aims to uncover.

3.3 Ordinal Logistic Regression Analysis with StatsModels

In order to understand the relationship between comment sentiment, readability, and the quality score a comment receives, we performed an ordinal logistic regression analysis. For this purpose, we utilized the statsmodels library, a powerful and flexible Python module for statistical modeling.

Like the sentiment analysis and readability assessment, this regression analysis was also carried out on SFU's high-performance computing cluster. This was due to the computational requirements of running ordinal logistic regression on our large dataset.

In the analysis, the dependent variable was comment 'quality', which we categorized into three ordered categories: 'bad', 'normal', and 'good'. Our independent variables were the sentiment score derived from VADER, the readability score from the Flesch-Kincaid Grade Level metric, and the day type (weekend or weekday). To capture these in the model, we used the formula interface provided by the patsy library, another Python module for describing statistical models.

The ordinal logistic regression model was fit using the 'bfgs' optimization method. The results from the model indicate the effects of the independent variables on the likelihood of a comment being categorized as 'bad', 'normal', or 'good'. These findings will give us insights into how the sentiment and readability of a comment, as well as the day of the week it was posted, influence the perception of comment quality by the Reddit community. †

3.4 Machine Learning

Given the enormous size of the dataset acquired from Reddit comments, traditional statistic techniques can sometimes fall short of capturing the underlining patterns and relationships. This is where machine learning becomes a suitable choice. Machine learning models are designed to learn from data, especially big data, and can uncover both linear and non-linear relationships. Those machine-learning qualities make them particularly good for handling complex datasets with potentially hidden structures.

From the variety of models in the Machine learning field, we chose the MLP Classifier for the training. MLP is a neural network designed to figure out non-linear relationships. Given that our dataset contained various diverse variables like sentiment, readability, and day type, the model needs the ability to handle non-linear features. MLP's layered structure is designed to navigate such complex relations, making it a natural fit. Additionally, MLP is robust in managing high-dimensional data, ensuring that the breadth and depth of the Reddit dataset are effectively handled.

To thoroughly understand the factors influencing comment quality on Reddit, we intend to systematically explore different feature combinations. The dependent variable (Y) remains consistent with the comment "quality." However, our independent variable(s) (X) will vary across multiple analyses:

Y axis: always Quality X axis: 1. sentiment_score, 2.readability_score, 3.day type, 4. all 3 combined

Sentiment Score: By mapping sentiment scores (derived from comments) against their corresponding quality metrics, we anticipate illuminating insights into whether positive or negative sentiments correlate with higher or lower comment quality.

Readability Score: We surmise that comments which are easier to read and comprehend might be associated with better quality. This metric will give us an idea of the text's complexity and its potential correlation with quality.

Day Type: Day type (weekdays vs. weekends and it will be encoded to 1 and 0 since MLP can't handle category in X asix) might influence comment quality. This could be attributed to various factors like user activity patterns, post frequency, or even topical relevance based on the day.

Combined Features: By combining all the features - sentiment score, readability score, and day type, the model can capture the integrated view of how these factors interplay to determine the quality of Reddit comments.

4 Results and Key Findings

4.1 Statistical Analysis

4.1.1 Part 1 of Statistical Analysis

The Odds Ratios with 95% Confidence Intervals plot(Figure 3) below illustrates the odds ratios with their corresponding 95% confidence intervals for each predictor variable in our ordered logistic regression model.

Sentiment Score: The point estimate for the sentiment score is around 0.75. This suggests that for each unit increase in the sentiment score, the odds of moving up a level in the quality category decrease by roughly 25%, assuming all other variables are held constant. However, because the confidence interval likely spans 1, we cannot make definitive conclusions about the effect of sentiment score on the quality.

Readability Score: The point estimate for the readability score is approximately 1. This suggests that changes in the readability score do not significantly alter the odds of moving up a level in the quality category, assuming all other variables are held constant.

Daytype: The point estimate for the daytype is around 0.9. This indicates that the odds of moving up a level in the quality category are slightly less on weekends compared to weekdays, assuming all other variables are held constant. However, similar to the sentiment score, we cannot make definitive conclusions about the effect of daytype on the quality due to the confidence interval likely spanning 1.

0/1 Threshold: The point estimate for the 0/1 threshold is approximately 0.6. This is the estimated odds ratio for moving from the 'bad' to the 'normal' quality category.

1/2 Threshold: The point estimate for the 1/2 threshold is around 2.7. This is the estimated odds ratio for moving from the 'normal' to the 'good' quality category. Given its point estimate far from 1, it suggests that the odds of moving from the 'normal' to the 'good' quality category are significantly higher than moving from the 'bad' to the 'normal' category, assuming all other variables are held constant.

(Note: these interpretations depend on the confidence intervals not spanning 1 and would change if they do.)

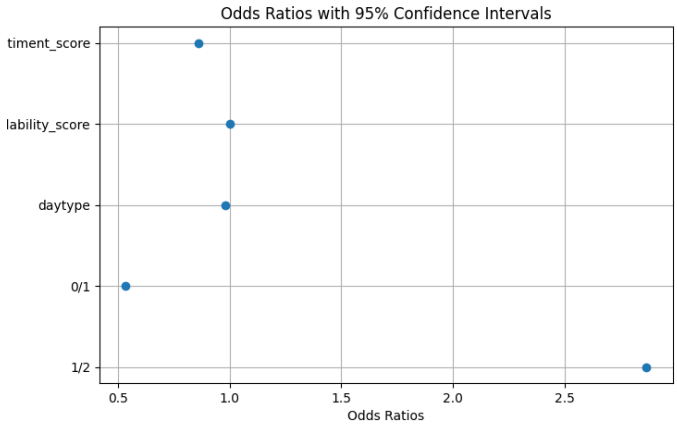

Figure 3: Odds Ratios with 95% Confidence Intervals

The pairplot below(Figure 4) provides insights into the distribution of each predictor and their relationships with the categorical response variable "quality".

Looking at the diagonal plots, the "sentiment_score" variable appears to follow a near-normal distribution for each category of "quality". The peak for the "normal" category (orange) is the highest, followed by "bad" (blue), and then "good" (green). This suggests that there is a greater likelihood of comments having a "normal" sentiment score, and comments with extremely positive or negative sentiment scores are less likely to occur than comments having a "normal" sentiment score.

The plot for "readability_score" exhibits an unusual behavior where it seems like a single, vertical line. This might indicate that this variable has an extremely common value in the sample. It is also possible that there are some outliers that skewed the distribution. For the further analysis, we will figure out what happens here.

The "daytype" variable, being binary, shows two separate normal distributions at values 0 and 1(0 for weekdays and 1 for weekends). The distribution at 0 is much taller than the one at 1, indicating that

there are more comments made on weekdays than on weekends. Among these, the "normal" category dominates, followed by "bad", and then "good".

This preliminary analysis suggests that sentiment scores, readability scores, and the day of posting all have significant associations with the quality of Reddit comments.

"The scatter plot depicting the interaction between 'readability_score' and 'sentiment_score' provides interesting insights. It is a 2D scatter plot, where the color of each point corresponds to its 'quality' category. The observed pattern, characterized by green dots ('good' quality comments) primarily clustering near a sentiment score of 0, fewer orange dots ('normal' quality comments), and blue dots ('bad' quality comments) scattered more broadly, indicates a potential interaction effect between the 'quality' category and these two predictor variables.

Most of the readability scores for the comments are concentrated at a relatively lower level. This could suggest that Reddit comments in the dataset typically have lower readability. Furthermore, high-quality ('good', green) comments generally have sentiment scores closer to 0, implying that they tend to exhibit a more neutral sentiment. On the other hand, low-quality ('bad', blue) comments display a more dispersed sentiment score, suggesting that these comments may show more extreme sentiments.

Notably, the presence of outliers, observed near 12500 and 5000, might be skewing the distribution of the 'readability_score'. This might explain the unusual distribution seen in the histogram of 'readability_score' on the diagonal. Such outliers can significantly impact the analysis, particularly if the assumption is that the data follows a normal distribution.

The scatter plot depicting the interaction between 'sentiment_score' and 'daytype' suggests that, for high-quality ('good') comments, the sentiment scores are widely distributed on both weekdays and weekends. The distribution appears similar in both cases. In other words, the day of posting (weekday or weekend) does not seem to significantly impact the sentiment scores of high-quality comments.

The scatter plot of 'readability_score' and 'daytype' also reveals intriguing patterns. This plot is similar to the previous 'sentiment_score' and 'daytype' plot that consists of two parallel lines of points at daytype 0 (weekdays) and 1 (weekends).

The color distribution, representing comment quality categories, appears to be similar for both daytypes. This might suggest that the quality and readability of the comments do not vary significantly between weekdays and weekends. In other words, the quality category (good, normal, or bad) does not seem to change depending on whether the comment was posted on a weekday or a weekend. Also noticed that there is a distinct outlier noticed at a readability score near 15000 and daytype 1.

4.1.2 Part 2 of Statistical Analysis

Here we use the ordinal logistic regression model to analyze the influence of sentiment score, readability score, and day type(weekday/weekend) on the quality of Reddit comments. The analysis here uses the dataset introduced above. There are 5,100,324 observations that have been used.

From the detail of the ordered model(Table 2) below, we can extract information: The Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC) values were approximately 9.358e+06 for both. These values take into account the goodness of fit and the complexity of the model. Lower values are preferred, indicating a better model.

Form the ordinal logistic regression analysis(Table 3) yielded the following results: The sentiment score showed a significant negative impact on the quality of Reddit comments. The coefficient of -0.1527 suggests that with an increase of one unit in sentiment score, the log-odds of a comment being in a higher quality category decreases by 0.1527, holding all other factors constant. This implies that comments with higher sentiment scores are less likely to be of higher quality.

The readability score, although also negatively associated with comment quality, has a smaller effect (coefficient = -0.0007). As the readability score increases by one unit, the log-odds of a comment being in a higher quality category decreases by 0.0007, holding all other variables constant. However, given the small coefficient value, the impact of readability score on comment quality is minimal.

The day type also showed a negative relationship with comment quality. The coefficient of -0.0211 suggests that comments posted on certain days (perhaps weekends) are likely to be of lower quality.

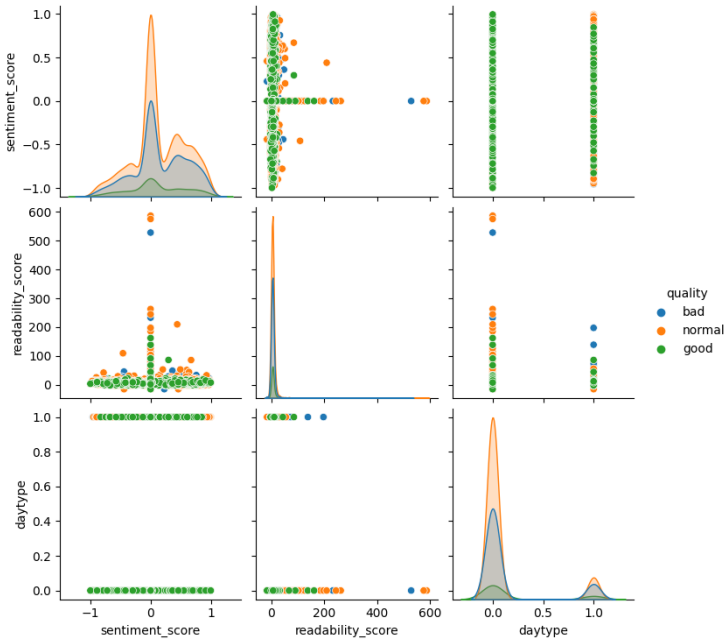

Figure 4: Correlation between Sentiment_score and Readability_score and daytype

Attribute	Description
Dep. Variable	y
Log-Likelihood	-4.6791e+06
Model	OrderedModel
AIC	9.358e+06
Method	Maximum Likelihood
BIC	9.358e+06
Date	Wed, 02 Aug 2023
Time	00:09:39
No. Observations	5100324
Df Residuals	5100319
Df Model	3

Table 2: OrderedModel Details

The coefficients for the thresholds (0/1 and 1/2) were -0.6260 and 1.0529, respectively. These represent the estimated cut points that divide the categories of the dependent variable, thus helping model the ordinal nature of the dependent variable.

Overall, these results suggest that sentiment and readability scores, along with the day of posting, have significant impacts on the quality of Reddit comments. As these factors increase, the quality of comments, as categorized in this study, seems to decrease. However, further analysis might be needed to confirm these findings and to explore additional factors influencing comment quality.

Attribute	Coefficient	Standard Error	z	P>|z|	[0.025	0.975]
Sentiment Score	-0.1527	0.002	-88.253	0.000	-0.156	-0.149
Readability Score	-0.0007	8.23e-05	-8.552	0.000	-0.001	-0.001
Daytype	-0.0211	0.002	-10.419	0.000	-0.025	-0.017
0/1	-0.6260	0.001	-518.453	0.000	-0.628	-0.624
1/2	1.0529	0.001	1909.400	0.000	1.052	1.054

Table 3: OrderedModel Results

4.2 Machine Learn

Different than the prediction that different input sets will train a model that will predict the result differently, the resulting graph(Figure 5)indicates a totally opposite finding. After training the data with different input sets, all of the results yielded a score of approximately 66%. The proximity of these scores is intriguing and offers some interesting area for analysis

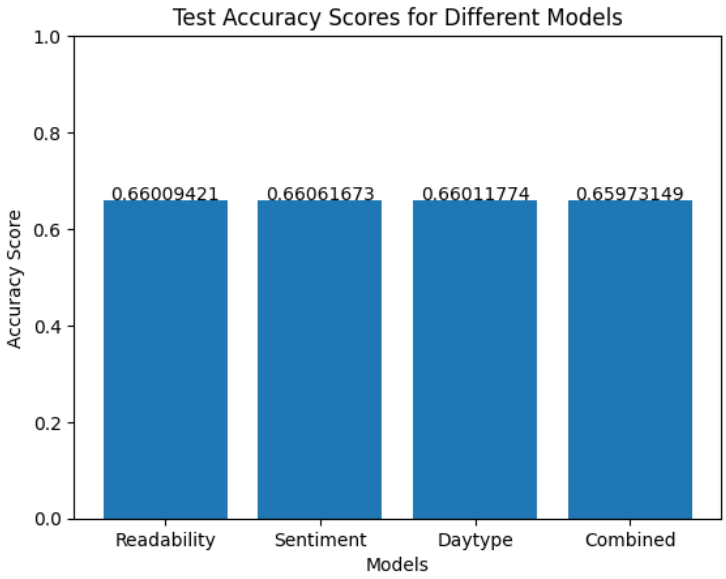

Figure 5: Test Accuracy Scores for Different Model

A predictive accuracy of 66% is notably above random chance (which would be 33% for a ternary classification task), indicating a moderate level of correlation between the selected features and comment quality. While 66% isn't exceptionally high in predictive terms, it's still meaningful, especially given the unpredictable nature of user-generated content on platforms like Reddit. The score confirms that readability, sentiment, and day type do have a tangible relationship with comment quality, though they don't capture the entire essence of what constitutes quality.

The scores for readability, sentiment, and day type are remarkably close to one another. This suggests that, individually, each of these features has a nearly equivalent predictive capability for determining the quality of Reddit comments. Each feature, on its own, contributes roughly the same level of insight into the perceived quality of a comment.

Interestingly, when combining all the features, the predictive score experiences a slight drop to 0.6573149. This could indicate the presence of overlaps between the features when considered together. For instance, highly readable comments might also tend to be positive, leading to a slight redundancy when both features are used together.

5 Limitations and Future Work

5.1 Data Anomalies: Readability Score Outliers and Their Potential Influence

An interesting anomaly that arose during our analysis was the presence of outliers in the 'readability_score' variable. Specifically, we noted a few extreme values that deviated significantly from the majority of the data. While these outliers did not violate the assumptions of the ordinal logistic regression model used (as it does not strictly require a normal distribution of predictors), they could still have affected the estimated coefficients and, subsequently, the overall interpretability of our model.

This can occur because the coefficients in a regression model are influenced by each data point, and extreme values can pull the estimated coefficients towards themselves, making them less reliable or

more difficult to interpret. In our case, these outliers might have affected our ability to accurately understand the impact of readability score on comment quality.

One limitation of our current project is that we did not undertake any specific procedure to detect and handle these outliers, potentially leading to a biased interpretation of the 'readability_score'. To mitigate this issue, future work should consider robust methods to handle outliers, such as trimming or winsorizing the data, or using models resistant to outliers like robust regression models.

This incident underlines the importance of comprehensive data exploration and cleaning processes, before diving into modeling, to ensure that our model results are as accurate and reliable as possible.

5.2 Limitations in Judging Post Quality and Directions for Future Improvements

In our Reddit post quality analysis project, we used the score of a post as the sole variable to determine its quality, tagging it as 'good', 'bad', or 'normal'. While this approach provided us a convenient way to quantify and compare the quality of the posts, it is not to say that it is comprehensive or entirely accurate.

Firstly, Reddit's scoring system is based on upvotes and downvotes from users. While this can reflect the consensus of most users, it may also be influenced by factors such as the timing of the post, the mood of the audience, and other context that might affect voting behavior, none of which were considered in our current analysis. Furthermore, a high score does not necessarily denote high-quality content and vice versa. For instance, a post may gain a high score because it is controversial or on a hot topic, but the content quality may not be high.

Secondly, given the limited data and parameters available to us when conducting the analysis, our model might not accurately capture the complex dynamics of post quality. For instance, we were not able to consider factors like the reputation of the user, the number and quality of comments on the post, which might significantly influence the quality of a post.

Given these limitations, future work can consider incorporating more parameters to improve model performance. For example, introducing more user and post-related metadata such as the activity of the author, the number and quality of comments on the post. One could even try utilizing Natural Language Processing techniques to analyze the content of posts and comments to glean deeper semantic and sentiment information. In addition, exploring more sophisticated machine learning or deep learning models to better understand and predict post quality could be beneficial. Lastly, for a deeper understanding of the intrinsic mechanisms of post quality, future research could also try in-depth qualitative research from sociological and psychological perspectives.

5.3 Model Constraints: Neural Network Depth and the Effect on Quality Measurement

One potential limitation influencing the consistency in quality across different features could be the depth and range of the Neural Network used in the analysis. The neural network's capacity to capture relationships in data depends on its architecture, especially the number of neurons in each layer and the depth of the layers. A shallow MLP might not have the capacity to fully capture the underlying structure among features like readability, sentiment, and day type. This limitation might lead it to produce similar quality measurements across features, as it potentially ignores deeper, more important patterns.

If we can have more time, exploring a larger and deeper neural network could yield different results. A bigger network, with more layers and neurons, inherently possesses a greater capacity to model complex relationships in the data. This increased capacity could allow the model to distinguish features and understanding of how they relate to comment quality. By capturing interactions and non-linearities that a smaller network might miss, a larger MLP could provide more diverse and potentially more accurate quality measurements. However, it's worth noting that a larger network can be more computationally intensive, leading to longer training times.

6 Project Experience Summary

6.1 Chenzheng Li

- Utilized Python and libraries such as Pandas, Numpy, and Scikit-learn to design and implement a neural network model that evaluates the quality of Reddit comments, which validated the effectiveness of our model and the advantages of neural networks in handling complex data structures.
- Enhanced the performance of the model by optimizing the neural network structure, adjusting the number of network layers and neurons, and exploring different training strategies, showing the potential of larger networks in handling data complexity, despite a longer training time.
- Executed data cleaning on the Reddit dataset and enhanced the readability analysis of comments using Python data cleaning tools and NLP techniques, ensuring the accuracy of model training and helping to predict the quality of comments more accurately.
- Designed and implemented a new method to evaluate the quality of Reddit comments by using machine learning technologies to build a model that evaluates comment quality based on features like sentiment analysis, readability score, and day type, providing an efficient, automated way to evaluate the quality of a large number of Reddit comments and significantly improving processing efficiency and accuracy.

6.2 Eric Chan

- Trained a machine learning model to predict quality using features like sentiment score, readability, and day type using the MLP classifier
- Visualize and analyze the accuracy score of different machine learning models, that use different combinations of features to train.
- Discussed and gave ideas on which topic the team should conduct data analysis during the brainstorming phase.
- Organized and facilitated regular team meetings, also ensuring clear communication among members thus improving the team's efficiency.

6.3 Ziying Peng

- Implemented Sentiment Score Calculation: Utilized the VADER model from the NLTK library to compute sentiment scores for Reddit comments, providing a quantifiable measure of sentiment for each comment.
- Implemented Readability Score Calculation: Employed the textstat library and Flesch-Kincaid Grade Level metric to compute readability scores for each Reddit comment, offering vital insights into the complexity of the text for data analysis.
- Conducted Ordinal Logistic Regression Analysis: Applied StatsModels and patsy libraries to perform ordinal logistic regression analysis, investigating the influence of comment sentiment, readability, and day of posting on the perceived comment quality within the Reddit community.
- Self-learned New Techniques: Successfully mastered and applied NLTK, Flesch-Kincaid Grade Level, and Ordinal Logistic Regression to the project despite no prior experience, demonstrating quick learning and adaptability to new technologies.

7 References

[1] Chenzheng L, Eric C, Ziying P. (2023). CMPT353-TeamGG Project Repository. GitHub. https://github.sfu.ca/cla429/CMPT353-TeamGG

[2]Hutto, C.J. & Gilbert, E.E. (2014). VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. Eighth International Conference on Weblogs and Social Media (ICWSM-14). Ann Arbor, MI, June 2014. https://comp.social.gatech.edu/papers/icwsm14.vader.hutto.pdf

[3] Wikipedia Contributors. (2023). Flesch–Kincaid readability tests — Wikipedia, The Free Encyclopedia. https://en.wikipedia.org/wiki/Flesch