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."

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

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(Table 1)

Attribute	Attributes	Description
1	archived	Indicates whether the post is archived or not
2	author	The username of the person who posted the comment
3	body	The text content of the comment
4	controversiality	A measure of how controversial the comment is, based on the number of upvotes and downvotes
5	created_utc	The time when the comment was created in Unix time
6	downs	The number of downvotes the comment received
7	edited	Indicates whether the comment has been edited after posting
8	gilded	The number of gold awards the comment received
9	id	A unique identifier for the comment
10	link_id	The unique identifier for the post to which the comment was made
11	name	A full name that combines the type (t1 for comment) and id to uniquely identify the comment
12	parent_id	The ID of the parent comment or post to which this comment was made
13	retrieved_on	The time when this data was collected in Unix time
14	score	The total score of the comment (ups - downs)
15	score_hidden	Indicates whether the score of the comment is hidden or not
16	subreddit	The subreddit where the comment was posted
17	subreddit_id	The unique identifier for the subreddit
18	ups	The number of upvotes the comment received
19	month	The month when the comment was posted

Table 1: Original Data

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.

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

Table 2: List of Attributes in the dataset after clean

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:

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".

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

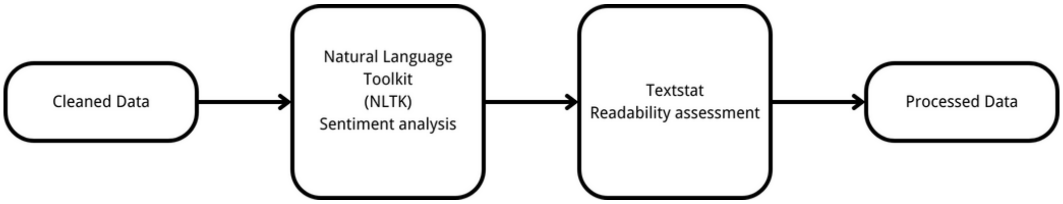
Cleaned Data
Natural Language Toolkit (NLTK) Sentiment analysis
Textstat Readability assessment
Processed Data

Figure 1: Process work flow

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.

Results and Key Findings

3.3 Stat

3.4 Machine Learn

Data Visualization

Discussion, Limitations, and Future Work

References

[1]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

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