# Temporal, Correlation and Sentiment Analysis of ChatGPT Subreddit

# A Comparative Study Using Multiple Sentiment Analysis Tools and Correlation Analysis

**ABSTRACT**

This paper presents an in-depth investigation into sentiment and temporal trends within the ChatGPT subreddit community spanning from January 2023 to January 2024. Leveraging data extracted through the PRAW interface from Reddit, a multi-faceted sentiment analysis was conducted employing various algorithms such as VADER, TextBlob, AFINN, SenticNet, and BingLiu. Additionally, temporal analysis was undertaken to identify any discernible patterns or fluctuations over the specified timeframe. Furthermore, correlation analysis was executed to explore the relationship between sentiment scores derived from these algorithms and the upvote ratio and score of posts within the community. The comprehensive findings offer a nuanced understanding of sentiment dynamics and temporal trends within the ChatGPT subreddit, shedding light on the complex interplay between sentiment and user engagement metrics on the platform. This study contributes significantly to the growing body of knowledge surrounding sentiment analysis in online communities, particularly within the context of Reddit, highlighting its relevance in deciphering user behavior and engagement patterns.

Keywords: AI, ChatGPT, Prompts , Time , Learning , Simple , Easy

**INTRODUCTION**

In the ever-evolving landscape of online communities, Reddit stands out as one of the most vibrant platforms, fostering diverse communities where users engage in discussions, share content, and express their opinions on a wide range of topics. Among these communities, the ChatGPT subreddit serves as a hub for enthusiasts, practitioners, and curious minds interested in the developments and applications of conversational AI, particularly the GPT (Generative Pre-trained Transformer) models.

Understanding the sentiment dynamics and temporal trends within such communities is crucial for gaining insights into user behavior, engagement patterns, and community dynamics. Sentiment analysis, a branch of natural language processing, offers a systematic approach to quantify and analyze sentiments expressed within text data. By leveraging sentiment analysis techniques, researchers can decipher the prevailing sentiments within online communities, discern sentiment trends over time, and explore their impact on user engagement metrics. In this context, our study focuses on analyzing sentiment and temporal trends within the ChatGPT subreddit community over a specific timeframe, from January 2023 to January 2024.

To achieve this, we employ various sentiment analysis algorithms, including VADER, TextBlob, AFINN, SenticNet, and BingLiu, to gauge the sentiment expressed in posts and comments within the ChatGPT subreddit. These algorithms offer complementary perspectives on sentiment analysis, each with its strengths and limitations. Additionally, temporal analysis is conducted to identify any temporal patterns or fluctuations in sentiment over the specified timeframe.

Furthermore, we investigate the correlation between sentiment scores derived from these algorithms and the upvote ratio and score of posts within the ChatGPT subreddit. Understanding how sentiment relates to post engagement metrics can provide valuable insights into the factors influencing user engagement and community dynamics. By integrating sentiment analysis with temporal and correlation analyses, our study aims to provide a comprehensive understanding of sentiment dynamics and their implications within the ChatGPT subreddit community.

In this paper, we present the methodology employed for data collection, sentiment analysis, temporal analysis, and correlation analysis. Subsequently, we discuss the findings of our study, highlighting significant trends, patterns, and correlations observed within the ChatGPT subreddit community. Finally, we conclude with implications, limitations, and avenues for future research, emphasizing the importance of sentiment analysis in understanding online community dynamics and fostering informed decision-making in community management and content moderation.

**METHODOLOGY**

Data Collection

Data was collected from the ChatGPT subreddit using the Python Reddit API Wrapper (PRAW). The data collection process involved extracting posts and associated comments made within the specified timeframe, from January 2023 to January 2024. The dataset includes textual content, post metadata (e.g., upvote ratio, score), and timestamps.

A graph of a number of tokens

Description automatically generated

Data Preprocessing

Before conducting sentiment analysis, textual data underwent preprocessing steps such as tokenization, removal of stop words, and lemmatization to standardize the text and enhance the accuracy of sentiment analysis results. Additionally, any irrelevant or non-textual content was filtered out to ensure the quality of the dataset. The pipeline consists of three functions:

str.lower : Converts text to lowercase.

Tokenize : Tokenizes the text, splitting it into individual words or tokens.

remove\_stop : Removes stop words from the text.

The prepare function takes a text input and a pipeline of functions as arguments. It applies each function in the pipeline sequentially to the text and returns the processed tokens.

Word Cloud

A word cloud is a visual representation of text data where the size of each word indicates its frequency or importance in the dataset. Typically, common words like "the" or "and" are excluded, allowing thematic or significant terms to stand out. Word clouds are often used to quickly convey the most prominent themes or topics within large volumes of text.

A close-up of words

Description automatically generated

Sentiment Analysis

Multiple sentiment analysis algorithms were employed to analyze the sentiment expressed within the textual content of posts and comments. These algorithms include:

VADER is a lexicon and rule-based sentiment analysis tool designed specifically for analyzing text and deriving sentiment scores. It is particularly adept at handling social media text due to its ability to recognize and interpret nuances such as slang, emojis, and emoticons. VADER utilizes a combination of sentiment lexicons (pre-defined dictionaries) and grammatical rules to assess the sentiment of a given piece of text. It assigns sentiment scores to individual words and then combines these scores to generate an overall sentiment score for the entire text, accounting for both the polarity (positive, negative, or neutral) and intensity of sentiment. VADER is known for its speed and effectiveness in analyzing short, informal text snippets often found in social media platforms.

TextBlob

TextBlob is a Python library built on top of NLTK (Natural Language Toolkit) and Pattern libraries, offering a simple API for common natural language processing (NLP) tasks such as part-of-speech tagging, noun phrase extraction, and sentiment analysis. Sentiment analysis in TextBlob is based on a sentiment lexicon approach, where each word in the text is assigned a polarity score (positive, negative, or neutral) based on its appearance in a predefined lexicon. These scores are then aggregated to compute the overall sentiment polarity of the text. TextBlob also provides the flexibility to train custom sentiment classifiers using user-provided datasets.

AFINN

AFINN is a lexicon-based sentiment analysis tool that assigns a pre-defined polarity score to words based on their appearance in a manually curated sentiment lexicon. Developed by Finn Arup Nielsen, AFINN contains over 2,477 English words and phrases, each associated with a sentiment score ranging from -5 (most negative) to +5 (most positive). Sentiment analysis with AFINN involves scanning a piece of text for words present in the lexicon and summing up their associated scores to compute the overall sentiment polarity. AFINN is lightweight and easy to use, making it suitable for applications requiring quick sentiment analysis of textual data.

SenticNet

SenticNet is a semantic resource for sentiment analysis and opinion mining that incorporates both conceptual and affective information. Unlike lexicon-based approaches, SenticNet represents concepts in a multi-dimensional semantic space, where each concept is associated with a vector of affective scores representing its valence, arousal, and dominance dimensions. These scores capture not only the polarity but also the intensity and other affective dimensions of sentiment associated with concepts. SenticNet is often used in applications requiring deeper semantic understanding and nuanced sentiment analysis, such as opinion mining in reviews or social media.

BingLiu

BingLiu Opinion Lexicon is a sentiment lexicon created by Bing Liu, consisting of positive and negative words commonly found in product reviews. The lexicon contains approximately 6,800 words and phrases, each manually annotated with its sentiment polarity (positive or negative). Sentiment analysis using BingLiu involves matching words from the input text against the lexicon and aggregating their associated sentiment polarities to compute the overall sentiment score. BingLiu is particularly useful for sentiment analysis tasks focusing on product reviews or domains where opinions are expressed through specific vocabulary.

Each algorithm provides sentiment scores or polarity values indicating the positivity or negativity of the text. These scores were computed for both posts and comments, capturing the overall sentiment of the content.

Temporal Analysis

Temporal analysis was conducted to examine how sentiment fluctuates over time within the ChatGPT subreddit community. The dataset was segmented into time intervals (e.g., daily, weekly), and sentiment scores derived from the algorithms were aggregated and analyzed for each interval. This analysis aimed to identify any temporal patterns or trends in sentiment expression.

Correlation Analysis

Correlation analysis was performed to explore the relationship between sentiment scores and post engagement metrics, namely upvote ratio and score. Pearson correlation coefficients were calculated to measure the strength and direction of the linear relationship between sentiment scores and engagement metrics. This analysis provided insights into how sentiment influences post engagement within the community.

We calculate the mean correlations of certain columns in one**\_**df. The columns listed are 'vader\_Score', 'textblob\_Score', 'afinn\_score', 'senticnet\_score', 'Upvote\_Ratio', and 'Score'. These columns likely represent different sentiment analysis metrics (VADER, TextBlob, AFINN, SenticNet) and engagement metrics (Upvote Ratio, Score).

Correlational analysis is a statistical method used to investigate the relationship between two or more variables within a dataset. It aims to quantify the extent to which changes in one variable are associated with changes in another. This analysis typically yields a correlation coefficient, such as Pearson's r, which ranges from -1 to 1. A positive correlation coefficient indicates a direct relationship, meaning that as one variable increases, the other tends to increase as well.

Correlational analysis plays a vital role in various fields, including psychology, economics, and epidemiology, where researchers seek to understand patterns and make predictions based on observed data. However, it's essential to interpret correlational findings cautiously, considering potential confounding variables and alternative explanations for the observed associations. Visual representations such as scatter plots are often utilized to aid in the interpretation of correlational relationships, allowing researchers to identify trends and outliers within the data. Overall, correlational analysis provides valuable insights into the interplay between variables but requires careful consideration of its limitations and the broader context of the research question at hand.

Statistical Analysis

Statistical techniques were employed to analyze the results of sentiment and correlation analyses. Descriptive statistics were used to summarize sentiment scores and engagement metrics, while inferential statistics were employed to assess the significance of observed correlations. By employing this comprehensive methodology, we aimed to provide a rigorous and systematic analysis of sentiment and temporal trends within the ChatGPT subreddit community, yielding valuable insights into the dynamics of sentiment expression and its implications for community engagement and interaction.

**RESULTS**

Sentiment Analysis

The sentiment analysis encompassed a comprehensive examination of posts and comments within the ChatGPT subreddit community, utilizing a range of sentiment analysis algorithms, including VADER, TextBlob, AFINN, SenticNet, and BingLiu. Each algorithm provided unique insights into the prevailing sentiments expressed within the community, offering a multifaceted perspective on sentiment polarity. Additionally, subjectivity scores were computed to gauge the degree of subjective opinions present in the text, contributing to a deeper understanding of sentiment nuances.

**Vader Algorithm**

A graph of a distribution of a positive value

Description automatically generated with medium confidence

The histogram you've provided represents the distribution of VADER Sentiment Compound Scores. VADER (Valence Aware Dictionary and sentiment Reasoner) is a tool used for text sentiment analysis that is sensitive to both polarity (positive/negative) and intensity (strength) of emotion. The Compound Score is a metric that calculates the sum of all the lexicon ratings which have been normalized between -1 (most extreme negative) and +1 (most extreme positive).

In this graph, the x-axis shows the range of possible compound scores, while the y-axis shows the count of occurrences for each score range.

**TextBlob and Afinn Algorithm**

A graph of a line graph

Description automatically generated

A graph of a bar graph

Description automatically generated with medium confidence

The uploaded image is a histogram depicting the distribution of AFINN Sentiment Scores. The AFINN sentiment analysis model assigns scores to words, with scores ranging from -5 (very negative) to +5 (very positive), and these scores are then aggregated for a piece of text to give an overall sentiment score.

The x-axis represents the AFINN sentiment scores, and the y-axis represents the count of occurrences of those scores. In this graph, there is a prominent peak at zero, indicating that many texts have a neutral sentiment. The distribution also shows some occurrences of texts with negative and positive sentiments, but these are significantly fewer compared to the neutral texts. The data suggests that the set of texts analyzed contains mostly neutral sentiments, with a smaller presence of both negative and positive sentiments.

**SenticNet Algorithm**

A graph of a box plot

Description automatically generated

In a box plot like this, the central rectangle spans the first quartile to the third quartile (the interquartile range or IQR), with a line at the median. The "whiskers" extend from either end of the box to show the range of the data, typically to 1.5 times the IQR above the third quartile and below the first quartile. Points outside of this range are often considered outliers and are plotted as individual points.

Here, the median is at 0, suggesting that the overall sentiment is neutral, with the box representing the middle 50% of scores very close to the median, indicating low variability in sentiment scores within the IQR. There are several outliers on both the negative and positive sides, suggesting there are some texts with strong sentiments.

**Temporal Analysis**

The temporal analysis unveiled intriguing patterns and fluctuations in sentiment and subjectivity levels over the specified timeframe, spanning from January 2023 to January 2024. By segmenting the data into temporal intervals, such as daily or weekly periods, trends in sentiment and subjectivity became apparent, reflecting the dynamic nature of community interactions. Certain time periods exhibited pronounced shifts in sentiment and subjectivity, potentially influenced by significant events, discussions, or community dynamics. Understanding these temporal trends provides valuable insights into the evolving sentiment landscape within the ChatGPT subreddit community.

Temporal analysis involves the examination of data over time to identify trends, patterns, and relationships. It is a crucial method used in various fields, including economics, finance, climatology, and epidemiology. This analysis allows researchers to understand how variables change over different time periods and to make predictions about future behavior.

Time series data, which records observations at regular intervals, is often analyzed using techniques such as trend analysis, seasonality decomposition, and forecasting models like ARIMA (Autoregressive Integrated Moving Average). Temporal analysis can reveal long-term trends, cyclic patterns, and seasonal fluctuations within the data. It also helps researchers understand the impact of temporal dependencies and autocorrelation on the variables under study. Moreover, temporal analysis enables the identification of critical events or anomalies that may influence the overall trajectory of the data. Interpretation of temporal trends requires consideration of historical context, external factors, and potential causative relationships. Overall, temporal analysis provides valuable insights into the dynamics of time-varying phenomena, aiding decision-making and policy formulation across various domains.

**Correlation Analysis**

The correlation analysis delved into the intricate relationship between sentiment/subjectivity scores and post engagement metrics, such as upvote ratio and score. Through rigorous statistical analysis, correlations between sentiment dynamics and post engagement were identified, shedding light on the factors influencing community interaction. Positive correlations between higher sentiment scores/subjectivity levels and increased post engagement metrics were observed in some instances, underscoring the role of sentiment expression in shaping community engagement. However, the strength and significance of these correlations varied across different sentiment analysis algorithms, highlighting the nuanced nature of sentiment dynamics within the community.

In summary, the results of the sentiment analysis, temporal analysis, correlation analysis, and word cloud visualizations offer a multifaceted understanding of sentiment dynamics, subjectivity levels, and their implications for community engagement within the ChatGPT subreddit. These findings provide valuable contributions to the field of online community research, informing future investigations and guiding community management strategies to foster a thriving and engaged community environment.

**CONCLUSION**

In conclusion, this study has provided valuable insights into the sentiment dynamics, subjectivity levels, and temporal trends within the ChatGPT subreddit community. Through a comprehensive analysis encompassing sentiment analysis, temporal analysis, correlation analysis, and word cloud visualizations, we have gained a nuanced understanding of the factors shaping community interactions and engagement patterns.

**Sentiment Analysis**

The sentiment analysis revealed a diverse range of sentiments expressed within the community, with varying degrees of positivity, negativity, and neutrality across different posts and comments. Subjectivity scores further highlighted the prevalence of subjective opinions within the textual content, reflecting the multifaceted nature of community discourse.

A collage of graphs and diagrams

Description automatically generated

VADER Sentiment Scores: The top-left graph is a histogram that represents the frequency of VADER sentiment scores. The scores are concentrated around zero, which may indicate a lot of neutral sentiment in the data.

TextBlob Sentiment Polarity Scores: The top-right graph is a density plot for TextBlob sentiment polarity scores, which shows the distribution of scores across a continuous range. The plot is bell-shaped and centered around zero, which suggests a normal distribution of sentiment scores with most of them being neutral.

AFINN Sentiment Scores: The bottom-left graph is another histogram showing the distribution of AFINN sentiment scores. Unlike the VADER histogram, the AFINN scores have a wider distribution and multiple peaks, suggesting a variation in sentiment with a significant number of positive, neutral, and negative scores.

SenticNet Sentiment Scores: The bottom-right graph appears to be a box plot combined with a strip plot, displaying the distribution of SenticNet sentiment scores. The box plot provides a summary of the distribution, indicating the median, interquartile range, and potential outliers. The strip plot overlays individual data points, giving a sense of the data's density at different score levels.

**Temporal Analysis**

Temporal analysis uncovered temporal patterns and fluctuations in sentiment and subjectivity over time, underscoring the dynamic nature of sentiment expression within the community. It involves studying how phenomena change over time and understanding the patterns or trends that emerge from these changes. It's used across various fields, such as finance, meteorology, and social sciences, to predict future events, identify cycles, or evaluate the impact of past events. These temporal trends may be influenced by external events, discussions, or community dynamics, emphasizing the need for continuous monitoring and analysis of sentiment dynamics.

A graph showing a number of blue lines

Description automatically generated

Volatility: The plot shows considerable volatility in sentiment over time. The length of the error bars suggests that there's a significant variation in sentiment on a day-to-day basis.

Trend: There doesn't seem to be a clear upward or downward trend over the year, indicating that there is no long-term increase or decrease in sentiment.

Mean Sentiment: The average sentiment score hovers around zero, which may suggest a balance between positive and negative sentiments in the dataset analyzed.

Error Bars: The error bars represent the range of sentiments for each day. Days with longer bars indicate more variability in sentiment on that day.

No Seasonality: From this graph, there doesn't appear to be any seasonal pattern, which would be seen as consistent fluctuations at regular intervals (e.g., more positive sentiment during certain months).

**Correlation Analysis**

Correlation analysis elucidated the relationship between sentiment/subjectivity scores and post engagement metrics, revealing positive correlations between higher sentiment scores/subjectivity levels and increased post engagement in certain instances. These findings underscore the importance of sentiment expression in driving community engagement and interaction.

A group of graphs showing different colored dots

Description automatically generated with medium confidence

Additionally, word cloud visualizations provided qualitative insights into the prevalent topics, themes, and sentiments expressed within the community, complementing the quantitative sentiment analysis results with visual representations of community discourse.

Overall, the findings of this study contribute to the growing body of knowledge surrounding online community dynamics and sentiment analysis. By understanding the sentiment dynamics, subjectivity levels, and temporal trends within the ChatGPT subreddit community, community managers and moderators can make informed decisions to foster a positive and engaging community environment.

Moving forward, future research could explore the impact of sentiment dynamics on community cohesion, member satisfaction, and long-term community sustainability. Additionally, investigating the effectiveness of sentiment-based interventions and community engagement strategies could further enhance our understanding of how sentiment influences online community dynamics. In conclusion, this study highlights the importance of sentiment analysis in understanding online community dynamics and underscores its relevance in guiding community management strategies and fostering thriving online communities.