Sentiment Analysis through My WhatsApp Messages

Motivation

This project uses sentiment analysis, an effective instrument to examine the complex dynamics of my messaging habits and emotional patterns. The ultimate goal is to improve my communication methods by exploring my emotions throughout the day. This will help me become more self-aware, which may lead to fresh insights into the things that are impacting my behaviour and mood.

I aim to find the little nuggets of information that could point to trends, triggers, and connections between particular communication styles and my emotional condition by carefully examining all of my texts. Not only do I want to maximize my relationships with other people, but I also want to design a customized road map for navigating everyday interactions in a way that supports my emotional health.

The project serves as an exploration of self-improvement and self-discovery, using data-driven insights to improve communication skills and get a deeper understanding of the complex dance between words and emotions. As I work through the threads of my messaging tapestry, I desire to equip myself with the knowledge necessary to create deep connections, successfully handle stress, and create an atmosphere of care for pleasant emotional experiences. My goal with this introspective inquiry is to become more resilient and emotionally intelligent enough to handle the challenges of everyday life, while also improving my communication abilities.

Data Source

The primary data source for this project is my personal WhatsApp messages. The dataset includes conversational data with timestamps, allowing for a detailed analysis of messaging intensity and sentiment over different periods. The data is collected and processed using Python scripts in a Jupyter Notebook environment.

An illustration of the data format - extracted as a text file (.txt)

```
01/04/18, 4:17 pm - Mike: Or kya
01/04/18, 4:17 pm - Mohit: Else kl
01/04/18, 4:17 pm - Mohit: Ek min
01/04/18, 4:18 pm - Mike: Ok
01/04/18, 7:04 pm - Mike: 1 kg papeeta poora Lana katwana mat
01/04/18, 7:04 pm - Mike: Half kg apples-4se 5 Chad jae
01/04/18, 7:05 pm - Mike: Or 250gms oranges
01/04/18, 7:05 pm - Mike: 
02/04/18, 1:21 am - Mohit: Mika
02/04/18, 1:21 am - Mohit: Photo bhejio
02/04/18, 1:21 am - Mohit: Bhai
02/04/18, 9:16 am - Mohit: You deleted this message
02/04/18, 9:16 am - Mohit: Mika
03/04/18, 1:09 am - Mike: <Media omitted>
```

Data Analysis

1. Data Collection

Conversational data is extracted from WhatsApp to provide a complete picture of conversations. For every interaction, time stamps are added, allowing for a more in-depth analysis of temporal trends.

2. Data Cleaning

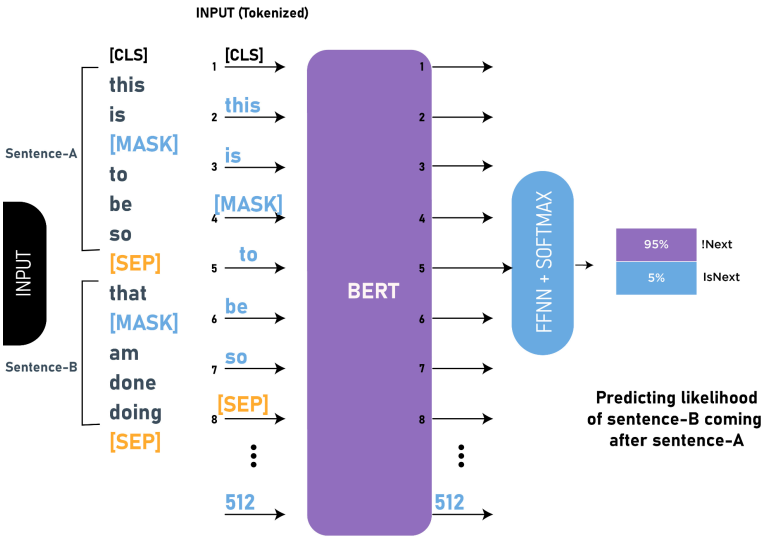
- Missing values are handled.
- Text data is converted to lowercase
- Special characters and links are removed.
- Tokenization is performed
- Stopwords are removed.
- Spell checking is implemented using the ZEMBEREK Turkish NLP library.

3. Message Intensity Analysis

I examine the rise and fall of messages over various time and date intervals, providing an analysis of message intensity. By exposing temporal nuances and possible relationships, an in-depth analysis of message counts enables a deep comprehension of messaging behaviours.

4. Sentiment Analysis

By utilizing an advanced Turkish model which is based on BERT, sentiment analysis goes beyond simple polarity recognition. Every message follows a comprehensive examination, resulting in sentiment scores that provide emotional subtleties. This method provides a better knowledge of the emotional spectrum hidden in every communication, improving sentiment analysis over a binary classification.



5. Hypothesis Testing

The Pearson correlation test is used to identify trends and connections between sentiment scores over various time intervals, and statistical precision is put to use. Detailed definitions and thorough testing are used to clarify the complex interactions between emotional states and temporal dynamics. This statistical investigation acts as a compass, directing the identification of trends and patterns that affect mental health and methods of communication.

Null Hypothesis(H0)

There is no significant correlation between the selected periods of the day.

Alternative Hypothesis(H1)

There is a significant correlation between the selected periods of the day.

Hypothesis Testing Technique:

In this section, the Pearson correlation test to assess the correlation between different variables is employed. The Pearson function is utilized to calculate both correlation coefficients and associated p-values for hypothesis testing.

Purpose:

The Pearson correlation test is employed to understand the strength and direction of the linear relationship between two variables. This analysis provides insight into whether changes in one variable are associated with systematic changes in another.

Method:

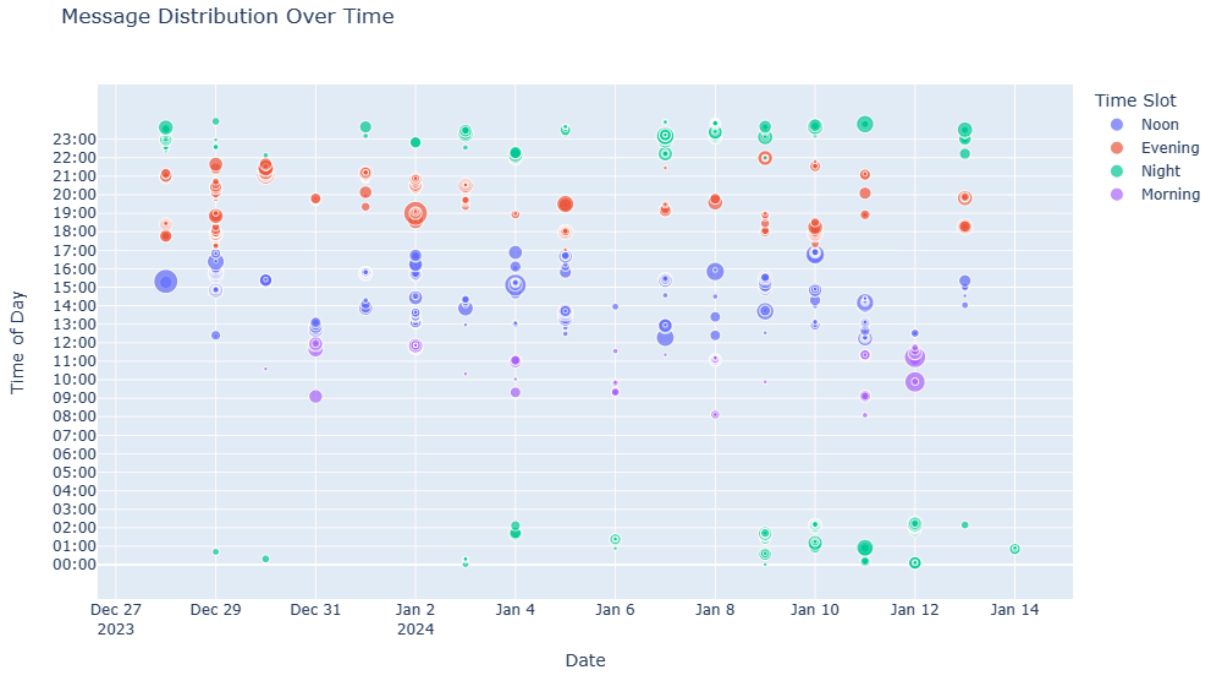
The code snippet utilizes the Pearson function from the **scipy.stats** library to conduct hypothesis testing based on the Pearson correlation coefficient. This coefficient is employed to measure the linear relationship between two variables, providing insights into the strength and direction of their association. The accompanying p-value assists in evaluating the statistical significance of the observed correlation.

In this context, the null hypothesis posits that there is no correlation between the specified pairs of variables. The code then calculates the p-value associated with the correlation coefficients for morning vs. night average sentiment and evening vs. noon average sentiment. The significance level (alpha) is set to 0.05, a commonly used threshold in hypothesis testing.

The subsequent evaluation of results involves comparing the computed p-values with the chosen significance level. If a p-value is less than the significance level, it suggests that there is a statistically significant correlation between the respective pairs of variables. Conversely, if the p-value exceeds the significance level, the conclusion is that there is no significant correlation.

Findings

Message Distribution Over Time

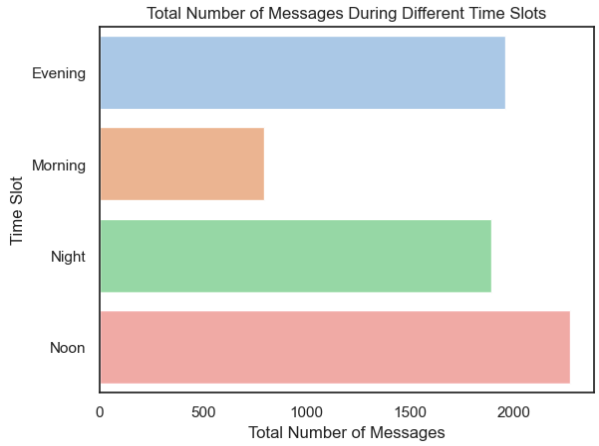


The total number of messages during different time slots follows a distinct ranking, revealing patterns in messaging habits.
Please be aware that the plotted graph is not visible when viewing the .ipynb file on GitHub.

- The x-axis represents dates from Dec 27, 2023, to Jan 14, 2024.
- The y-axis represents the times of day in hours from 00:00 to 23:00.
- Messages are represented as coloured dots scattered across the graph, colour-coded based on the time of day they were sent:
    - Noon: Green dots
    - Evening: Red dots
    - Night: Blue dots
    - Morning: Purple dots

The scatter plot shows the distribution of messages sent over different times of the day during the specified date range. Messages appear to be evenly distributed throughout different times of the day but are more concentrated during evening hours. This graph provides an overview of the messaging activity over time.

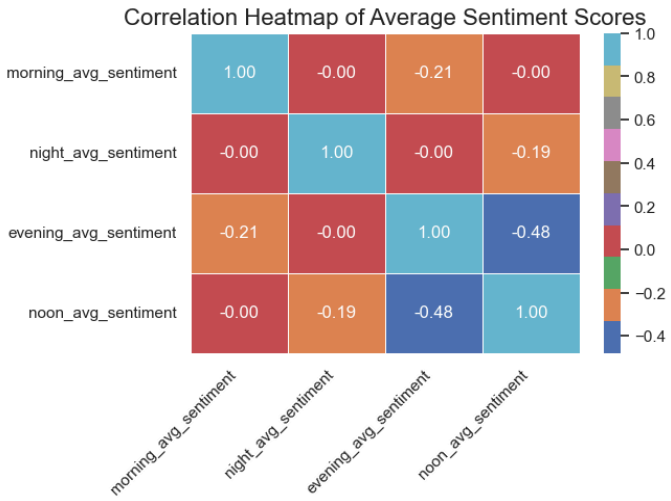
Total Number of Messages During Different Time Slots



- Evening: There are slightly over 1000 messages sent during this time, represented by a blue bar.
- Morning: The orange bar indicates around 750 messages sent during this time.
- Night: The green bar shows that around 1250 messages are sent during this time.
- Noon: This time slot has the highest number of messages, nearly 2000, represented by a red bar.

This graph provides an overview of the total number of messages sent during different times of the day. It appears that the most active time for sending messages is at Noon, followed by Night, Evening, and Morning.

Sentiment Correlations



- Most time slots show no significant correlation in sentiment scores, except for a significant correlation between evening and noon sentiment. This implies a potential pattern or similarity in sentiment during these specific time slots
    - A moderate negative correlation (-0.48) between evening_avg_sentiment and noon_avg_sentiment, indicated by green colour.
    - A weak negative correlation (-0.21) between morning_avg_sentiment and evening_avg_sentiment, indicated by light green colour.

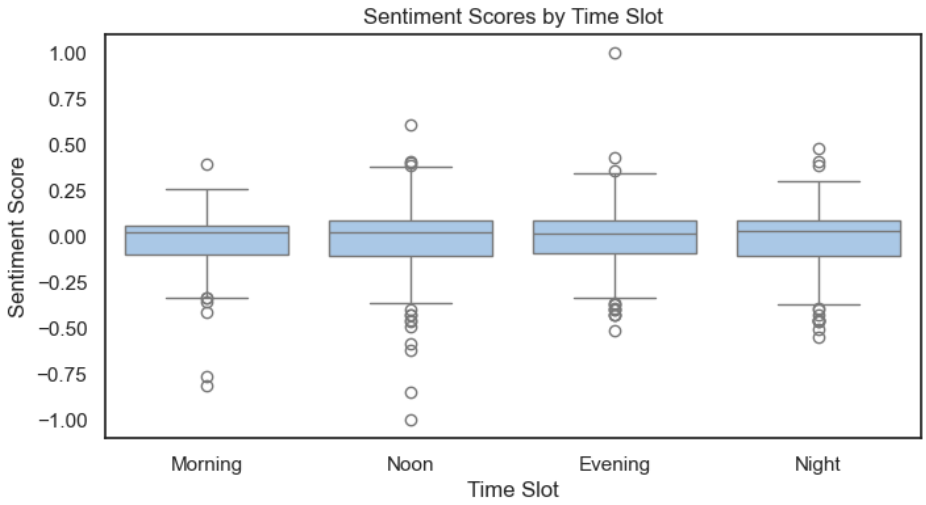
This heatmap provides an overview of how sentiment scores during different times of the day correlate with each other. It appears that there's a negative correlation between the sentiments in the evening and at noon, as well as in the morning and the evening.

Sentiment Scores Table

In this analysis, a table containing sentiment scores for every message is presented. The sentiment scores were calculated using the BERT-Turkish model. Tables serve as the foundation for hypothesis testing to explore patterns and trends in sentiment across different time slots.

- circles (outliers) plotted outside of the boxes to represent individual scores that fall outside of the interquartile range

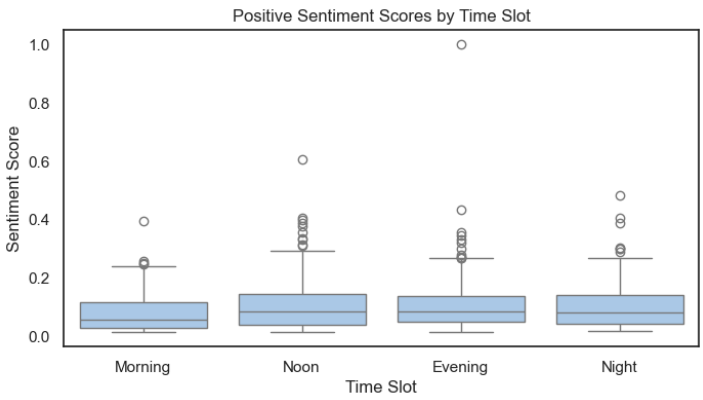
Sentiment Scores Overview



- Sentiment scores for each time slot are visually represented with blue boxes showing the interquartile range (IQR) – the statistical dispersion between the upper and lower quartiles.
- Black whiskers extend from each box, indicating the range where most values lie.
- All four boxes have positive median values (middle lines), representing the center point between the higher and lower halves of the sentiment data.

This box plot graph provides an overview of the distribution of sentiment scores during different times of the day. It appears that the median sentiment score is positive throughout the day but varies slightly. The spread of sentiments shows that there are both positive and negative sentiments at all times of the day, with outliers present in each time slot. This could indicate that while the overall sentiment is positive, there are instances of extreme sentiments (both positive and negative) at all times of the day. This could be useful for understanding the variability and distribution of sentiments during different times of the day and could be used for further analysis or decision-making. For example, this could be useful for understanding communication patterns, and emotional well-being, and even for scheduling important conversations when sentiments tend to be more positive.

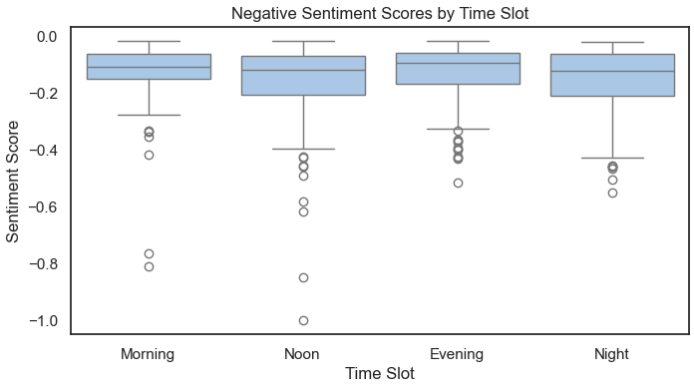
Positive Sentiment Scores



- Sentiment scores appear to be higher during Evening and Night compared to Morning and Noon.

This box plot graph provides an overview of the distribution of positive sentiment scores during different times of the day. It appears that there are instances of very positive sentiments during the evening and nighttime slots.

Negative Sentiment Scores



- In all time slots except for “Night,” there are more outliers below the boxes indicating instances of very negative sentiments.

This box plot graph provides an overview of the distribution of negative sentiment scores during different times of the day. It appears that there are instances of very negative sentiments in all time slots except for “Night.”

Limitations and Future Work

Limitations

Channel-Specific Focus

While WhatsApp messages offer a rich source of data, it's important to acknowledge that this analysis might not encapsulate sentiments expressed through other communication channels. Expanding the scope to include diverse platforms could provide a better understanding of emotional expression.

Sentiment Model Variations

Despite its strength, the sentiment analysis model might not be able to properly capture the complex range of feelings. The analysis would be more accurate and comprehensive if the model were to be continuously improved and adjusted to account for variations in emotional expression.

Future Work

Data Source Diversification

To achieve a more comprehensive sentiment analysis, future work could involve integrating data from various sources beyond WhatsApp. This might include emails, social media interactions, or other messaging platforms, offering a broader view of emotional states across different contexts.

Model Refinement

The design is essential to continue improving sentiment analysis models. This involves not just addressing the subtleties of emotional expression but also investigating sophisticated methods of adding modifications unique to each user to customize the model to suit the style of communication.

Long-Term Trends Analysis

Analyzing the data over a longer time frame makes it possible to spot trends and changes in messaging practices. This long-term method may reveal trends that develop over time, offering a more profound understanding of how communication techniques and emotional health have changed.

Conclusion

Gaining insight into individual messaging behaviours and sentiment patterns is essential for self-discovery and better communication. Although the basis for such insights is established by this research, the necessity for ongoing progress is driven by the realization of its limitations. The complex relationship between words and emotions can be better understood by combining various data sources, improving sentiment algorithms, and conducting long-term studies. The exploration continues beyond this point, diving further into an area of never-ending research, establishing the framework for more studies into emotional health and behaviour that are dedicated to constant improvement and growth.