ANALYSIS OF USAGE OF EMOTICONS ON TWITTER

Devangi Joshi   
3rd *Year,CSE  
2110110183*

Khushi Goyal  
3rd Year,CSE *2110110286*

*Abstract*— This research project aims to analyze the usage of emoticons (emojis) in tweets on Twitter to gain a deeper understanding of emotional expressions and trends in social media conversations. Emojis play a crucial role in conveying sentiments, and this analysis can provide valuable insights into the emotional tone of Twitter conversations. The study can have various applications, including marketing, sentiment analysis, and understanding social trends. Emojis are not just simple pictograms; their meanings can evolve within the context of online discussions, and this project seeks to decipher these nuances.

# INTRODUCTION

*A. Background*

The rise of social media has fundamentally transformed the way people communicate and express themselves, and Twitter stands out as a prominent platform in this digital landscape. With its character-limited format, Twitter has become a hub for succinct and rapid communication. A distinctive feature of Twitter conversations is the widespread utilization of emojis—small digital icons that encompass a rich spectrum of emotions, ideas, and sentiments. Emojis, while visually simple, hold immense power in shaping the emotional tone and context of tweets. In the context of Twitter, where brevity is valued, these compact symbols serve as essential tools for conveying complex emotions and adding depth to textual exchanges

*B. Emojis on Twitter*

The versatile nature of emojis on Twitter is such that they serve not only to enhance the emotional quality of tweets but also to simplify communication. However, the true meaning and interpretation of an emoji are far from straightforward. Their significance depends heavily on the specific tweet in which they are embedded, as well as the ever-evolving trends within the realm of online discourse. It is this nuanced relationship between emojis and the emotional landscape of Twitter that this research seeks to explore.

*C. Research Objectives*

This research endeavour is motivated by the need to unravel the intricate relationship between emojis and the emotional expressions within Twitter. While emojis are often used to infuse a light-hearted tone into textual content, their meanings are highly context-dependent.

We aim to understand our twitter dataset using various analysis techniques and algorithms available, creating visualizations to understand the meaning and usage of emoticons.

*D. Significance of the Study*

Emojis are not static symbols; their meanings evolve over time, are shaped by cultural influences, and adapt to the dynamic nature of online communication. By delving into this dynamic aspect of emojis, this research seeks to provide valuable insights into the emotional fabric of Twitter conversations. This understanding has broad-reaching implications, ranging from informing marketing strategies and sentiment analysis to identifying emerging trends within the realm of online discourse..

# DATA EXTRACTION

*A. Data Collection*

The primary source of data for this research is Twitter, a microblogging platform renowned for its dynamic and concise communication. This dataset is inclusive, encompassing a diverse range of tweets to provide a representative sample of Twitter conversations. For the collection of Twitter data, we have utilized Kaggle, a well- established platform for accessing and sharing datasets. The dataset is available in the form of a spreadsheet in .csv format.

## Data Import

## We have imported the dataset (the .csv file) into the root folder of our project using a piece of python code utilising python libraries numpy and pandas.

## The file contains about 20000-28000 tweets. The document has only one column titled ‘text’ which contains the raw tweet.

## We will be employing various preprocessing and cleaning techniques on this data first.

# DATA PRE-PROCESSING

Cleaning and Preparation: Prepare the data for analysis by cleaning and pre-processing it. This includes steps such as:

Text Extraction: refers to the process of identifying and selecting tweets that contain emoticons. Emoticons are those small pictorial representations used to express emotions or sentiments in a tweet. This step involves scanning through the dataset and identifying tweets that include emoticons.

Case Standardization: is the practice of converting all text in the dataset to a common case, typically lowercase. This step ensures consistency in text analysis. For example, if we have tweets with a mix of uppercase and lowercase letters, converting everything to lowercase standardizes the text for further processing, making it easier to compare and analyse.

Removal of Non-English Text: involves filtering out any text that is not written in the English language. This is particularly important if we want to focus on analysing English-language tweets. Tweets in various languages may be present in our dataset, and this step helps to isolate and work with English text exclusively.

Removal of Irrelevant Characters: In this step, "Removal of Irrelevant Characters" entails

eliminating any characters, spaces, tabs, or symbols that are not relevant to the analysis of emotions

conveyed by emoticons. This ensures that the text data is clean and only includes essential components for emoticon-based analysis.

Data Labelling: involves the process of associating emotions or sentiments with tweets using emoticons. For example, if a tweet contains a smiling emoticon, it could be labelled as conveying a positive emotion. Data labelling is crucial for supervised machine learning techniques where the model needs labelled data to learn how to classify emotions based on the presence of emoticons.

Emoticon Extraction: focuses on identifying and isolating emoticons from the tweets. Emoticons are specific combinations of characters or symbols that represent emotions, and this step is necessary to separate them from the

rest of the text. Extracting emoticons allows for a more focused analysis of their role in conveying emotions in tweets.

Since, in our project we already have the dataset with tweets that contain at least one emoji, so we don’t require this step.

These pre-processing steps are essential for getting the data ready for further analysis and classification. They help ensure that the text data is clean, standardized, and relevant to the research goals, particularly in the context of analysing the usage of emoticons in Twitter data.

Lemmatization:

*Stemming and Lemmatization* are methods for simplifying words. Words can appear in different forms (like "running," "ran," "runs"), but these methods help make them look the same, so it's easier to analyse. It's like making sure all variations of the word "run" are treated as "run." Stemming is a bit like cutting off parts of the word to get to its root, while lemmatization is a bit more intelligent, understanding the word's actual base form.

**The Pre-processing steps taken are:**

1. Lower Casing: Each text is converted to lowercase.
2. Replacing URLs**:** Links starting with **"http" or "https" or "www"** are replaced by **"URL"**.
3. Replacing Usernames: Replace @Usernames with word **"USER"**. *(e.g.: "@Kaggle" to "USER")*
4. Removing Non-Alphabets**:** Replacing characters except Digits and Alphabets with a space.
5. Removing Consecutive letters**:** 3 or more consecutive letters are replaced by 2 letters. *(e.g.: "Heyyyy" to "Heyy")*
6. Removing Short Words**:** Words with length less than 2 are removed.
7. Removing Stopwords**:** Stopwords are the English words which does not add much meaning to a sentence. They can safely be ignored without sacrificing the meaning of the sentence. *(e.g.: "the", "he", "have")*
8. Lemmatizing**:** Lemmatization is the process of converting a word to its base form. *(e.g.: “Great” to “Good”)*

*Our python code inputs the raw data .csv file, makes the above pre-processing modifications, removes special characters but preserves emojis and then creates a new cleaned, pre-processed file and saves it in the root folder itself.*

*For further analysis we will be using the new .csv document*.

# DATA ANALYSIS

We wanted to perform various analysis to better understand what the dataset implies by the results of those analysis. So, from the very basic frequency analysis, to using NLTK for sentiment analysis, we get results which we visualize using graphs employing the python library matplotlib.

## Emoticon Frequency analysis

* Use NumPy to count the frequency of each emoticon in the dataset.
* Create a NumPy array or Pandas Series where emoticons are the index, and frequencies are the values.
* Visualize the distribution of different emoticons in the dataset using Matplotlib. This will help us understand the prevalence of various emoticons.

We can modify the code to get top N frequent emojis in the dataset.

## K-means clustering (analysing the clusters)

Matrix X represents the occurrence of words and emojis in the text.

The number of clusters is set to num\_clusters.

Adding a new column 'cluster' to the Data Frame, indicating the cluster label for each row.

Then printing the top 10 terms (words or emojis) in each cluster based on their occurrence.

Visualizing the clusters in a scatter plot using the reduced components.

Each point is coloured based on its cluster label.

This code performs k-means clustering on a dataset containing pre-processed text and emojis. It prints the top terms in each cluster and visualizes the clusters in a 2D space after reducing the dimensionality using TruncatedSVD. The clusters represent groups of similar text and emoji patterns in the data.

## Sentiment analysis

*Sentiment Classification:*

* Apply sentiment analysis techniques to classify tweets into positive, negative, or neutral categories based on the usage of emoticons.
* Utilize pre-trained models or natural language processing libraries like NLTK to perform sentiment analysis.

*Sentiment Labeling:*

* Add a new column to our Data Frame to store the assigned sentiment labels (positive, negative, neutral) for each tweet.
* This enables us to analyse emoticon usage patterns within different sentiment categories.
* For instance, we can explore which emoticons are commonly used in positive or negative tweets.

Sentiment analysis is performed using the *SentimentIntensityAnalyzer* from the NLTK library. This analyzer assigns a sentiment score to each tweet, indicating its positivity, negativity, or neutrality.

The overall distribution of sentiments in the dataset is analyzed, and a bar chart is created to visualize the counts of positive, negative, and neutral sentiments.

The code then attempts to analyze the distribution of sentiments based on the presence of emojis in tweets. It checks for the presence of emojis using the *is\_emoji* function from the emoji library.

Finally, the code generates two subplots in a single figure. The first subplot shows the overall distribution of sentiments, and the second subplot shows the distribution of sentiments based on the presence of emojis.

## Contextual analysis

* Examine how the meaning of an emoticon changes in different contexts.
* For example, explore how the same emoticon may convey different emotions when used with different text.

*Contextual Segmentation:*

* To divide our dataset into segments or categories based on different contexts.
* Algorithm: Topic Modelling using Latent Dirichlet Allocation (LDA).
* Justification: LDA can help identify different contexts within the dataset by discovering topics based on the text content of the tweets. This can be useful for segmenting the dataset into different contextual categories.
* The identified topics are assigned to each tweet, and the distribution of tweets across topics is displayed.

The LDA algorithm has identified five topics in the dataset. Each topic is represented by a list of terms (words) that are indicative of that topic.

Topic 0: This topic seems to be related to cryptocurrency and gaming, with terms like 'eth', 'crypto', 'game', 'buy', etc.

Topic 1: This topic involves terms related to URLs, shopping, reading, and real-time insights.

Topic 2: This topic is associated with events or activities happening in April, such as Easter, streaming, sales, and happiness.

Topic 3: This topic includes terms related to checking, live streaming, free products, and codes, possibly related to online activities.

Topic 4: This topic appears to be about general discussions, with terms like 'user', 'new', 'know', 'people', etc.

Tweets assigned to Topic 1 are the most common, followed by those in Topic 4, Topic 0, Topic 2, and Topic 3.

Interpretation:

Topic 1: This topic seems to capture tweets related to URLs, shopping, and real-time insights. Users in this segment might be sharing links, insights, or engaging in e-commerce activities.

Topic 4: This topic represents more general discussions and user interactions. Users in this segment may be discussing various topics without a specific theme.

Topic 0: This topic is centered around cryptocurrency, gaming, and online activities. Users in this segment might be interested in crypto, gaming, and online purchases.

Topic 2: This topic suggests tweets related to events or activities happening in April, including Easter celebrations, streaming, and sales.

Topic 3: This topic involves tweets related to checking, live streaming, free products, and codes, indicating engagement in online activities.

*Emoticon Meaning Analysis:*

* For focusing on a specific emoticon and extracting tweets containing that emoticon, we can continue to use Pandas for data filtering.
* Define the emoticon we want to analyse and filter tweets accordingly.

WITH VISUALIZATION

The code leverages spaCy's word vectors to obtain embeddings for selected emojis, applies t-SNE for dimensionality reduction, and creates an interactive 2D scatter plot to visualize the spatial relationships between the emojis. The visualization provides insights into the similarity or dissimilarity between the selected emojis in the reduced dimensional space.

WITHOUT VISUALIZATION

The provided Python code is focused on obtaining word embeddings for a selection of emojis using a pre-trained spaCy model with word vectors. The embeddings are then saved to a DataFrame and exported to a CSV file for potential further analysis.

The code utilizes spaCy's pre-trained model to obtain word embeddings for each selected emoji.

The resulting embeddings are stored in emoji\_embeddings, forming a matrix of vectors.

The DataFrame has two columns: 'Emoji' and 'embedding', where 'embedding' contains the word vectors converted to a list.

In this hypothetical example, each emoji ('👉', '👈', ..., '💙') is associated with a list of numerical values representing its word embedding. The actual numbers in the embedding list are the components of the vector in the high-dimensional space. These values capture the semantic meaning of the emojis based on the pre-trained spaCy model.

This CSV file essentially provides a record of the embeddings for the selected emojis, making it convenient for further analysis or visualization of the spatial relationships between these emojis in a vector space.

For further contextual analysis, we perform LDA again but this time mentioning the context of the tweet in an added column and saving the changes to a new .csv document.

The DataFrame with the new 'Context' column is saved to a new CSV file named 'Emoji\_Preserved\_Cleaned\_Contexts.csv' using to\_csv, excluding the index column.

*Sentiment Analysis within Contexts:*

* Apply sentiment analysis techniques to the tweets within each context category. We can use Natural Language Processing (NLP) libraries like NLTK or spaCy to perform sentiment analysis.
* NLTK provides tools for sentiment analysis, including pre-trained models that can classify tweets as positive, negative, or neutral based on their text content.

We extend the analysis of tweets by incorporating sentiment analysis using the NLTK library. The sentiment scores are assigned to each tweet based on its textual content, and the distribution of sentiments within different contexts is displayed.

A new column 'Sentiment' is added to the DataFrame to capture the sentiment of each tweet.

The get\_sentiment function is defined to assign sentiments ('positive', 'negative', 'neutral') based on the compound score obtained from the sentiment analyzer.

The sentiments are determined by comparing the compound

score to predefined thresholds.

The code groups the DataFrame by 'Context' and 'Sentiment', counting the occurrences of each sentiment within each context.

The resulting distribution is displayed in a tabular format, indicating how many tweets with each sentiment are present in each context.

The DataFrame with the added 'Sentiment' column is saved to a new CSV file named 'Emoji\_Preserved\_Cleaned\_Contexts\_Sentiments.csv' using to\_csv, excluding the index column.

*Emoticon Frequency in Contexts***:**

* To calculate the frequency of the chosen emoticon in each context category using Pandas. We will use Pandas functions like groupby and value\_counts to count the frequency of the emoticon in each context.
* A list of target emojis (target\_emojis) is specified. These are the emojis for which we want to analyze the frequency in each context.
* For each target emoji, a new column is added to the DataFrame to store the frequency of that emoji in each tweet.
* The str(text).count(emoji) function is used to count the occurrences of each target emoji in the 'Text' column for each tweet.
* These new columns are named '👉\_Frequency', '👈\_Frequency', ..., '💙\_Frequency'.
* The DataFrame is then grouped by 'Context', and the sum of emoji frequencies is calculated for each context.
* This is achieved using the groupby and sum functions, resulting in a new DataFrame (emoticon\_frequency\_by\_context) that shows the total frequency of each target emoji in each context.

# CONCLUSION

The analysis of emoji usage in the Twitter dataset has provided valuable insights into the patterns, sentiments, and contexts associated with emojis in a diverse range of tweets. The project involved several key steps and analyses, each contributing to a comprehensive understanding of how emojis are employed in this particular dataset.

1. Data Loading and Pre-processing:

The project commenced with the loading of the Twitter dataset, named 'Emoji\_Preserved\_Cleaned\_Processed\_Data.csv,' into a Pandas DataFrame. This step ensured that the subsequent analyses were conducted on clean and processed data.

2. Topic Modelling with Latent Dirichlet Allocation (LDA):

Latent Dirichlet Allocation (LDA) was applied to uncover latent topics within the tweet dataset. The resulting topics provided a thematic categorization of tweets, revealing the underlying structures and prevalent themes in the data.

3. Contextual Analysis of Emojis:

Building upon the topic modelling, the project delved into the analysis of emoji usage within different contexts. Each tweet was assigned a topic, and specific emojis were targeted for analysis. The frequency of these emojis within each context was quantified, shedding light on the role and prevalence of emojis across different thematic areas.

4. Sentiment Analysis:

Sentiment analysis was conducted to understand the emotional tone associated with the tweets. The NLTK library was employed to assign sentiments (positive, negative, neutral) based on the textual content. The sentiment distribution within each context was visualized, providing additional layers of insight into the emotional aspects of tweets.

5. Visualization of Emoji and Sentiment Distribution:

Seaborn and Matplotlib were utilized to create visualizations that enhanced the interpretation of the results. A countplot showcased the distribution of sentiments within each context, offering a clear and intuitive representation of the emotional tones associated with different topics.

6. Emoji Embeddings and t-SNE Visualization:

The analysis extended to understanding the spatial relationships between selected emojis using word embeddings. The t-distributed Stochastic Neighbour Embedding (t-SNE) technique reduced the dimensionality of emoji embeddings, allowing for a visual exploration of their proximity in a 2D space.

7. Conclusion and Future Work:

In conclusion, the project has provided a multi-faceted analysis of emoji usage in the given Twitter dataset. Insights into thematic categorizations, sentiment distributions, and spatial relationships between emojis contribute to a nuanced understanding of the dataset.

Future work could involve exploring more advanced techniques for sentiment analysis, such as deep learning approaches. Additionally, the analysis could be extended to incorporate user-specific behaviours, temporal trends, or even the impact of specific events on emoji usage.

8. Final Outputs and Contributions:

The project has generated several outputs, including CSV files with enriched information (e.g., contexts, sentiments), visualizations depicting sentiment distributions, and spatial embeddings of selected emojis. These outputs serve as valuable resources for further exploration and deeper analyses.

In summary, the analysis of emoji usage in the Twitter dataset has not only uncovered patterns and themes but has also contributed to the broader understanding of how emojis play a role in expressing sentiments and conveying contextual information on the platform.

# REFERENCES

<https://visualstudio.microsoft.com/visual-cpp-build-tools/>

<https://www.nltk.org/data.html>

<https://www.kaggle.com/code/pzivic/twitter-sentiment-analysis-pa3lo>

<https://www.kaggle.com/code/karan842/twitter-sentimental-analysis>

<https://www.kaggle.com/code/virajanand3317/naive-bayes-classifier>

[*https://www.kaggle.com/code/stoicstatic/twitter-sentiment-analysis-for-beginners*](https://www.kaggle.com/code/stoicstatic/twitter-sentiment-analysis-for-beginners)

# An empirical study of emoji usage on Twitter in linguistic and national contexts- <https://www.sciencedirect.com/science/article/pii/S2468696421000318>

Emoji Sentiment Analysis on Twitter Data" by Novak et al. (2015)- <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0144296>

Twitter Sentiment Analysis Using Emoticons and Emoji Ideograms

<https://www.researchgate.net/publication/308413240_TWITTER_SENTIMENT_ANALYSIS_USING_EMOTICONS_AND_EMOJI_IDEOGRAMS>

An empirical study of emoji usage on Twitter in linguistic and national contexts

<https://www.sciencedirect.com/science/article/pii/S2468696421000318>

Sentiment Analysis of Tweets Including Emoji Data

<https://ieeexplore.ieee.org/document/8560896>

Sentiment analysis of text and emoji data for twitter network

<https://bjeps.alkafeel.edu.iq/cgi/viewcontent.cgi?article=1034&context=journal>

Sentiment analysis of social media Text\-Emoticon Post with Machine learning Models Contribution

<https://iopscience.iop.org/article/10.1088/1742-6596/2070/1/012079/pdf>

Twitter Sentiment Analysis via Bi-sense Emoji Embedding and Attention-based LSTM

<https://dl.acm.org/doi/pdf/10.1145/3240508.3240533>

Investigating Redundancy in Emoji Use: Study on a Twitter Based Corpus

<https://aclanthology.org/W17-5216.pdf>