C&S Analytics Project

Social Media Sentiments Analysis

Team Bosch B2: Becky Wang, Chieh-Yi Chou, Hsing-Chia Tsai, Jiajian Wang, Soujanya Bharadwaj, Yung-Hsin Lin

University of California, Irvine MSBA

Instructor: Professor Sanjeev Dewan

Table of Contents

1. Introduction

In today's digital age, social media platforms such as Twitter, Facebook, and Instagram have become crucial spaces where people express their opinions, share experiences, and react to various topics. These platforms generate vast amounts of unstructured data every day, creating valuable insights for businesses, marketers, and decision-makers. However, extracting meaningful information from this large pool of data can be overwhelming without proper analysis.

Social Media Sentiment Analysis is a process that uses NLP and machine learning techniques to evaluate the emotional tone of social media content, such as posts, comments, and reviews. This analysis categorizes the sentiments expressed as positive, negative, or neutral. By applying sentiment analysis, businesses can understand customer opinions, monitor brand reputation, and make informed decisions based on real-time public feedback.

1.1 Relevance and Impact

The importance of Social Media Sentiment Analysis lies in its ability to unlock valuable insights from the vast data generated on social media platforms. In today's competitive business environment, it's essential for companies to understand public perception and customer sentiment to stay ahead of the curve. Sentiment analysis enables organizations to track customer reactions to new products or marketing campaigns, detect emerging trends, and address customer concerns before they escalate.

By leveraging sentiment analysis, companies can refine their strategies, enhance customer engagement, and improve brand loyalty. It also helps businesses identify potential opportunities or risks, giving them a better understanding of customer preferences and needs. In short, Social Media Sentiment Analysis is a powerful tool that drives data-driven decision-making and supports companies in building stronger relationships with their audience.

2. Data Analysis

The Social Media Sentiment Analysis Dataset, which we sourced from Kaggle, captures a vibrant tapestry of emotions, trends, and interactions across various social media platforms. This dataset provides a snapshot of user-generated content, encompassing text, timestamps, hashtags, countries, likes, and retweets. Each entry unveils unique stories—moments of surprise, excitement, admiration, thrill, contentment, and more—shared by individuals worldwide. By analyzing this dataset, we can gain a deeper understanding of the sentiments driving online conversations, helping us track public reactions to various events, products, or trends and make informed, data-driven decisions.

2.1 Data Summary

The dataset provides a detailed exploration of factors influencing Social Media Sentiments among users.

Table 2-1 Data features with description

Feature	Description
Text	User-generated content showcasing sentiments
Sentiment	Categorized emotions (positive, negative, neutral)
Timestamp	Date and time information
User	Unique identifiers of users contributing
Platform	Social media platform where the content originated (Twitter ,Facebook, Instagram)
Hashtags	Identifies trending topics and themes
Likes	Quantifies user engagement(likes)
Retweets	Reflects content popularity(retweets)

Country	Geographical origin of each post on social media platform
Year	Year of the post
Month	Month of the post
Day	Day of the post
Hour	Hour of the post

2.2 Data Pre-processing

Missing Value Analysis

As part of the data preprocessing step, the dataset was examined for missing values. A column-wise check was performed to identify any features with incomplete data. The analysis revealed that the dataset contained no missing values across all features. This result indicates that the dataset is complete and requires no removal of rows or columns due to missing entries. The completeness of the dataset ensures reliability for further analysis.

Text Normalization

Performing text preprocessing and standardization on the "Text" column, ultimately generating the "Clean_Text" column. The program initializes the Porter Stemmer and loads the NLTK English stopwords library. The function executes a series of text-cleaning operations, including converting the input text to lowercase and using regular expressions to remove content within square brackets, URLs, extra spaces, HTML tags, punctuation, newline characters, alphanumeric words, and non-ASCII characters. Next, the program utilizes NLTK's word tokenization process to segment text into individual words, removes stopwords to eliminate common but non-informative terms, and applies the Porter Stemmer to perform word stemming, reducing words to their root form for consistency and normalization.

VADER Sentiment Analysis

We use VADER's polarity_scores method to extract the compound score for sentiment classification and evaluation of "Text". Based on the score range, a compound score of 0.05 or higher is classified as positive, a score of -0.05 or lower is classified as negative, and scores in between are categorized as neutral.

	Clean_Text	Vader_Score	Sentiment
0	enjoy beauti day park	0.4939	positive
1	traffic terribl morn	0.0000	neutral
2	finish amaz workout	0.0000	neutral
3	excit upcom weekend getaway	0.0000	neutral
4	tri new recip dinner tonight	0.0000	neutral
..	...	...	...
95	confus reign tri make sens recent event	0.0000	neutral
96	excit build surpris birthday parti	0.0000	neutral
97	wit act kind made day	0.5267	positive
98	pride complet challeng fit challeng	0.5994	positive
99	moment shame speak injustic	-0.4767	negative

3. Exploratory Data Analysis (EDA)

3.1 Sentiment Distribution

The pie chart visualizes the proportion of positive, negative, and neutral sentiments across Facebook, Instagram, and Twitter. It highlights platform-specific sentiment trends, providing insights into user engagement and public perception.

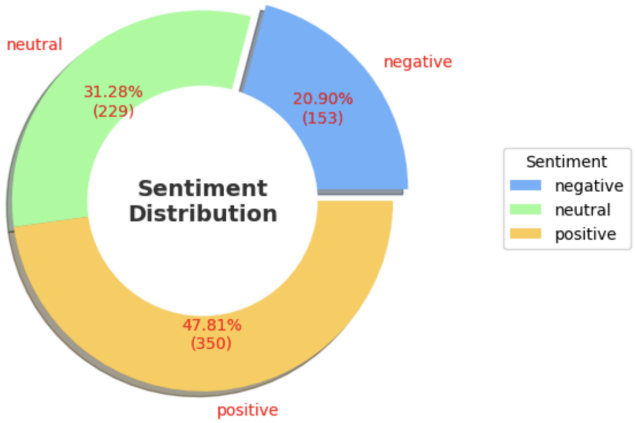


Figure 3-1 Sentiment Distribution

3.2 Country-Based Analysis

The bar chart illustrates the relationship between sentiment and country, highlighting the USA, Canada, and the UK as the countries with the highest positive sentiment at least three times more than others, followed by moderate neutral sentiment, and relatively lower negative sentiment. This suggests that discussions or engagements from these countries tend to be more optimistic compared to others. Additionally, the significant number of neutral sentiments indicates a balanced perspective, while the lower proportion of negative sentiments reflects a generally favorable outlook. In contrast, countries like France, Brazil, Germany, and Italy exhibit considerably lower engagement across all sentiment categories, suggesting either lower participation or a more evenly distributed sentiment profile. Therefore, we have decided to focus our future analyses primarily on the USA, Canada, and the UK while excluding countries with lower engagement and sentiment representation.

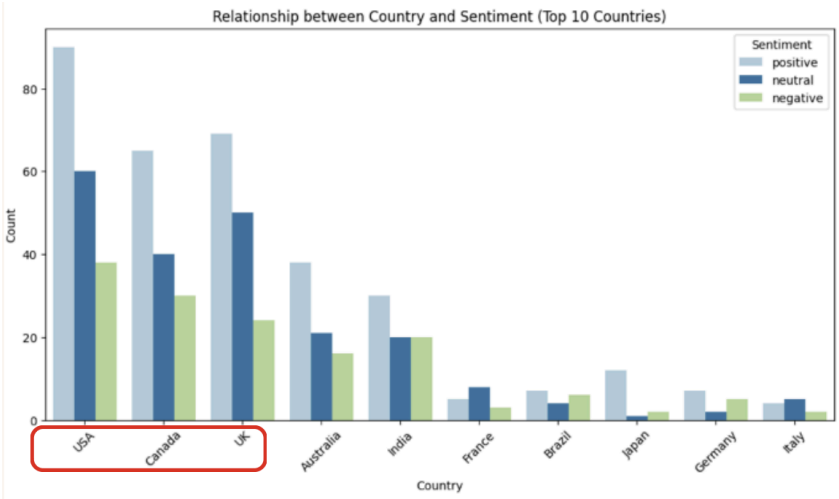


Figure 3-2 Relationship between Country and Sentiment (Top 10 Countries)

3.3 Platform-Based Analysis

Instagram, with 36% usage, is the most popular platform among users from the three largest countries(Figure 3-3)—USA, Canada, and the UK. The high engagement from these regions, coupled with their predominantly positive sentiment, significantly contributes to Instagram having the highest positive sentiment overall. This indicates that users from these countries not only prefer Instagram but also engage with content in a more favorable and optimistic manner.

Given this trend, Instagram's strong presence in these English-speaking markets suggests its potential as a primary platform for positive brand interactions and audience engagement.

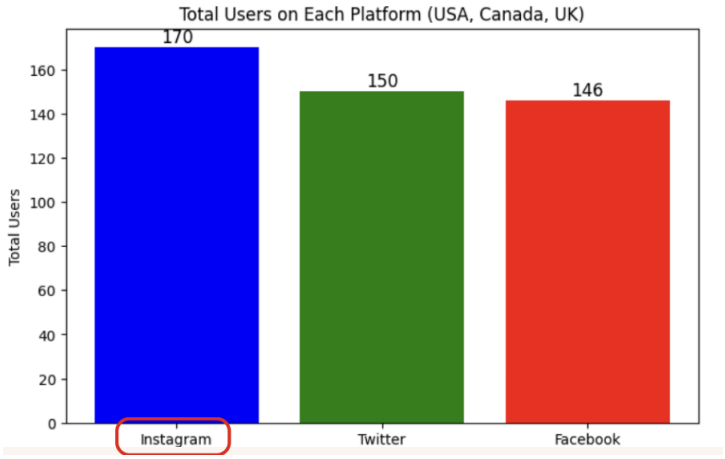


Figure 3-3 Total Users on Each Platform

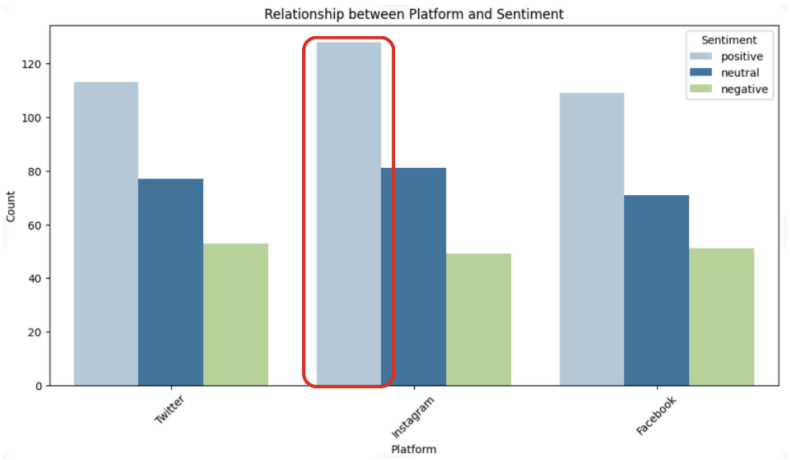


Figure 3-4 Relationship between Platform and Sentiment

3.4 Time-Based Analysis

In addition to analyzing platforms and countries, we first conduct an analysis based on months to explore the relationship between time and sentiment. June records the highest number of positive sentiments, indicating a peak in positive expressions, which may be attributed to seasonal events, holidays, or increased engagement during this period. In contrast, September exhibits the highest count of negative sentiments, suggesting a shift in sentiment, possibly influenced by external factors such as post-summer transitions, back-to-school periods, or broader societal trends.

This pattern highlights the seasonal variation in sentiment trends, providing valuable insights for optimizing content strategies and engagement planning.

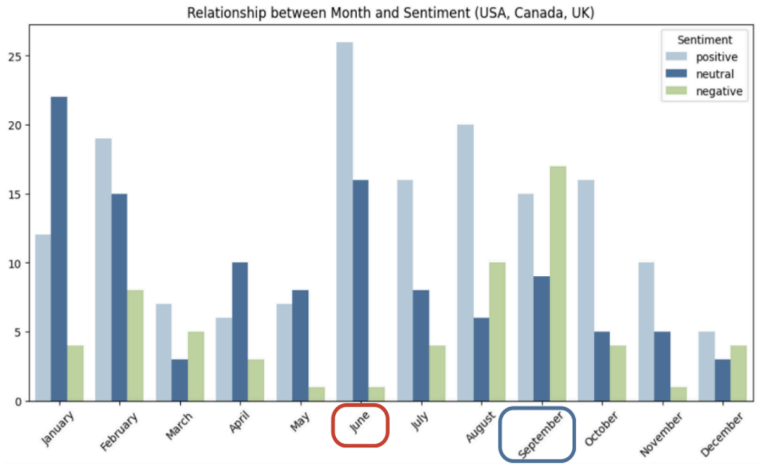


Figure 3-5 Relationship between Month and Sentiment

Next, we delve deeper into the impact of weekdays on sentiment to understand how sentiment varies across different days of the week. Our analysis reveals that Sunday records the highest count of positive sentiments, while Saturday has the lowest, which contradicts common sense. Typically, Saturday is expected to have higher positive sentiment, as it is a day of leisure and social activities, while Sunday might be associated with increased negativity due to the anticipation of returning to school or work. Given this discrepancy, we have decided to validate our findings by utilizing a model to check for accuracy and ensure the reliability of our sentiment analysis.

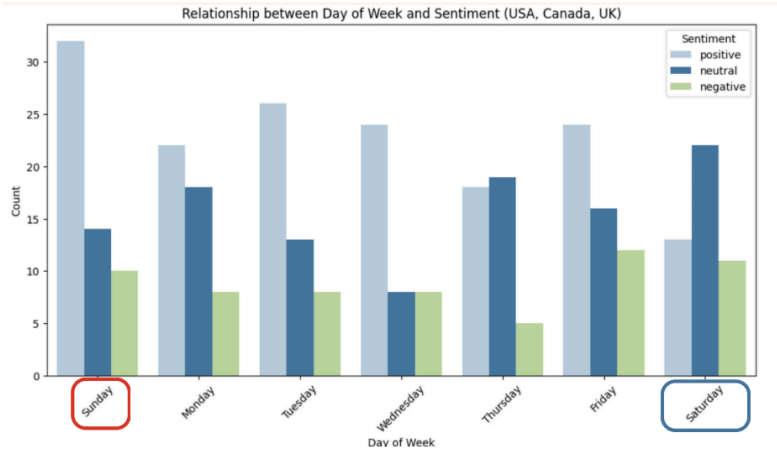
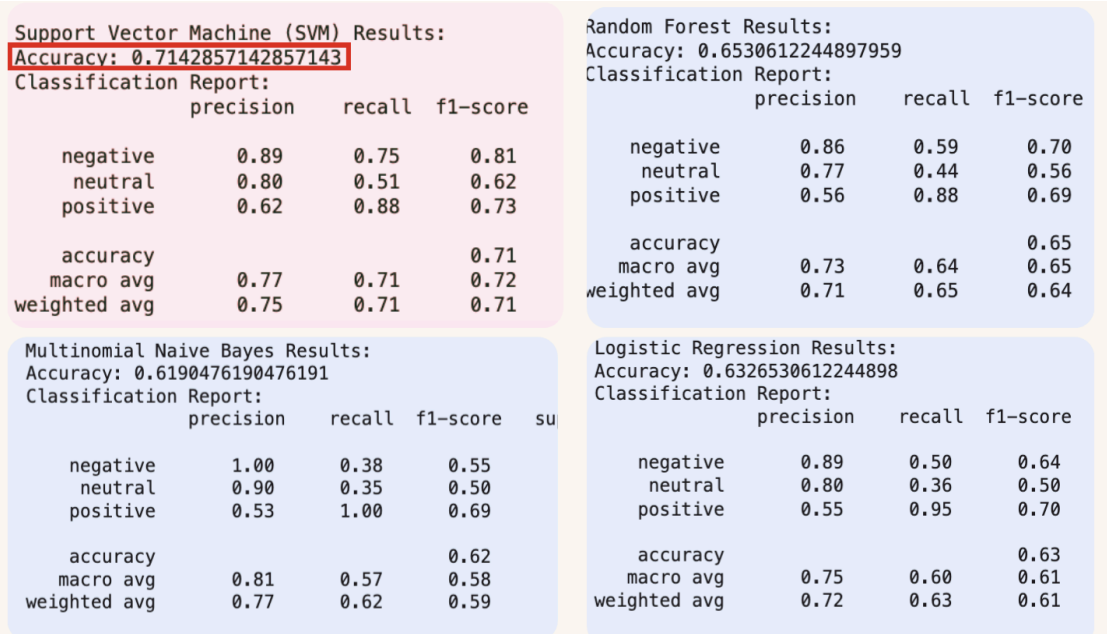


Figure 3-6 Relationship between Day of Week and Sentiment

4. Model Description

While testing accuracy across various models, including SVM, Naive Bayes, Random Forest, and Logistic Regression, we found that the SVM model achieved the highest accuracy at 0.71, outperforming Naive Bayes (0.62), Random Forest (0.65), and Logistic Regression (0.63). Based on this result, we selected SVM as our model for prediction. Additionally, we observed that across all models, the accuracy for negative sentiment was consistently high, reaching around 90%, while the accuracy for neutral and positive sentiments remained relatively low. To address this, we conducted a deeper analysis of the text distribution within these two sentiment categories.



Support Vector Machine (SVM) Results:
Accuracy: 0.7142857142857143
Classification Report:

	precision	recall	f1-score
negative	0.89	0.75	0.81
neutral	0.80	0.51	0.62
positive	0.62	0.88	0.73
accuracy			0.71
macro avg	0.77	0.71	0.72
weighted avg	0.75	0.71	0.71

Random Forest Results:
Accuracy: 0.6530612244897959
Classification Report:

	precision	recall	f1-score
negative	0.86	0.59	0.70
neutral	0.77	0.44	0.56
positive	0.56	0.88	0.69
accuracy			0.65
macro avg	0.73	0.64	0.65
weighted avg	0.71	0.65	0.64

Multinomial Naive Bayes Results:
Accuracy: 0.6190476190476191
Classification Report:

	precision	recall	f1-score	su
negative	1.00	0.38	0.55	
neutral	0.90	0.35	0.50	
positive	0.53	1.00	0.69	
accuracy			0.62	
macro avg	0.81	0.57	0.58	
weighted avg	0.77	0.62	0.59	

Logistic Regression Results:
Accuracy: 0.6326530612244898
Classification Report:

	precision	recall	f1-score
negative	0.89	0.50	0.64
neutral	0.80	0.36	0.50
positive	0.55	0.95	0.70
accuracy			0.63
macro avg	0.75	0.60	0.61
weighted avg	0.72	0.63	0.61

Figure 4-1 Model Analysis

4.1 Text Classification

We found lower accuracy, particularly in distinguishing positive and neutral sentiments, likely due to the classification criteria. When analyzing the frequency of common words, we observed that many words typically associated with positive sentiment in common sense were instead classified as neutral by the VADER package, which calculates sentiment scores numerically.

As a result, numerous positive words were misclassified as neutral, impacting the model's accuracy for these two categories. For example, the word "new," which is generally perceived as positive, was classified as neutral in this dataset.

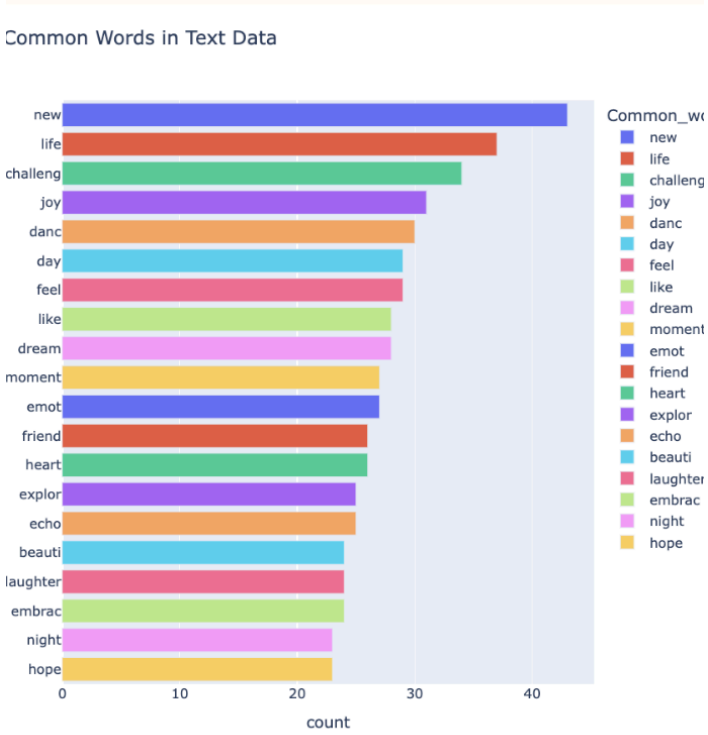


Figure 4-2 Common words in Text Data

Common_neutral words	count
new	22
explor	13
excit	12
life	12
beauti	12
night	12
danc	12
attend	11
seren	11
feel	10

Common_positive words	count
joy	30
friend	24
laughter	24
new	21
challeng	20
life	20
hope	20
dream	20
embrac	19
like	19

Figure 4-3 Common words and count

5. Temporal Sentiment Patterns and Business Implications

Our analysis reveals significant temporal patterns in social media sentiment that have important implications for businesses and marketers:

Seasonal Variations

June exhibited the highest positive sentiment, while September showed the peak in negative sentiment. This pattern suggests a potential correlation with summer activities and the back-to-school period, respectively. Businesses could leverage this insight by:

- Intensifying positive marketing campaigns and product launches in June to capitalize on the optimistic mood.
- Implementing supportive or empathetic messaging in September to counteract the negative sentiment trend.

Weekly Sentiment Fluctuations

Contrary to expectations, Sunday recorded the highest positive sentiment, while Saturday showed the lowest. This unexpected finding warrants further investigation and could inform social media engagement strategies:

- Businesses might consider increasing their social media activity on Sundays to align with the positive sentiment peak.
- Content strategies for Saturdays may need to be reevaluated to better resonate with users' moods.

Platform-Specific Trends

Instagram emerged as the most popular platform among users in the USA, Canada, and the UK, with the highest positive sentiment. This suggests that:

- Brands targeting these markets should prioritize Instagram for positive engagement and brand-building activities.
- Content strategies for Instagram should focus on maintaining and amplifying the existing positive sentiment.

By incorporating these temporal and platform-specific insights into their social media strategies, businesses can optimize their engagement timing, tailor their messaging, and potentially improve their overall social media performance and brand perception.

6. Conclusion

The SVM model, with an accuracy of 71%, proved to be the most effective in predicting sentiment performance. Our analysis reveals that English-speaking countries (USA, Canada, and the UK) exhibit a more optimistic outlook, characterized by higher positive sentiment and lower negativity compared to other regions. Additionally, Instagram emerged as the most popular platform in these countries, garnering the highest positive sentiment, while neutral and negative sentiments remained consistent across all platforms.

In terms of time-series analysis, some anomalies were observed due to classification criteria, particularly in distinguishing positive and neutral sentiments. Notably, the accuracy for negative sentiment classification remained exceptionally high at around 90%, reinforcing the reliability of negative sentiment predictions. This also supports the conclusion that June had the lowest negative sentiment, while September exhibited the highest negative sentiment, likely influenced by the beginning and end of summer break period. These sentiment patterns provide valuable insights for businesses, enabling them to strategically allocate resources and optimize engagement to maximize their impact.