# Introduction

The dataset of this experiment includes user posts on social media, recording the user's text content, emotional tendencies, posting time, user ID, social platform, related tags, reposts, likes, country, as well as specific information such as year, month, day, and hour. The dataset covers users from multiple countries around the world, involving social platforms such as Twitter, Instagram, and Facebook. Emotional tags include various emotional states such as positive, negative, and neutral. These data provide us with rich materials for analyzing user emotional expression.

The prediction task of this experiment is to explore the impact of different factors (such as time, location, platform) on user emotional expression. We aim to predict the distribution of user emotions under specific conditions through statistical analysis and machine learning methods. For example, we may predict the probability of a user posting positive emotional posts on a specific social platform during a certain period of time. In addition, we will also explore the differences in emotional expression among users in different countries, as well as the possible cultural and social factors behind these differences.

Through analysis of the dataset, we found that posts with positive emotions accounted for a higher proportion, indicating that social media users tend to share positive emotions and experiences. There are significant differences in the distribution of emotions on different social platforms, with a relatively high proportion of negative emotional posts on Twitter, which may be related to the real-time and news nature of the platform. Time factor analysis shows that during weekends and holidays, users express positive emotions more frequently, which may be related to an increase in leisure time and frequent social activities. Regional analysis reveals differences in emotional expression among users from different countries, which may be related to differences in cultural backgrounds and social environments in different regions. These findings are of great significance for understanding the emotional dynamics of global social media users and provide valuable references for content recommendation and sentiment analysis on social platforms.

# Data Description

The data for this experiment is sourced from the Kaggle official website. The CSV file with the data name "sentiment dataset. csv" contains post data collected from social media platforms, covering public posts from different users, including text content, emotional tendencies, timestamps, and other information.

Data Description:

1. Text content: The original text posted by users on social media, reflecting their views, emotions, and experiences.
2. Sentiment: The emotional tendency of a post, marked as "Positive", "Negative", "Neutral", and other categories.
3. Timestamp: The specific date and time when the post was published, in the format of "YYYY-MM-DD HH: MM: SS".
4. User: The social media user ID or nickname who posted the post.
5. Platform: A social media platform for posting, such as Twitter, Instagram, Facebook, etc.
6. Hashtags: tags or topics included in a post, used for classification and searching.
7. Retweeks: The number of times a post has been forwarded by other users.
8. Likes: The number of likes received by a post.
9. Country: The country or region where the user is located.
10. Year, Month, Day, Hour: The time information of the post publication, which records the year, month, day, and hour respectively.

This dataset provides researchers with a comprehensive perspective to analyze and understand the emotional expressions of social media users and their influencing factors. Through in-depth analysis of these data, the impact of different cultures, times, and social environments on user emotions can be revealed, providing support for social media analysis, sentiment analysis research, and marketing strategies.

# Models and Methods

This experiment utilized data visualization methods and machine learning classification models. Machine learning classification models include Logistic Regression RandomForestClassifier、SVC。

Firstly, remove invalid information such as stop words in the comments, and use SentimentSentityAnalyzer to perform sentiment analysis on each user comment, determining whether the statement belongs to positive, negative, or neutral categories. The results are as follows:A diagram of a positive distribution

Description automatically generated

Emotion analysis chart

This graph is used to display statistical data on emotional distribution. The figure is divided into three parts, representing negative, neutral, and positive emotions respectively. According to the data in the figure, positive emotions account for the largest proportion, at 47.81%, with a total of 350 samples; Next is neutral emotion, accounting for 31.28% with 229 samples; The proportion of negative emotions is the smallest, accounting for 20.90%, with 153 samples. This distribution indicates that in the analyzed dataset, the expression of positive emotions is significantly higher than that of negative and neutral emotions, which may mean that users are more inclined to share positive experiences and emotions on social media. In addition, the graph clearly displays the relative frequencies of various emotions through different colors and percentages, providing an intuitive visual tool for understanding emotional tendencies on social media.

Next, further analyze the changes in the data of the three emotional categories each year and draw a bar chart for analysis.A graph of a number of years

Description automatically generated

The results show the distribution of positive, neutral, and negative emotions in different years from 2010 to 2023. The vertical axis of the chart represents the number of emotional posts, while the horizontal axis represents the year. It can be observed that the number of posts with positive emotions dominated in all years, especially in 2022 and 2023, with a significant increase and reaching its peak in 2023. The number of posts with neutral emotions is relatively stable, without significant fluctuations, while the number of posts with negative emotions is relatively small but also shows a trend of increasing year by year.

Since 2015, the number of posts with various emotions has started to increase, which may be related to the popularity of social media and the growth of user base. From 2018 to 2020, the number of posts on various emotions remained relatively stable, but starting from 2021, the number of posts on positive emotions began to significantly increase, which may be related to global events, changes in social media algorithms, or shifts in user behavior patterns.

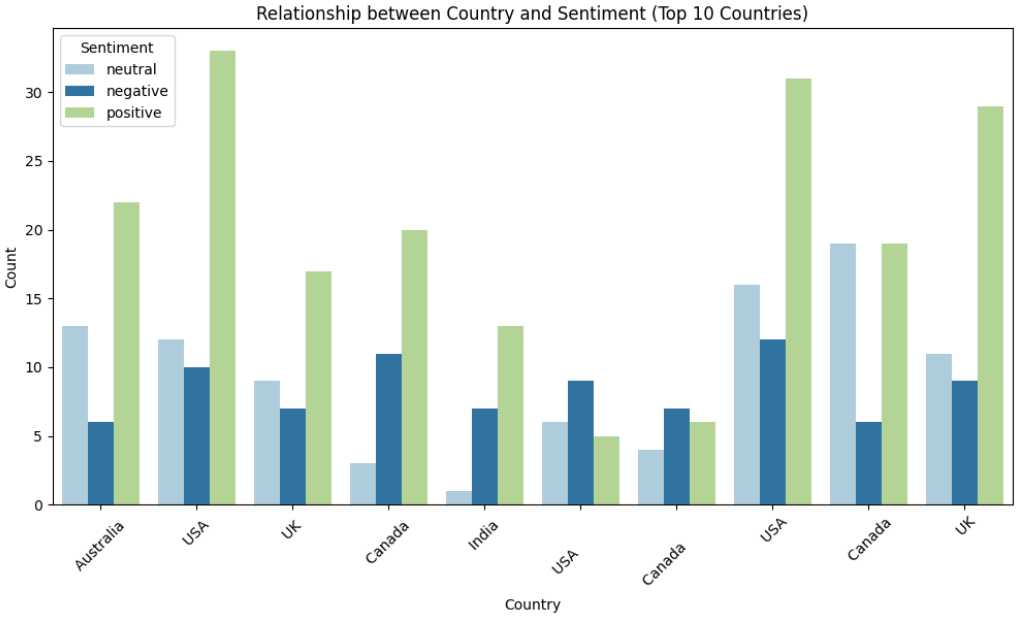
Analyze the proportion of emotion categories on different social platforms:

A graph of a diagram

Description automatically generated with medium confidence

The result table shows the distribution of users' emotional tendencies on different social media platforms (Twitter, Instagram, Facebook). In the chart, positive emotions are represented in light blue, neutral emotions are represented in dark blue, and negative emotions are represented in light green. From the graph, all three platforms have the highest number of positive emotional posts, especially Instagram, which has the highest number of positive emotional posts, close to 130. The number of positive emotional posts on Twitter and Facebook is similar, but slightly lower than on Instagram. Neutral sentiment posts are evenly distributed across the three platforms, with Facebook slightly more than Twitter and Instagram in terms of quantity. The number of posts with negative emotions is the lowest on all platforms, and the difference in quantity among the three platforms is not significant. This emotional distribution may reflect users' behavior habits and preferences on different social media platforms, for example, Instagram may be more inclined to share positive moments of life, while Twitter may be more used for daily communication and information sharing. Overall, users tend to express and spread positive content on social media.

Next, analyze the sentiment categories of users in different countries:



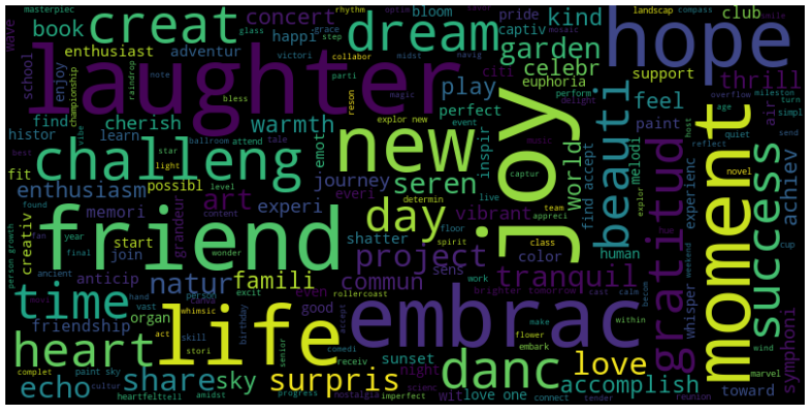
In the chart, positive emotions are represented in light green, neutral emotions are represented in light blue, and negative emotions are represented in dark blue. From the graph, it can be seen that there are significant differences in emotional expression among different countries. The United States (USA) has the highest number of positive emotional expressions, exceeding 30, indicating a relatively positive emotional tendency. Following closely behind is India, with a relatively high number of posts expressing positive emotions. Australia and Canada have similar numbers of positive emotional expressions, both around 20, showing a relatively optimistic social media atmosphere.

The UK has a higher number of neutral emotions expressed, slightly higher than positive emotions, which may reflect users' tendency towards more neutral or conservative expression. Canada also has a relatively large number of neutral emotional posts, but positive emotional expressions still dominate.

It is worth noting that although the majority of posts express positive emotions in all countries, the expression of negative emotions is reflected in all countries. Among them, India has a relatively large number of posts expressing negative emotions, which may be related to the country's socio-economic situation, population pressure, and other factors.

Overall, this chart reveals the emotional tendencies of different countries on social media, with positive emotions generally dominating, but the distribution of neutral and negative emotions also provides an important perspective for us to understand the emotional dynamics of social media users in various countries. These pieces of information are of significant value in fields such as social media strategy, international market analysis, and cross-cultural communication research.

Draw a word cloud map for the comment text as follows:



From the graph, it can be seen that words such as life, new, friend, and joy are the most common.

Next, the data will be divided into training and testing sets for model building. Treat data text as features and data sentiment categories as labels. Build Logistic Regression, RandomForestClassifier, and SVC models separately for training and analysis.

Firstly, TF-IDF is used to process the raw text data. In machine learning and natural language processing (NLP) tasks, TF-IDF (Term Frequency Across Document Frequency) is a commonly used text feature extraction method. It can convert textual data into numerical features, making it easier for machine learning models to process.

Firstly, a 'TfidfVectorizer' object was created, which is a class in the 'sklearn. feature.exe. text' module specifically designed to convert text data into TF-IDF representations` The max\_features=5000 parameter specifies the maximum number of features (i.e. words) in the TF-IDF vector, which means that the vectorizer will consider the 5000 words with the highest TF-IDF values in the text. Then two operations were performed: first, 'fit', which learns the vocabulary and TF-IDF values for each word based on the training set 'X\_train' (which contains the text content of the training data); Next is' transform ', which converts the training data into TF-IDF vectors. Then, a transformation operation was performed on the test set 'X\_test', using the vocabulary and TF-IDF values already learned from the training set to transform the text data of the test set. Note that 'fit' is not used here because the vocabulary has already been defined on the training set, which ensures that the same feature space is used for training and testing.

Train the processed data separately into Logistic Regression, Random Forest Classifier, and SVC models, and then test the trained models using test data.

# Results and Interpretation:

The results of the Logistic Regression model are as follows:

A screenshot of a computer screen

Description automatically generated

The overall accuracy of the model is 0.7755, which means approximately 77.55% of the predictions are correct. This result indicates that the model is able to make accurate sentiment classification in most cases.

The classification report provides more detailed performance metrics, including precision, recall, F1 score, and support.

1. Negative emotion: The model has an accuracy of 0.81 in identifying negative emotions, a recall rate of 0.57, and an F1 score of 0.67. This indicates that the model is able to identify posts with negative emotions with an accuracy of 81%, but the recall rate is relatively low, meaning that a considerable portion of negative emotion posts were not correctly identified.
2. Neutral sentiment: For neutral sentiment, the model has an accuracy of 0.71, a recall of 0.60, and an F1 score of 0.65. This shows that the model performs poorly in identifying neutral sentiment posts, with some margin of error.
3. Positive sentiment: The model performs the best on positive sentiment, with an accuracy of 0.79, a recall rate of up to 0.92, and an F1 score of 0.85. This indicates that the model can accurately identify posts with positive emotions in most cases.

Macro AVG provides a simple average performance metric that does not consider the number of samples in each category. Its accuracy, recall, and F1 score are 0.77, 0.69, and 0.72, respectively. Weighted average (avg) takes into account the sample size of each category, providing a more comprehensive model performance metric with accuracy, recall, and F1 score of 0.77.

The classification results of RandomForestClassifier are as follows:

A screenshot of a computer

Description automatically generated

The performance of the random forest model in sentiment analysis tasks is slightly better than the previous logistic regression model. The overall accuracy of the model is 0.7823, or 78.23%, indicating a high level of accuracy in predicting emotional categories. The classification report shows that the model performs the best in identifying negative emotions, with an accuracy of 0.89, a recall rate of 0.70, and an F1 score of 0.78. This indicates that the model can identify posts with negative emotions with high accuracy and make reliable predictions about negative emotions. In the recognition of neutral emotions, the accuracy of the model is 0.66, the recall rate is 0.72, and the F1 score is 0.69, indicating that the model has some error in predicting neutral emotions, but can still recognize neutral emotions well. In terms of positive emotion recognition, the accuracy of the model is 0.82, the recall rate is 0.83, and the F1 score is 0.83, indicating that the model performs well in positive emotion recognition and can accurately identify most positive emotion posts. The accuracy, recall, and F1 score of macro avg and weighted avg are 0.79, 0.75, and 0.77, respectively. The random forest model provides a relatively reliable prediction in sentiment analysis tasks, especially in the recognition of negative and positive emotions.

The SVC model results are as follows:

A screenshot of a computer screen

Description automatically generated

The performance of support vector machine (SVM) models in sentiment analysis tasks shows different strengths and weaknesses. The overall accuracy of the model is 0.7007, which is 70.07%, slightly lower than the aforementioned random forest model. In the recognition of negative emotions, the SVM model has a high accuracy of 0.82, but a low recall rate of only 0.39. This indicates that the model can accurately identify posts with negative emotions, but at the same time, it misses a considerable number of posts with negative emotions. In terms of neutral emotion recognition, the accuracy of the model is 0.70, the recall rate is 0.40, and the F1 score is 0.51, indicating that there is a certain degree of error in neutral emotion prediction, and there are many neutral emotion posts that have not been correctly recognized. In terms of recognizing positive emotions, the model performs well with an accuracy of 0.69, a recall rate of up to 0.93, and an F1 score of 0.79, indicating that the model is reliable in recognizing positive emotions and can comprehensively identify posts with positive emotions. The accuracy, recall, and F1 score of macro average and weighted average are 0.73, 0.57, and 0.61, respectively, as well as 0.71, 0.70, and 0.67

# Conclusion and Next Steps:

Summary:

After comparing the performance of three different machine learning models for sentiment analysis tasks - logistic regression, random forest, and support vector machine - we can see that each model has its own advantages and areas for improvement. The logistic regression model provides a balanced performance but has shortcomings in recalling negative emotions. The random forest model shows good performance in all categories, especially in the recognition of negative and positive emotions. Support vector machine performs best in recognizing positive emotions but has significant shortcomings in recalling negative emotions.

Next task:

In response to the shortcomings of the existing model, the next task should focus on optimizing the model and adjusting its parameters. For support vector machines, one can try adjusting the kernel function, penalty parameter C, or performing feature selection to improve the recall rate of negative emotions. At the same time, to further improve the generalization ability of the model, more feature engineering can be considered, such as using word embeddings to capture the semantic information of the text. In addition, ensemble learning methods such as random forest or gradient boosting can also be attempted to combine the advantages of multiple models and improve overall prediction performance. Finally, expanding and diversifying the training dataset is also a key step in improving model performance, especially by increasing the number of samples in categories where the model performs poorly. Through these methods, we can expect to achieve more accurate and robust sentiment analysis in future model iterations.