In [1]:
```
import numpy as np
import pandas as pd
```

In [2]:
```
import matplotlib.pyplot as plt
import seaborn as sns
import plotly.express as px
```

In [3]:
```
df=pd.read_csv('c:\\Users\\KAUSHIK\\Downloads\\social media sentiment analysis.csv\\sentimentdataset.csv')
```

In [4]:
```
df
```

Out[4]:

	Unnamed: 0.1	Unnamed: 0	Text	Sentiment	Timestamp	User	Platform	H
0	0	0	Enjoying a beautiful day at the park! ...	Positive	2023-01-15 12:30:00	User123	Twitter	#Natı
1	1	1	Traffic was terrible this morning. ...	Negative	2023-01-15 08:45:00	CommuterX	Twitter	#Traffic #
2	2	2	Just finished an amazing workout! 💪 ...	Positive	2023-01-15 15:45:00	FitnessFan	Instagram	#Fitness #
3	3	3	Excited about the upcoming weekend getaway! ...	Positive	2023-01-15 18:20:00	AdventureX	Facebook	#Travel #A
4	4	4	Trying out a new recipe for dinner tonight. ...	Neutral	2023-01-15 19:55:00	ChefCook	Instagram	#Cookir
...	...	...	...	...	...	...	...	
727	728	732	Collaborating on a science project that receiv...	Happy	2017-08-18 18:20:00	ScienceProjectSuccessHighSchool	Facebook	#ScienceFa #HighSchoo
728	729	733	Attending a surprise birthday party organized ...	Happy	2018-06-22 14:15:00	BirthdayPartyJoyHighSchool	Instagram	#SurpriseCe #HighSchoolF
729	730	734	Successfully fundraising	Happy	2019-04-05 17:30:00	CharityFundraisingTriumphHighSchool	Twitter	#Commur #HighSchoolPhil

	Unnamed: 0.1	Unnamed: 0	Text	Sentiment	Timestamp	User	Platform	H
			for a school charity ...					
730	731	735	Participating in a multicultural festival, cel...	Happy	2020-02-29 20:45:00	MulticulturalFestivalJoyHighSchool	Facebook	#CulturalCe #HighSch
731	732	736	Organizing a virtual talent show during challe...	Happy	2020-11-15 15:15:00	VirtualTalentShowSuccessHighSchool	Instagram	#VirtualEnter #HighSchool

732 rows × 15 columns

In [5]:
```
df.shape
```

Out[5]:
```
(732, 15)
```

In [6]:
```
df
```

Out[6]:

	Unnamed: 0.1	Unnamed: 0	Text	Sentiment	Timestamp	User	Platform	H
0	0	0	Enjoying a beautiful day at the park! ...	Positive	2023-01-15 12:30:00	User123	Twitter	#Natı
1	1	1	Traffic was terrible this morning. ...	Negative	2023-01-15 08:45:00	CommuterX	Twitter	#Traffic #
2	2	2	Just finished an amazing workout! 💪 ...	Positive	2023-01-15 15:45:00	FitnessFan	Instagram	#Fitness #
3	3	3	Excited about the upcoming weekend getaway! ...	Positive	2023-01-15 18:20:00	AdventureX	Facebook	#Travel #A
4	4	4	Trying out a new recipe for dinner tonight. ...	Neutral	2023-01-15 19:55:00	ChefCook	Instagram	#Cookir
...	...	...	...	...	...	...	...	
727	728	732	Collaborating on a science project that receiv...	Happy	2017-08-18 18:20:00	ScienceProjectSuccessHighSchool	Facebook	#ScienceFa #HighSchoo
728	729	733	Attending a surprise birthday party organized ...	Happy	2018-06-22 14:15:00	BirthdayPartyJoyHighSchool	Instagram	#SurpriseCe #HighSchoolF
729	730	734	Successfully fundraising	Happy	2019-04-05 17:30:00	CharityFundraisingTriumphHighSchool	Twitter	#Commur #HighSchoolPhil

	Unnamed: 0.1	Unnamed: 0	Text	Sentiment	Timestamp	User	Platform	H
			for a school charity ...					
730	731	735	Participating in a multicultural festival, cel...	Happy	2020-02-29 20:45:00	MulticulturalFestivalJoyHighSchool	Facebook	#CulturalCe #HighSch
731	732	736	Organizing a virtual talent show during challe...	Happy	2020-11-15 15:15:00	VirtualTalentShowSuccessHighSchool	Instagram	#VirtualEnter #HighSchool

732 rows × 15 columns

In [7]:
```
df.drop(df.columns[[0,1]],axis=1)
```

Out[7]:

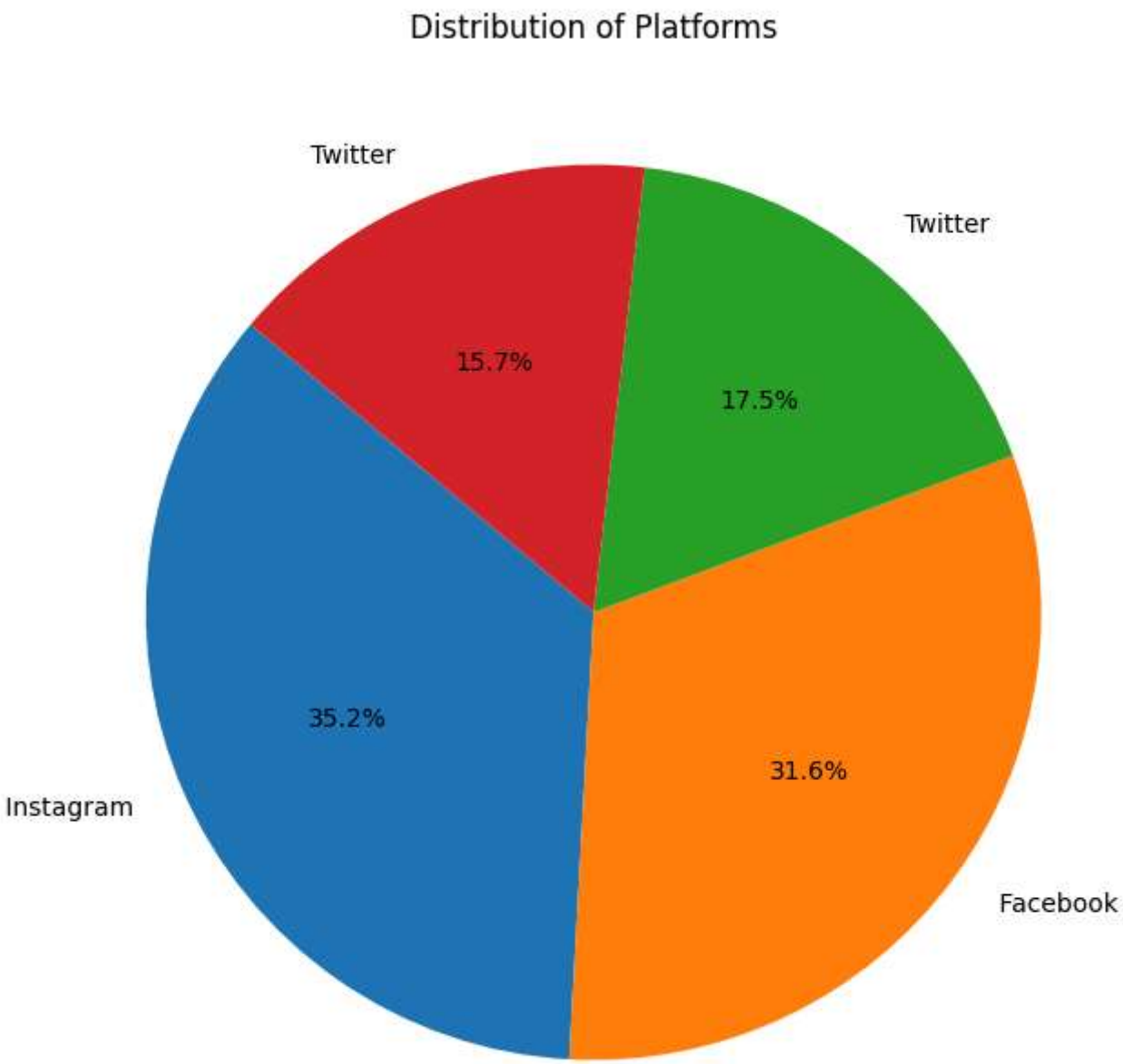
	Text	Sentiment	Timestamp	User	Platform	Hashtags	Retweets	Lik
0	Enjoying a beautiful day at the park! ...	Positive	2023-01-15 12:30:00	User123	Twitter	#Nature #Park	15.0	3(
1	Traffic was terrible this morning. ...	Negative	2023-01-15 08:45:00	CommuterX	Twitter	#Traffic #Morning	5.0	1(
2	Just finished an amazing workout! 💪 ...	Positive	2023-01-15 15:45:00	FitnessFan	Instagram	#Fitness #Workout	20.0	4(
3	Excited about the upcoming weekend getaway! ...	Positive	2023-01-15 18:20:00	AdventureX	Facebook	#Travel #Adventure	8.0	15
4	Trying out a new recipe for dinner tonight. ...	Neutral	2023-01-15 19:55:00	ChefCook	Instagram	#Cooking #Food	12.0	25
...	...	...	...	...	...	...	...	
727	Collaborating on a science project that receiv...	Happy	2017-08-18 18:20:00	ScienceProjectSuccessHighSchool	Facebook	#ScienceFairWinner #HighSchoolScience	20.0	39
728	Attending a surprise birthday party organized ...	Happy	2018-06-22 14:15:00	BirthdayPartyJoyHighSchool	Instagram	#SurpriseCelebration #HighSchoolFriendship	25.0	48

	Text	Sentiment	Timestamp	User	Platform	Hashtags	Retweets	Lik
729	Successfully fundraising for a school charity ...	Happy	2019-04-05 17:30:00	CharityFundraisingTriumphHighSchool	Twitter	#CommunityGiving #HighSchoolPhilanthropy	22.0	42
730	Participating in a multicultural festival, cel...	Happy	2020-02-29 20:45:00	MulticulturalFestivalJoyHighSchool	Facebook	#CulturalCelebration #HighSchoolUnity	21.0	43
731	Organizing a virtual talent show during challe...	Happy	2020-11-15 15:15:00	VirtualTalentShowSuccessHighSchool	Instagram	#VirtualEntertainment #HighSchoolPositivity	24.0	47

732 rows × 13 columns

In [8]:
```
# Calculate the size of each category
category_sizes = df['Platform'].value_counts()

# Create a pie chart
plt.figure(figsize=(8, 8))
plt.pie(category_sizes, labels=category_sizes.index, autopct='%1.1f%%', startangle=140)
plt.title('Distribution of Platforms')
plt.show()
```



In [9]:
```
from sklearn.preprocessing import LabelEncoder, StandardScaler
```

```
# Drop rows with missing values
df = df.dropna()

# Encode categorical variables
label_encoder = LabelEncoder()
df['Sentiment'] = label_encoder.fit_transform(df['Sentiment'])

# Normalize numerical features
scaler = StandardScaler()
numerical_columns = df.select_dtypes(include=['float64', 'int64']).columns
df[numerical_columns] = scaler.fit_transform(df[numerical_columns])

print(df.head())
```

```
   Unnamed: 0.1  Unnamed: 0  \
0     -1.733763   -1.741727
1     -1.729032   -1.737017
2     -1.724301   -1.732306
3     -1.719570   -1.727595
4     -1.714839   -1.722884

                                          Text  Sentiment  \
0   Enjoying a beautiful day at the park!        ...   1.025909
1   Traffic was terrible this morning.           ...   0.768055
2   Just finished an amazing workout! 💪           ...   1.025909
3   Excited about the upcoming weekend getaway!  ...   1.025909
4   Trying out a new recipe for dinner tonight.  ...   0.795197

             Timestamp            User     Platform  \
0  2023-01-15 12:30:00   User123          Twitter
1  2023-01-15 08:45:00   CommuterX        Twitter
2  2023-01-15 15:45:00   FitnessFan      Instagram
3  2023-01-15 18:20:00   AdventureX       Facebook
4  2023-01-15 19:55:00   ChefCook        Instagram

                                Hashtags  Retweets     Likes  \
0   #Nature #Park                          -0.922303 -0.916295
1   #Traffic #Morning                      -2.339444 -2.336727
2   #Fitness #Workout                      -0.213733 -0.206079
3   #Travel #Adventure                     -1.914302 -1.981619
4   #Cooking #Food                         -1.347445 -1.271403

        Country      Year     Month       Day      Hour
0     USA        0.902984 -1.502582 -0.058718 -0.856774
1     Canada     0.902984 -1.502582 -0.058718 -1.829867
2   USA          0.902984 -1.502582 -0.058718 -0.126954
3     UK         0.902984 -1.502582 -0.058718  0.602866
4    Australia   0.902984 -1.502582 -0.058718  0.846139
```

In [10]:

```python
import re

# Function to preprocess text
def preprocess_text(text):
    # Convert text to lowercase
    text = text.lower()
    # Remove URLs
```

```
    text = re.sub(r'http\S+|www\S+|https\S+', '', text, flags=re.MULTILINE)
    # Remove special characters and numbers
    text = re.sub(r'\W+|\d+', ' ', text)
    # Remove extra spaces
    text = re.sub(r'\s+', ' ', text).strip()
    return text

# Apply preprocessing to the 'Text' column
df['Cleaned_Text'] = df['Text'].apply(preprocess_text)

# Display the first few rows of the updated dataframe
print(df[['Text', 'Cleaned_Text']].head())
```

```
                                        Text  \
0   Enjoying a beautiful day at the park!        ...
1   Traffic was terrible this morning.           ...
2   Just finished an amazing workout! 💪           ...
3   Excited about the upcoming weekend getaway!  ...
4   Trying out a new recipe for dinner tonight.  ...

                               Cleaned_Text
0       enjoying a beautiful day at the park
1          traffic was terrible this morning
2           just finished an amazing workout
3  excited about the upcoming weekend getaway
4  trying out a new recipe for dinner tonight
```

In [11]:
```
pip install xgboost
```

```
Requirement already satisfied: xgboost in c:\users\kaushik\appdata\local\programs\python\python312\lib\site-packages
(3.0.0)
Requirement already satisfied: numpy in c:\users\kaushik\appdata\local\programs\python\python312\lib\site-packages (f
rom xgboost) (2.1.3)
Requirement already satisfied: scipy in c:\users\kaushik\appdata\local\programs\python\python312\lib\site-packages (f
rom xgboost) (1.15.2)
Note: you may need to restart the kernel to use updated packages.
```

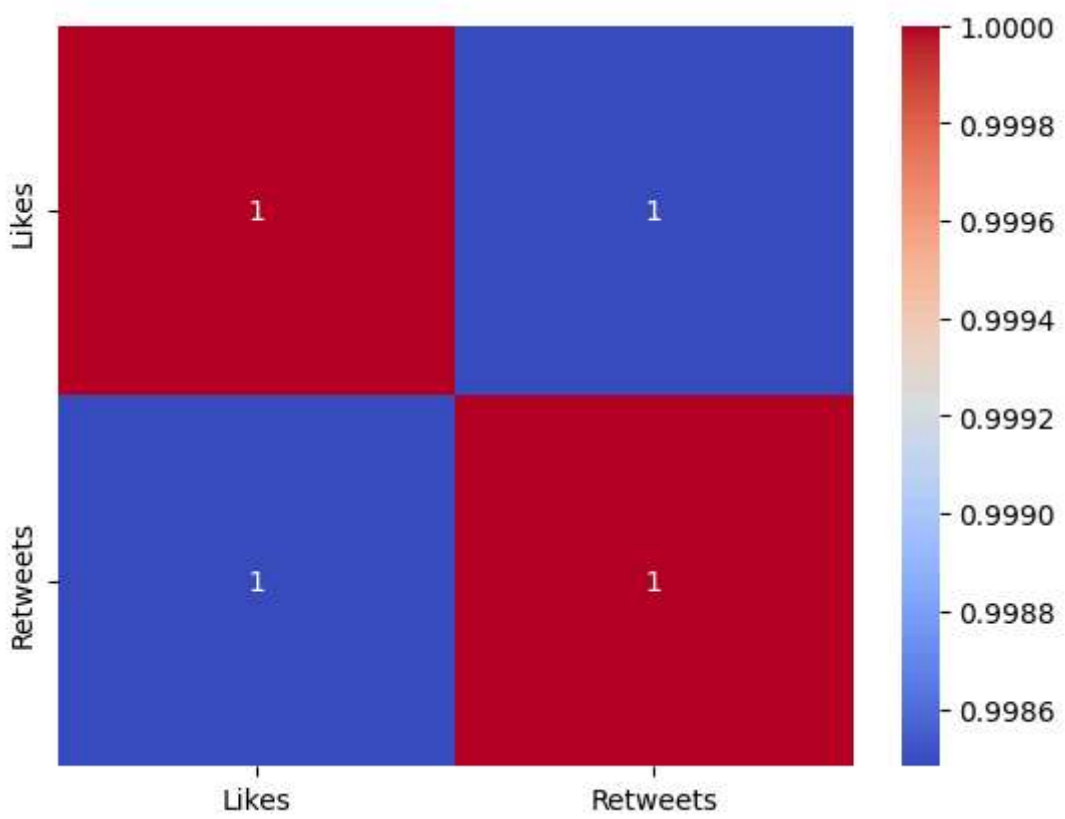
In [12]:
```
px.scatter(df,x='Retweets',y='Likes',title='Retweets vs Likes')
```

In [13]:
```
df['Retweets'].describe()
```

Out[13]:
```
count    7.320000e+02
mean     1.553099e-16
std      1.000684e+00
min     -2.339444e+00
25%     -5.325894e-01
50%      6.969545e-02
75%      4.948377e-01
max      2.620549e+00
Name: Retweets, dtype: float64
```

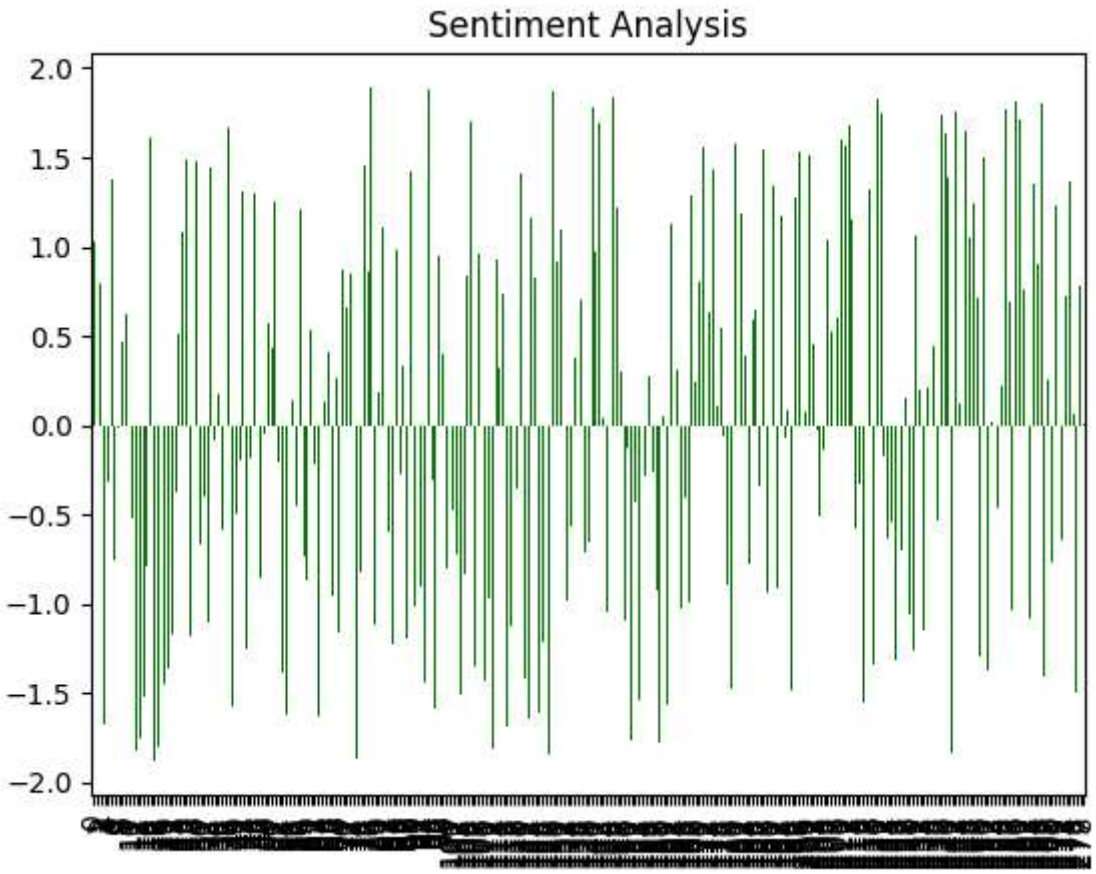
In [14]:
```
# Display the correlation matrix
sns.heatmap(df[['Likes','Retweets']].corr(),annot=True, cmap='coolwarm')
```

Out[14]: <Axes: >



In [15]:
```
label=pd.Series(df['Sentiment'].unique())
label.plot(kind='bar',color='green',title='Sentiment Analysis')
```

Out[15]: <Axes: title={'center': 'Sentiment Analysis'}>



In [16]:
```
numeric_df = df.select_dtypes(include=['float64', 'int64'])
print(numeric_df.head())
```

```
   Unnamed: 0.1  Unnamed: 0  Sentiment  Retweets     Likes      Year  \
0     -1.733763   -1.741727   1.025909 -0.922303 -0.916295  0.902984
1     -1.729032   -1.737017   0.768055 -2.339444 -2.336727  0.902984
2     -1.724301   -1.732306   1.025909 -0.213733 -0.206079  0.902984
3     -1.719570   -1.727595   1.025909 -1.914302 -1.981619  0.902984
4     -1.714839   -1.722884   0.795197 -1.347445 -1.271403  0.902984

      Month       Day      Hour
0 -1.502582 -0.058718 -0.856774
1 -1.502582 -0.058718 -1.829867
2 -1.502582 -0.058718 -0.126954
3 -1.502582 -0.058718  0.602866
4 -1.502582 -0.058718  0.846139
```

In [17]:
```
numeric_df.describe()
```

Out[17]:

	Unnamed: 0.1	Unnamed: 0	Sentiment	Retweets	Likes	Year	Month	Day
count	7.320000e+02	732.000000	7.320000e+02	7.320000e+02	7.320000e+02	7.320000e+02	7.320000e+02	7.320000e+02
mean	-7.765494e-17	0.000000	-8.250838e-17	1.553099e-16	1.941374e-16	-3.990493e-14	-1.941374e-17	-3.761411e-17
std	1.000684e+00	1.000684	1.000684e+00	1.000684e+00	1.000684e+00	1.000684e+00	1.000684e+00	1.000684e+00
min	-1.733763e+00	-1.741727	-1.878339e+00	-2.339444e+00	-2.336727e+00	-3.739260e+00	-1.502582e+00	-1.711852e+00
25%	-8.644320e-01	-0.866719	-8.197813e-01	-5.325894e-01	-5.789426e-01	-5.253989e-01	-9.159739e-01	-7.672039e-01
50%	1.680429e-04	0.003578	8.064882e-03	6.969545e-02	6.985732e-03	1.887924e-01	-3.606196e-02	-5.871788e-02
75%	8.647681e-01	0.864454	8.494824e-01	4.948377e-01	5.041370e-01	9.029837e-01	8.438500e-01	7.678492e-01
max	1.729368e+00	1.725330	1.894469e+00	2.620549e+00	2.634785e+00	9.029837e-01	1.723762e+00	1.830578e+00

In [18]:
```
numeric_df.drop(columns=['Unnamed: 0.1', 'Unnamed: 0'], inplace=True)
```

In [19]:
```
numeric_df
```

Out[19]:

	Sentiment	Retweets	Likes	Year	Month	Day	Hour
0	1.025909	-0.922303	-0.916295	0.902984	-1.502582	-0.058718	-0.856774
1	0.768055	-2.339444	-2.336727	0.902984	-1.502582	-0.058718	-1.829867
2	1.025909	-0.213733	-0.206079	0.902984	-1.502582	-0.058718	-0.126954
3	1.025909	-1.914302	-1.981619	0.902984	-1.502582	-0.058718	0.602866
4	0.795197	-1.347445	-1.271403	0.902984	-1.502582	-0.058718	0.846139
...	...	...	...	...	...	...	...
727	0.008065	-0.213733	-0.277101	-1.239590	0.550546	0.295525	0.602866
728	0.008065	0.494838	0.362094	-0.882495	-0.036062	0.767849	-0.370227
729	0.008065	0.069695	-0.064036	-0.525399	-0.622670	-1.239528	0.359592
730	0.008065	-0.072019	0.006986	-0.168303	-1.209278	1.594416	1.089412
731	0.008065	0.353124	0.291072	-0.168303	1.430458	-0.058718	-0.126954

732 rows × 7 columns

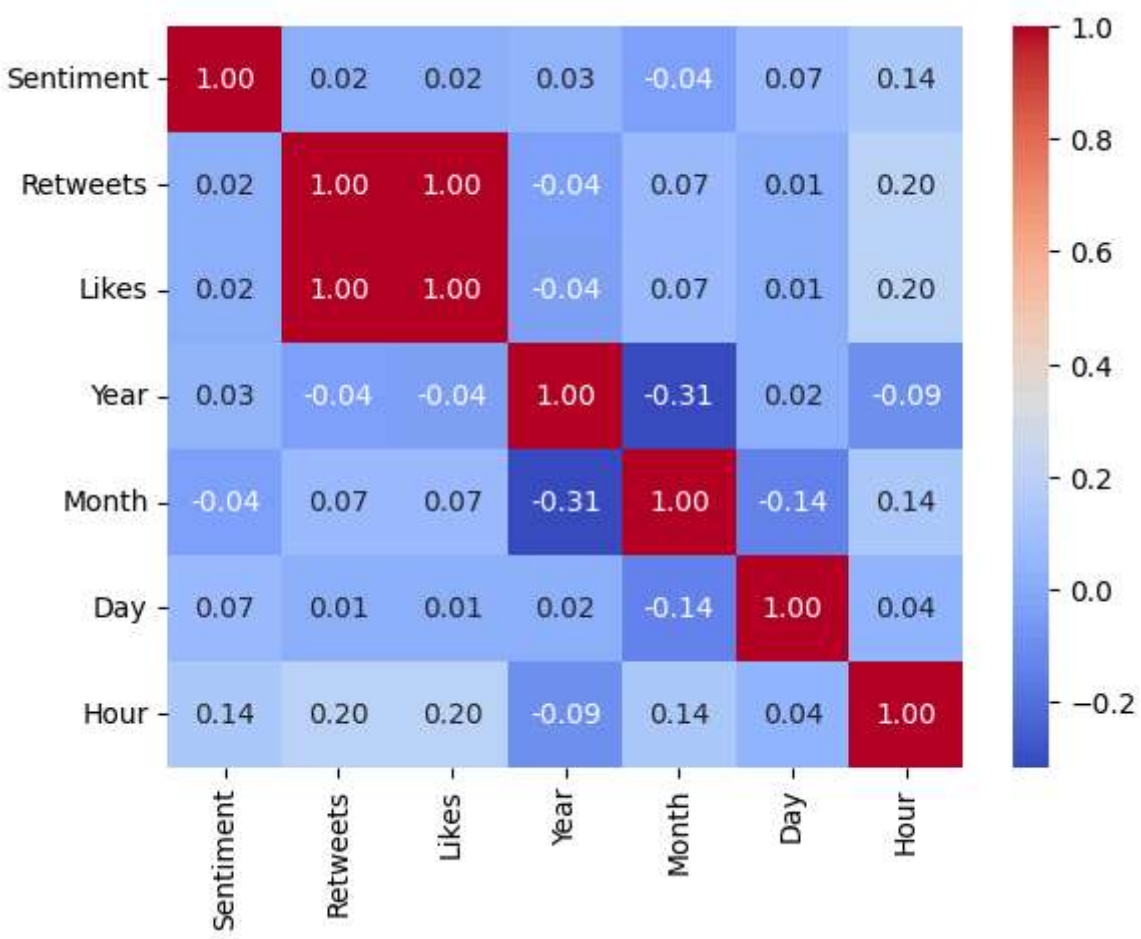
In [20]:
```
numeric_df.corr()
```

Out[20]:

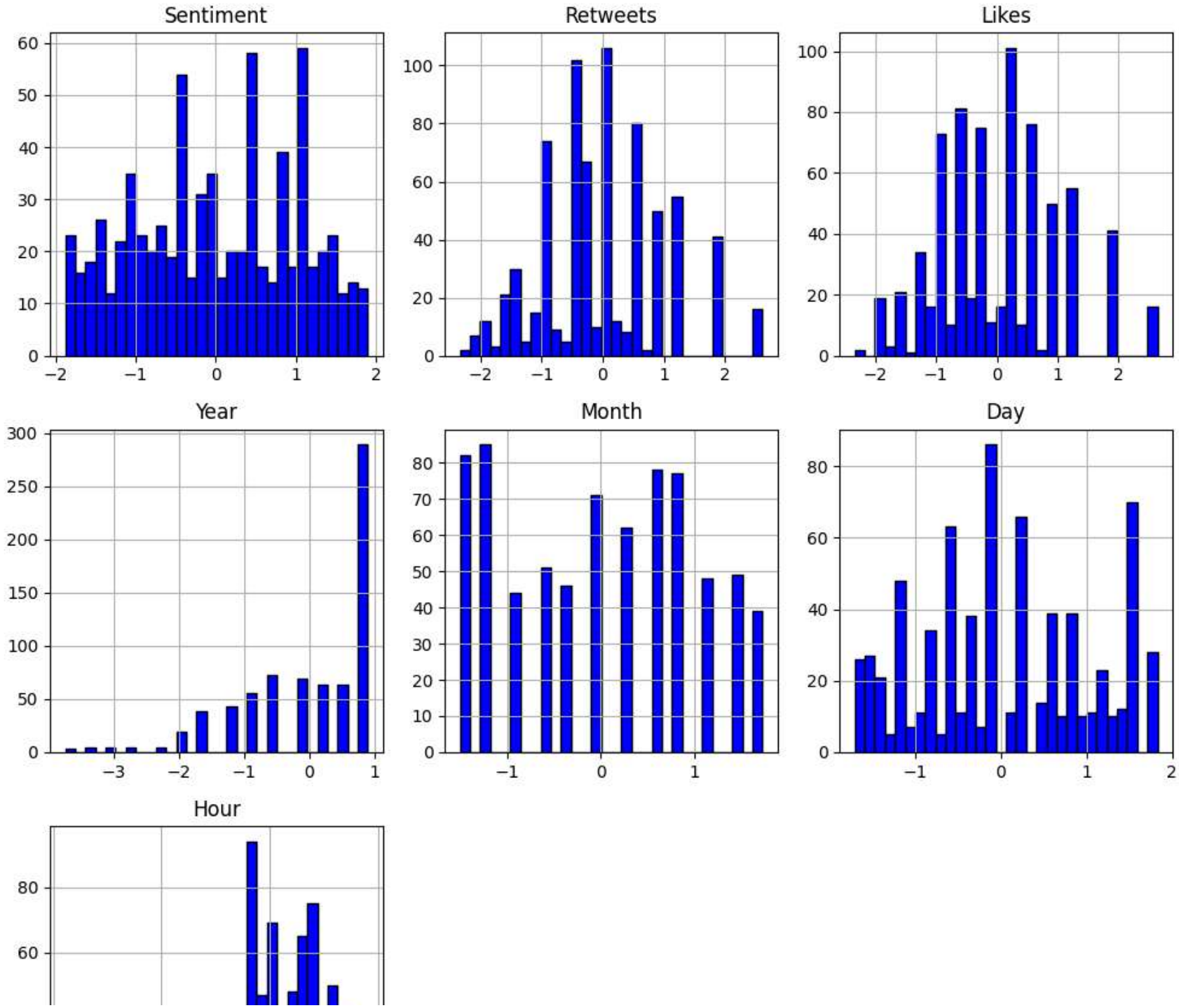
	Sentiment	Retweets	Likes	Year	Month	Day	Hour
Sentiment	1.000000	0.021697	0.021539	0.029799	-0.042200	0.066255	0.140351
Retweets	0.021697	1.000000	0.998482	-0.039982	0.073265	0.009213	0.196955
Likes	0.021539	0.998482	1.000000	-0.043415	0.066643	0.011489	0.195331
Year	0.029799	-0.039982	-0.043415	1.000000	-0.314845	0.021973	-0.087470
Month	-0.042200	0.073265	0.066643	-0.314845	1.000000	-0.135873	0.137835
Day	0.066255	0.009213	0.011489	0.021973	-0.135873	1.000000	0.044072
Hour	0.140351	0.196955	0.195331	-0.087470	0.137835	0.044072	1.000000

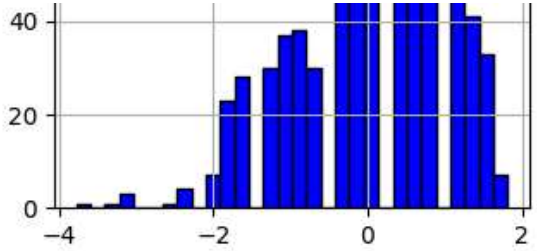
In [21]:
```
sns.heatmap(numeric_df.corr(), annot=True, cmap='coolwarm', fmt='.2f')
```

Out[21]: <Axes: >



In [22]:
```
numeric_df.hist(figsize=(10, 10), bins=30, layout=(3, 3), color='blue', edgecolor='black')
plt.tight_layout()
plt.show()
```


Sentiment
Retweets
Likes
Year
Month
Day
Hour



In [ ]:
```
#using textblob for sentiment analysis
from textblob import TextBlob
blob = TextBlob(' '.join(df['Text']))
print(blob.sentiment)
```

```
Sentiment(polarity=0.14225630225204094, subjectivity=0.5174179053269956)
```

In [32]:
```
polarity = blob.sentiment.polarity
subjectivity = blob.sentiment.subjectivity
if polarity > 0:
    print("The sentiment is positive.")
elif polarity < 0:
    print("The sentiment is negative.")
else:
    print("The sentiment is neutral.")
# Display the polarity and subjectivity
print(f"Polarity: {polarity}, Subjectivity: {subjectivity}")
```

```
The sentiment is positive.
Polarity: 0.14225630225204094, Subjectivity: 0.5174179053269956
```