In [39]:

```python
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

from sklearn.preprocessing import StandardScaler
from sklearn.feature_extraction.text import TfidfVectorizer, CountVectorizer
from sklearn.model_selection import train_test_split, GridSearchCV, cross_val_score
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import classification_report, confusion_matrix, accuracy_score
from sklearn.pipeline import Pipeline

from xgboost import XGBClassifier

import nltk
from nltk.tokenize import word_tokenize
from nltk.stem import WordNetLemmatizer
from nltk.corpus import stopwords
from nltk.sentiment.vader import SentimentIntensityAnalyzer

from textblob import TextBlob
```

```
c:\Users\Fagan\anaconda3\envs\learn-env\lib\site-packages\xgboost\compat.
py:93: FutureWarning: pandas.Int64Index is deprecated and will be removed
from pandas in a future version. Use pandas.Index with the appropriate dt
ype instead.
  from pandas import MultiIndex, Int64Index
```

In [2]:

```python
df = pd.read_csv('bfro_reports_geocoded.csv')
df.head()
```

Out[2]:

	observed	location_details	county	state	season	title	latitude	longitude
0	I am not sure how relevant this report will be...	We were on our way to Rapid City, so we were h...	Washakie County	Wyoming	Summer	NaN	NaN	NaN
1	I don't know if what I saw was two bigfoots or...	Heading to the deep mine Poca #2, the airshaft...	Wyoming County	West Virginia	Winter	Report 13237: Daylight sighting near an abando...	37.58135	-81.29745
2	My family and I went to Ludlow, Vermont for Co...	It's off Rt 100 outside of Ludlow Vermont. It ...	Windsor County	Vermont	Fall	Report 13285: Evening sighting by motorists on...	43.46540	-72.70510
3	It was spring break 1984 and I was 16 at the t...	Wythe county Virginia near Wytheville, looking...	Wythe County	Virginia	Spring	Report 2285: Boy sees "Bigfoot" in the woods w...	37.22647	-81.09017
4	It was the winter of 1996 and we were on our w...	Hwy 182, Wood County Between Quitman, Texas an...	Wood County	Texas	Winter	Report 2048: Night time road crossing observation	32.79430	-95.54250

5 rows × 29 columns

In [3]:

```python
df.shape
```

Out[3]:

```
(5082, 29)
```

In [4]:

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5082 entries, 0 to 5081
Data columns (total 29 columns):
 #   Column              Non-Null Count  Dtype  
---  ------              --------------  -----  
 0   observed            5043 non-null   object 
 1   location_details    4319 non-null   object 
 2   county              5082 non-null   object 
 3   state               5082 non-null   object 
 4   season              5082 non-null   object 
 5   title               4104 non-null   object 
 6   latitude            4104 non-null   float64
 7   longitude           4104 non-null   float64
 8   date                4104 non-null   object 
 9   number              5082 non-null   float64
 10  classification      5082 non-null   object 
 11  geohash             4104 non-null   object 
 12  temperature_high    4102 non-null   float64
 13  temperature_mid     3964 non-null   float64
 14  temperature_low     4102 non-null   float64
 15  dew_point           3951 non-null   float64
 16  humidity            3951 non-null   float64
 17  cloud_cover         3939 non-null   float64
 18  moon_phase          4104 non-null   float64
 19  precip_intensity    3524 non-null   float64
 20  precip_probability  3964 non-null   float64
 21  precip_type         1309 non-null   object 
 22  pressure            3678 non-null   float64
 23  summary             3964 non-null   object 
 24  conditions          3964 non-null   object 
 25  uv_index            394 non-null    float64
 26  visibility          3916 non-null   float64
 27  wind_bearing        3955 non-null   float64
 28  wind_speed          3966 non-null   float64
dtypes: float64(17), object(12)
memory usage: 1.1+ MB
```

In [5]:

```
df.columns
```

Out[5]:

```
Index(['observed', 'location_details', 'county', 'state', 'season', 'titl
e',
       'latitude', 'longitude', 'date', 'number', 'classification', 'geoh
ash',
       'temperature_high', 'temperature_mid', 'temperature_low', 'dew_poi
nt',
       'humidity', 'cloud_cover', 'moon_phase', 'precip_intensity',
       'precip_probability', 'precip_type', 'pressure', 'summary',
       'conditions', 'uv_index', 'visibility', 'wind_bearing', 'wind_spee
d'],
      dtype='object')
```

In [6]:

```python
print(df['observed'].isna().sum())
print(df['classification'].isna().sum())
```

```
39
0
```

In [7]:

```python
df['classification'].value_counts()
```

Out[7]:

```
Class B    2550
Class A    2502
Class C      30
Name: classification, dtype: int64
```

In [8]:

```python
df['observed'] = df['observed'].astype(str)
```

In [9]:

```python
sia = SentimentIntensityAnalyzer()

def get_sentiment_scores(text):
    sentiment_scores = sia.polarity_scores(text)
    return sentiment_scores['pos'], sentiment_scores['neu'], sentiment_scores['neg']
```

In [10]:

```
df['positive_score'], df['neutral_score'], df['negative_score'] = zip(*df['observed'].apply(get_sentiment_scores))
df.head()
```

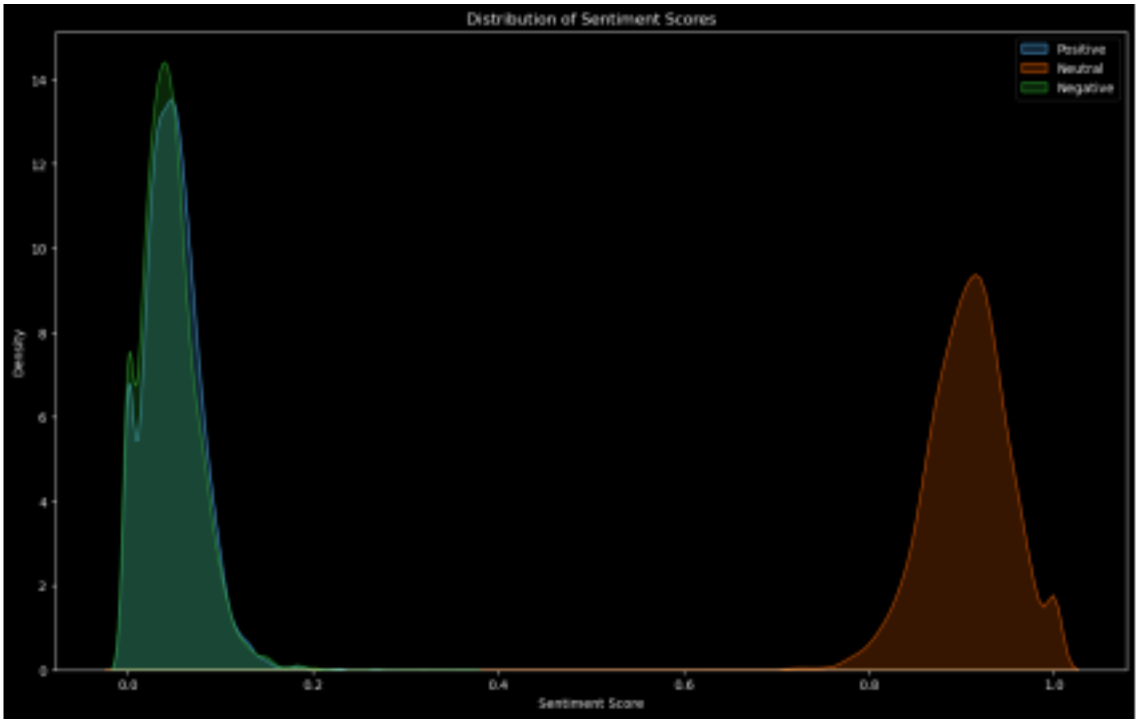
Out[10]:

	observed	location_details	county	state	season	title	latitude	longitude
0	I am not sure how relevant this report will be...	We were on our way to Rapid City, so we were h...	Washakie County	Wyoming	Summer	NaN	NaN	NaN
1	I don't know if what I saw was two bigfoots or...	Heading to the deep mine Poca #2, the airshaft...	Wyoming County	West Virginia	Winter	Report 13237: Daylight sighting near an abando...	37.58135	-81.29745
2	My family and I went to Ludlow, Vermont for Co...	It's off Rt 100 outside of Ludlow Vermont. It ...	Windsor County	Vermont	Fall	Report 13285: Evening sighting by motorists on...	43.46540	-72.70510
3	It was spring break 1984 and I was 16 at the t...	Wythe county Virginia near Wytheville, looking...	Wythe County	Virginia	Spring	Report 2285: Boy sees "Bigfoot" in the woods w...	37.22647	-81.09017
4	It was the winter of 1996 and we were on our w...	Hwy 182, Wood County Between Quitman, Texas an...	Wood County	Texas	Winter	Report 2048: Night time road crossing observation	32.79430	-95.54250

5 rows × 32 columns

In [11]:

```python
plt.figure(figsize=(15, 9))
sns.kdeplot(data=df, x='positive_score', label='Positive', shade=True)
sns.kdeplot(data=df, x='neutral_score', label='Neutral', shade=True)
sns.kdeplot(data=df, x='negative_score', label='Negative', shade=True)
plt.xlabel('Sentiment Score')
plt.ylabel('Density')
plt.title('Distribution of Sentiment Scores')
plt.legend()
plt.show()
```



In [12]:

```python
# Define a function to calculate the sentiment polarity and subjectivity using TextBlob
def get_sentiment(text):
    blob = TextBlob(text)
    sentiment_polarity = blob.sentiment.polarity
    sentiment_subjectivity = blob.sentiment.subjectivity
    return sentiment_polarity, sentiment_subjectivity
```

In [13]:

```python
# Apply the get_sentiment function to the 'observed' column of the dataframe and create new columns for the sentiment polarity and subjectivity
df[['sentiment_polarity', 'sentiment_subjectivity']] = df['observed'].apply(lambda x: pd.Series(get_sentiment(x)))
df
```

Out[13]:

	observed	location_details	county	state	season	title	latitude	longitu
0	I am not sure how relevant this report will be...	We were on our way to Rapid City, so we were h...	Washakie County	Wyoming	Summer	NaN	NaN	N
1	I don't know if what I saw was two bigfoots or...	Heading to the deep mine Poca #2, the airshaft...	Wyoming County	West Virginia	Winter	Report 13237: Daylight sighting near an abando...	37.58135	-81.29
2	My family and I went to Ludlow, Vermont for Co...	It's off Rt 100 outside of Ludlow Vermont. It ...	Windsor County	Vermont	Fall	Report 13285: Evening sighting by motorists on...	43.46540	-72.70
3	It was spring break 1984 and I was 16 at the t...	Wythe county Virginia near Wytheville, looking...	Wythe County	Virginia	Spring	Report 2285: Boy sees "Bigfoot" in the woods w...	37.22647	-81.09
4	It was the winter of 1996 and we were on our w...	Hwy 182, Wood County Between Quitman, Texas an...	Wood County	Texas	Winter	Report 2048: Night time road crossing observation	32.79430	-95.54
...	...	...	...	...	...	...	...	
5077	while camping protecting other equipment befor...	(withheld)	Rio Arriba County	New Mexico	Summer	NaN	NaN	N
5078	I was on my way to work on a Saturday morning ...	Laurel, Maryland. It was sighted off of Rt 19...	Prince George's County	Maryland	Spring	NaN	NaN	N
5079	On the twenty sixth and again on the twenty se...	head n.on highway 441 from Orlando,then go eas...	Lake County	Florida	Summer	NaN	NaN	N

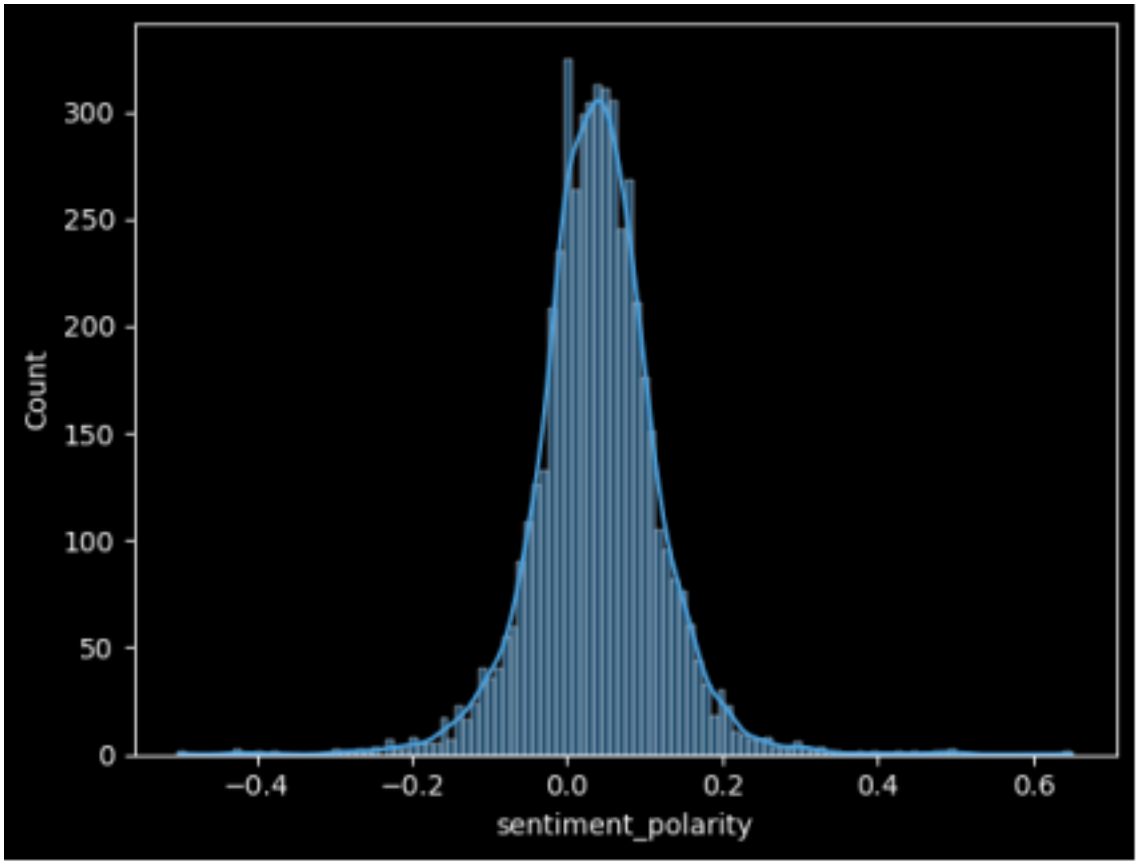
	observed	location_details	county	state	season	title	latitude	longitu
5080	I was hunting on me and my neighbor's property...	It was on my neighbor's property in the woods ...	White County	Illinois	Fall	NaN	NaN	N
5081	I was riding with a friend in the summer of 19	This happened on the Mississippi River Road in...	Calhoun County	Illinois	Summer	NaN	NaN	N

In [14]:

```
sns.histplot(df['sentiment_polarity'], kde=True)
```

Out[14]:

```
<AxesSubplot:xlabel='sentiment_polarity', ylabel='Count'>
```

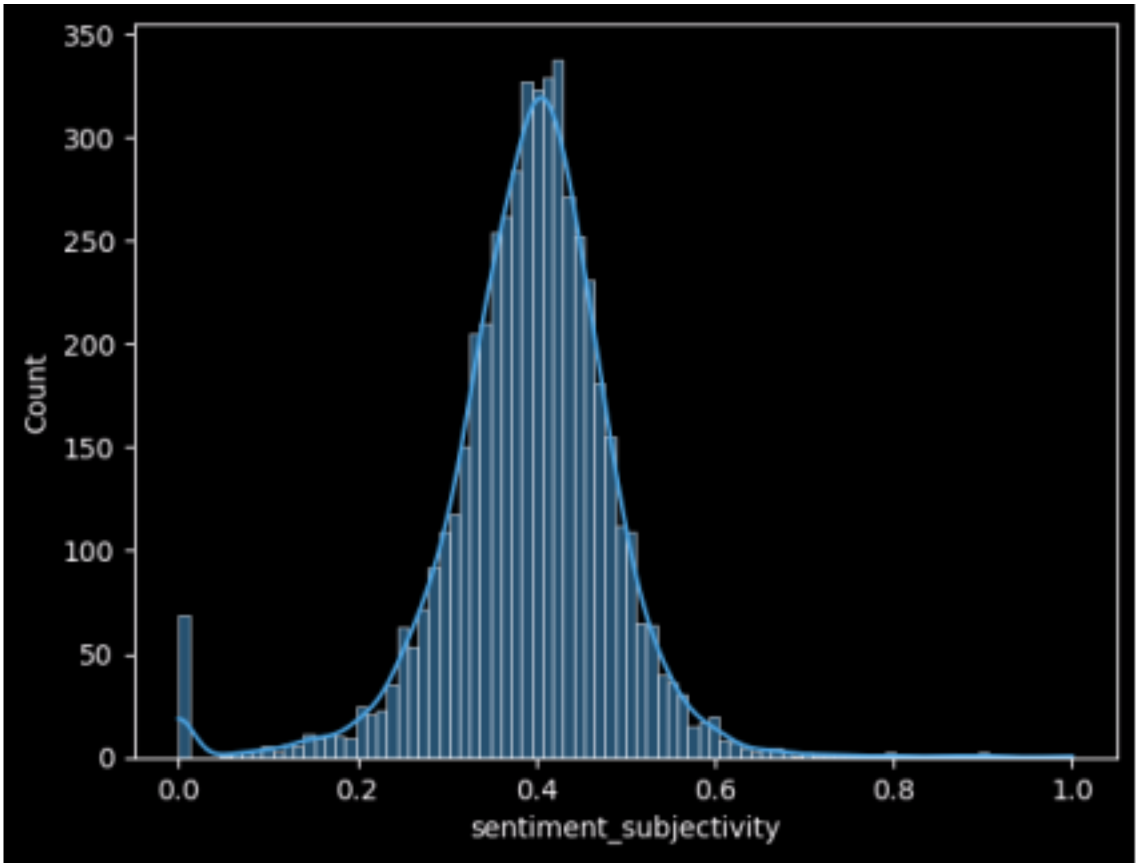


In [15]:

```python
sns.histplot(df['sentiment_subjectivity'], kde=True)
```

Out[15]:

```
<AxesSubplot:xlabel='sentiment_subjectivity', ylabel='Count'>
```



In [16]:

```python
# create preprocess_text function
def preprocess_text(text):

    # Tokenize the text
    tokens = word_tokenize(text.lower())

    # Remove stop words
    filtered_tokens = [token for token in tokens if token not in stopwords.words('english')]

    # Lemmatize the tokens
    lemmatizer = WordNetLemmatizer()
    lemmatized_tokens = [lemmatizer.lemmatize(token) for token in filtered_tokens]

    # Join the tokens back into a string
    processed_text = ' '.join(lemmatized_tokens)
    return processed_text
```

In [17]:

```python
df['observed'] = df['observed'].apply(preprocess_text)
df
```

Out[17]:

	observed	location_details	county	state	season	title	latitude	lon
0	sure relevant report , however thought importa...	We were on our way to Rapid City, so we were h...	Washakie County	Wyoming	Summer	NaN	NaN	
1	n't know saw two bigfoot something else , one ...	Heading to the deep mine Poca #2, the airshaft...	Wyoming County	West Virginia	Winter	Report 13237: Daylight sighting near an abando...	37.58135	-81
2	family went ludlow , vermont columbus day week...	It's off Rt 100 outside of Ludlow Vermont. It ...	Windsor County	Vermont	Fall	Report 13285: Evening sighting by motorists on...	43.46540	-72
3	spring break 1984 16 time . dad brother trip v...	Wythe county Virginia near Wytheville, looking...	Wythe County	Virginia	Spring	Report 2285: Boy sees "Bigfoot" in the woods w...	37.22647	-81
4	winter 1996 way home church one sunday evening...	Hwy 182, Wood County Between Quitman, Texas an...	Wood County	Texas	Winter	Report 2048: Night time road crossing observation	32.79430	-95
...	...	...	...	...	...	...	...	
5077	camping protecting equipment feat starting . e...	(withheld)	Rio Arriba County	New Mexico	Summer	NaN	NaN	
5078	way work saturday morning 7 a.m. misty , foggy...	Laurel, Maryland. It was sighted off of Rt 19...	Prince George's County	Maryland	Spring	NaN	NaN	
5079	twenty sixth twenty seventh approximately 5:00...	head n.on highway 441 from Orlando,then go eas...	Lake County	Florida	Summer	NaN	NaN	
5080	hunting neighbor 's property right daylight	It was on my neighbor's property in the woods ...	White County	Illinois	Fall	NaN	NaN	
5081	riding friend summer 1974 , back road night st...	This happened on the Mississippi River Road in...	Calhoun County	Illinois	Summer	NaN	NaN	

5082 rows × 34 columns

In [25]:

```
# Drop rows with NaN in 'observed' and 'classification'
df.dropna(subset=['observed', 'classification'], inplace=True)
# Drop rows with 'classification' equal to Class C
df = df[df['classification'] != 'Class C']
```

In [32]:

```python
# Create a pipeline
pipeline = Pipeline([
    ('vect', CountVectorizer()),
    ('clf', LogisticRegression(max_iter=2000))
])

# Define the hyperparameters to tune
parameters = {
    'vect__max_df': [0.6, 0.7, 0.8],
    'vect__min_df': [0.01, 0.02, 0.03],
    'clf__C': [0.01, 0.1, 1.0]
}

# Create a GridSearchCV object
grid_search = GridSearchCV(pipeline, parameters, cv=5, n_jobs=-1)

# Split the data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(df['observed'], df['classification'], test_size=0.2, random_state=42)

# Fit the GridSearchCV object to the training data
grid_search.fit(X_train, y_train)

# Print the best hyperparameters and the corresponding score
print('Best hyperparameters:', grid_search.best_params_)
print('Best score:', grid_search.best_score_)

# Evaluate the performance of the best estimator on the test set
y_pred = grid_search.best_estimator_.predict(X_test)
print('Accuracy:', accuracy_score(y_test, y_pred))
print('Classification report:\n', classification_report(y_test, y_pred))

```

```
Best hyperparameters: {'clf__C': 0.01, 'vect__max_df': 0.6, 'vect__min_df
': 0.01}
Best score: 0.795096623382981
Accuracy: 0.7655786350148368
Classification report:
               precision    recall  f1-score   support

     Class A       0.77      0.75      0.76       495
     Class B       0.77      0.78      0.77       516

    accuracy                           0.77      1011
   macro avg       0.77      0.77      0.77      1011
weighted avg       0.77      0.77      0.77      1011
```

In [33]:

```
pipeline
```

Out[33]:

```
Pipeline(steps=[('vect', CountVectorizer()),
                ('clf', LogisticRegression(max_iter=2000))])
```

In [34]:

```
# Changing from count vectorizer to TF-IDF vectorizer
pipeline.steps[0] = ('vect', TfidfVectorizer())
pipeline
```

Out[34]:

```
Pipeline(steps=[('vect', TfidfVectorizer()),
                ('clf', LogisticRegression(max_iter=2000))])
```

In [35]:

```
# Fit the GridSearchCV object to the training data
grid_search.fit(X_train, y_train)

# Print the best hyperparameters and the corresponding score
print('Best hyperparameters:', grid_search.best_params_)
print('Best score:', grid_search.best_score_)

# Evaluate the performance of the best estimator on the test set
y_pred = grid_search.best_estimator_.predict(X_test)
print('Accuracy:', accuracy_score(y_test, y_pred))
print('Classification report:\n', classification_report(y_test,
y_pred))
```

```
Best hyperparameters: {'clf__C': 1.0, 'vect__max_df': 0.7, 'vect__min_df
': 0.02}
Best score: 0.8054892973846209
Accuracy: 0.781404549950544
Classification report:
               precision    recall  f1-score   support

     Class A       0.78      0.77      0.78       495
     Class B       0.78      0.79      0.79       516

    accuracy                           0.78      1011
   macro avg       0.78      0.78      0.78      1011
weighted avg       0.78      0.78      0.78      1011
```

In [42]:

```python
# Changing from Logistic Regression to XGBoost
pipeline.steps[1] = ('clf', XGBClassifier())

# Define the hyperparameters to tune
parameters = {
    'vect__max_df': [0.6, 0.7, 0.8],
    'vect__min_df': [0.01, 0.02, 0.03],
    'clf__learning_rate': [0.01, 0.1, 1.0],
    'clf__max_depth': [3, 5, 7],
    'clf__n_estimators': [50, 100, 200]
}

# Create a GridSearchCV object
grid_search = GridSearchCV(pipeline, parameters, cv=5, n_jobs=-1)

# Fit the GridSearchCV object to the training data
grid_search.fit(X_train, y_train)

# Print the best hyperparameters and the corresponding score
print('Best hyperparameters:', grid_search.best_params_)
print('Best score:', grid_search.best_score_)

# Evaluate the performance of the best estimator on the test set
y_pred = grid_search.best_estimator_.predict(X_test)
print('Accuracy:', accuracy_score(y_test, y_pred))
print('Classification report:\n', classification_report(y_test, y_pred))
```

```
Best hyperparameters: {'clf__learning_rate': 0.1, 'clf__max_depth': 7, 'c
lf__n_estimators': 200, 'vect__max_df': 0.6, 'vect__min_df': 0.01}
Best score: 0.8099432131099389
Accuracy: 0.7695351137487636
Classification report:
               precision    recall  f1-score   support

     Class A       0.76      0.77      0.77       495
     Class B       0.78      0.77      0.77       516

    accuracy                           0.77      1011
   macro avg       0.77      0.77      0.77      1011
weighted avg       0.77      0.77      0.77      1011
```

In [43]:

```python
confusion_matrix(y_test, y_pred, labels=['Class A', 'Class B'])
```

Out[43]:

```
array([[380, 115],
       [118, 398]], dtype=int64)
```

In [45]:

```python
# Create a new DataFrame with rows where date is not missing
tsdf = df.dropna(subset=['date']).copy()

# Convert the date column to datetime and set it as the index
tsdf['date'] = pd.to_datetime(tsdf['date'])
tsdf.set_index('date', inplace=True)

# Sort the DataFrame by the index
tsdf.sort_index(inplace=True)

tsdf
```

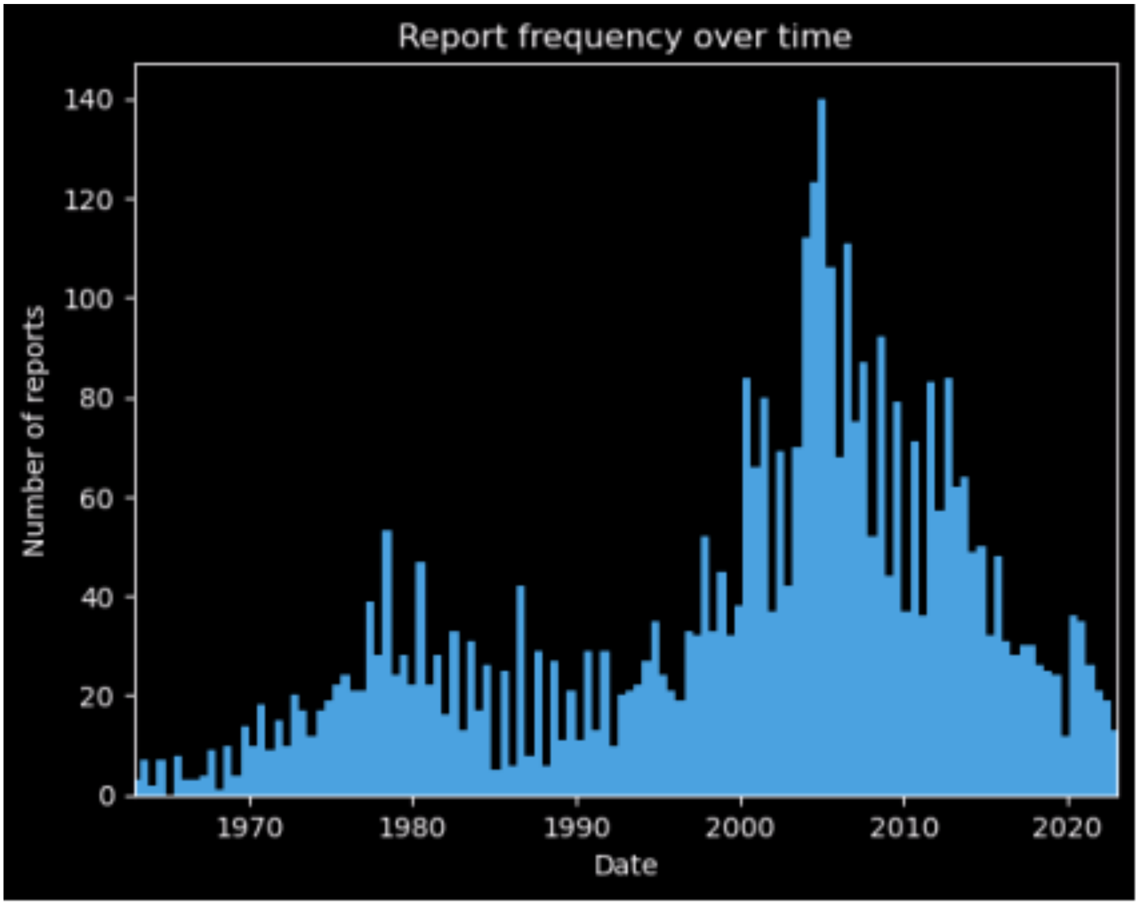
Out[45]:

	observed	location_details	county	state	season	title	latit
date							
1869-11-10	article , titled `` wild man - ? `` , original...	NaN	Stanislaus County	California	Fall	Report 14338: Old newspaper article (Titusvill...	37.39
1921-01-14	nan	NaN	Clearfield County	Pennsylvania	Winter	Report 14358: Old newspaper article (Clearfiel...	41.01
1925-10-14	today 's report bigfoot body , memory jarred s...	NaN	Avoyelles Parish	Louisiana	Fall	Report 24413: Woman recounts a tale her Grandf...	31.08
1930-09-30	second time listed report . two men lawyer hen...	Leon County Texas, 1930's along the Trinity ri...	Leon County	Texas	Fall	Report 2477: Nine foot tall brown/black creatu...	31.40
1937-08-16	nan	NaN	Warrick County	Indiana	Summer	Report 14336: Old newspaper article (The Hammo...	38.01
...	...	...	...	...	...	...	
2022-12-12	without doubt bigfoot chito nature preserve ' ...	Chito Nature Preserve - there are tens of thou...	Hillsborough County	Florida	Winter	Report 75271: Possible howls 20 miles SE of Ta...	27.82

	observed	location_details	county	state	season	title	latit
date							
2022-12-17	heard strange “ mournful howling ” (’ heard d...	Whitestone Dr, Canton GA [Investigator (MM) no...	Cherokee County	Georgia	Winter	Report 75283: RECENT !! Property owner reports...	34.17
2022-12-20	shortly dawn back porch coffee watching dog le...	Shady Drive, off of Warren Road	Indiana County	Pennsylvania	Winter	Report 75305: Possible trackway found and phot...	40.60
2023-02-09	husband heard strange sound 2 night february 2...	At the bottom of the woods next to the small c...	Cleburne County	Alabama	Winter	Report 75577: Daylight sighting, 2 witnesses, ...	33.65
	driving along					Report	

In [50]:

```
plt.hist(tsdf.index, bins=300)
plt.xlim('1963', '2023')
plt.xlabel('Date')
plt.ylabel('Number of reports')
plt.title('Report frequency over time')
plt.show()
```

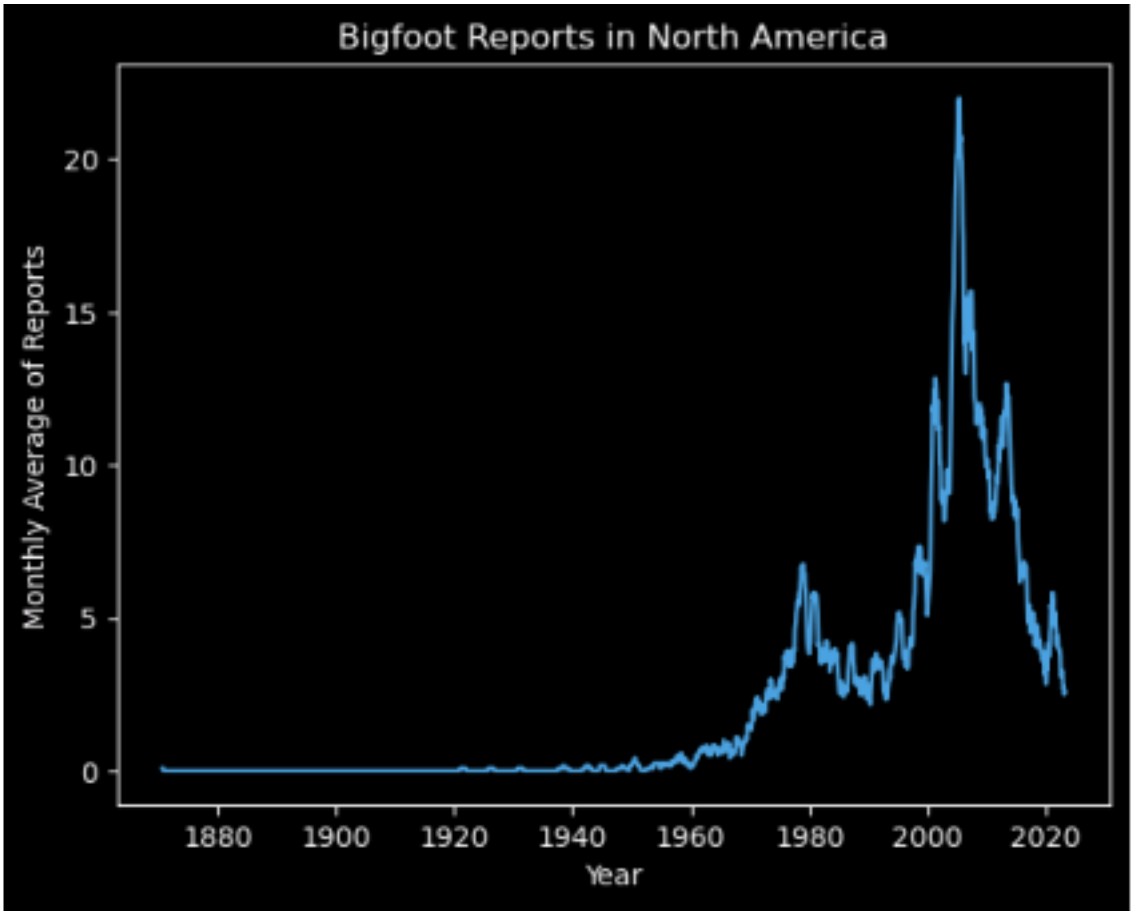


In [52]:

```python
# Count the number of reports per month
monthly_counts = tsdf.resample('M').count()['observed']

# Calculate the rolling monthly average with a window of 12 months
rolling_avg = monthly_counts.rolling(window=12).mean()

# Plot the rolling monthly average
plt.plot(rolling_avg)
plt.xlabel('Year')
plt.ylabel('Monthly Average of Reports')
plt.title('Bigfoot Reports in North America')
plt.show()
```



In [53]:

```python
# Extract the month from the index of tsdf and count the occurrences of each month
common_months = tsdf.index.month.value_counts()

# Print the most common month
print(f"The most common month is {common_months.index[0]}, with {common_months.iloc[0]} occurrences.")
```

```
The most common month is 10, with 550 occurrences.
```

In [ ]:

```
```