MACHINE LEARNING-BASED SENTIMENT ANALYSIS FOR TWEETS SAUDI TOURISM

SARAH M ALRASHIDI, FATMH N ALANAZI, HANAN A ALBALAWI, OHOOD M ALBALAWI, AWAD M AWADELKARIM

Faculty of Computers and Information Technology, University of Tabuk, Tabuk 47731, Saudi Arabia
E-mail: awad@ut.edu.sa

ABSTRACT

The growth of tourism in Saudi Arabia is a superior aspect of the recent economic success and realization of the Kingdom of Saudi Arabia's Vision 2030. This places a premium on research in such field and elevates it to national priority research, thus, this research project contributes to such context and places a premium on sentiment analysis in the tourism industry, namely concentrating on tweets about Saudi tourism. Therefore, this paper intends to demonstrate machine learning-based sentiment analysis models for tweets on Saudi tourism. The research studies and analyzes tweets related to tourism collected about six touristic places in Saudi Arabia from twenty accounts and nineteen touristic places from 144 hashtags. Following the preprocessing and feature extraction stage, such tweets are labeled as positive or negative using various machine learning algorithms. Two base classifier models of Support Vector Machine (SVM) and Naïve Bayes (NB) are applied. Over and above, a vital and important contribution of this project is creating the First Dataset for Tweets Saudi Tourism (FDTST) in both Arabic and English languages collected from Twitter libraries. Utterly, numerous classification models are developed and evaluated based on their performance computation, and the experimental results show that the developed models have achieved righteous and reliable upshots. Finally, the developed predictive models aid to appoint and specify several valuable recommendations and insights for continuous improvements and sustainable growth in the Saudi tourism industry.

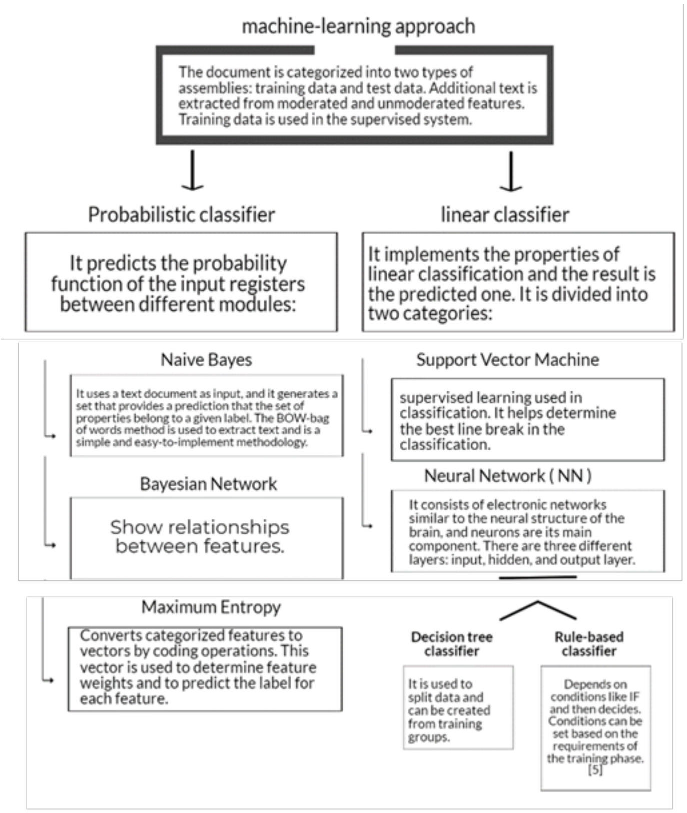
Keywords: *Sentiment Analysis, Machine Learning, Saudi Tourism, Tweets Saudi Tourism, Classification Models.*

1. INTRODUCTION

Sentiment analysis identifies positive, negative, or neutral feelings in a text by utilizing natural language processing, linguistic calculations, and text data analysis. This information can be gathered through blog posts, comments, reviews, and tweets. Polarity classification is a vital stage in assessing the emotional content of documents or sentences. The term "polarity evaluation" refers to determining whether a document or statement represents a positive, negative, or neutral point of view. Artificial intelligence and machine learning are two critical components that play a significant part in collecting information from the text that conveys the writer's emotion and establishes the writer's attitude, whether positive, negative, or neutral. Saudi Arabia has the potential to become a major tourist destination in the future. Apart from its historical and cultural values, Saudi Arabia has created growing support for tourism growth. Sentiment Analysis is a strong and wonderful field that focuses on massive text analysis methods rather than manual analysis; When it comes to sentiment analysis, several constraints apply; nevertheless, no comparable studies exist for Saudi Arabia's tourism sentiment analysis.

The approaches for sentiment analysis are grouped into three sections. Using machine learning, lexical-based, and mixed techniques, self-rating, sentiment identification, utility measure review, and spam detection. When implementing sentiment analysis, both machine learning models and lexical techniques can be used. Machine learning provides the highest level of accuracy, whereas semantic routing produces an acceptable overall outcome. Classification models require both training and testing data. The Decision Tree (DT), Support vector machine (SVM), Neural Network (NN), Naive Bayes (NB), and Maximum Entropy (ME) are some of the most frequently used classification algorithms for supervised learning.

Most classifications were performed using hyper red back-propagation neural network and semantic routing algorithms [1]. In a lexicon-based approach, the semantic orientation of the entire text is considered to be equal to the sum of the individual semantic orientations of words and phrases [2][3]. Based on the work in [4], the authors collected the key factors that connect the words in a document to its categories to develop a text classifier and model the text documents as collections of transactions. The data were collected from various positions and frequently require pre-processing before the full analysis can begin. The most common pre-processing procedures are feature filtering, feature extraction, and preparation.



The document is categorized into two types of assemblies: training data and test data. Additional text is extracted from moderated and unmoderated features. Training data is used in the supervised system.

It predicts the probability function of the input registers between different modules:

It implements the properties of linear classification and the result is the predicted one. It is divided into two categories:

It uses a text document as input, and it generates a set that provides a prediction that the set of properties belong to a given label. The BOW-bag of words method is used to extract text and is a simple and easy-to-implement methodology.

supervised learning used in classification. It helps determine the best line break in the classification.

Show relationships between features.

It consists of electronic networks similar to the neural structure of the brain, and neurons are its main component. There are three different layers: input, hidden, and output layer.

Converts categorized features to vectors by coding operations. This vector is used to determine feature weights and to predict the label for each feature.

It is used to split data and can be created from training groups.

Depends on conditions like IF and then decides. Conditions can be set based on the requirements of the training phase. [5]

Figure 1: Sections of Machine-learning Approaches and Algorithms.

The machine learning approach encompasses both supervised and unsupervised learning techniques in terms of methods and approaches for sentiment analysis. Support vector machine, Maximum Entropy, and Naive Bayes are all examples of supervised approaches. On the other hand, unsupervised techniques include exploiting sentiment lexicons, grammatical analysis, and synthetic patterns. Sections on machine learning techniques and algorithms are depicted in Figure 1. The accuracy and precision of data supervised learning methods are determined using the NB and SVM methodologies. In feelings, analytical texts may contain factual facts distinct from personal opinions, and it is critical to recognize the distinction. SVM is frequently used to classify texts [5][6]. Sentiment analysis has numerous applications in a variety of fields [7].

The creation of the first dataset for tweets related to Saudi tourism (FDTST) in both Arabic and English is a significant and essential contribution made to this paper. The dataset is compiled from Twitter library resources. As tourism is vital to a nation's progress and wealth, this activity can help the development of the Kingdom's 2030 vision. A proposed model uses textual data from Saudi Arabian tourists to identify and track tourist sentiment, thereby increasing tourist satisfaction by resolving any problems they encounter, resolving problems affecting the hospitality and tourism sectors due to social media, and measuring tourists' satisfaction with visited places and proposed services in Saudi Arabia. Numerous classification models are constructed and evaluated based on their performance computation, and the experimental findings demonstrate that the generated models have produced accurate and trustworthy outcomes.

2. RELATED WORK

This section summarizes the most significant and pertinent research. According to [8], much of the current research in sentiment categorization is directed toward the vast volume of rich online opinion sources such as discussion forums, review sites, social media posts and blogs, and easily available news organizations. The work hypothesized that human emotions could be classified into distinct primary emotions; as a result, people prefer to categorize information about halal tourism using the NRC Dictionary's eight emotion categories, and the researchers discovered that sentiment analysis is a branch of linguistics that detects the thoughts, sentiments, and moods expressed in a text. According to [9], the process of categorizing, recognizing, and quantifying views on anything is referred to as opinion mining (OM) or sentiment analysis (SA). The study's primary objective was to familiarize readers with sentiment analysis (SA), survey, opinion mining (OM) procedures, and the numerous approaches employed in this field. Additionally, it discusses application areas and challenges associated with sentiment analysis and provides insight into previous studies. According to [10], there will be numerous obstacles when dealing with emotion analysis. According to research, the creation of emotion categorization algorithms or opinion exploration remains an unexplored area of study. Naive Bayes and Support Vector Machines (SVM) are often used in supervised learning courses for opinion mining and emotion analysis. Machine Learning Methodologies and Lexicon-based Approaches are the two most often employed approaches in the categorization process. Machine Learning models are a subset of supervised learning, namely text classification. As a result, it is referred to as Supervised Machine Learning. It incorporates a variety of approaches, including SVM, NB, K-Nearest Neighbor (KNN), ME, and NN. According to [11], sentiment analysis enables the selection of "subjective data from enormous volumes of textual data through machine learning, data processing techniques, linguistic communication processes, data collection, and data management." Texts are classified as good, negative, or neutral in sentiment analysis. Additionally, it is defined by the range of values between 1 (obviously positive) and -1. (Clearly negative).

The work cited in [12] tries to develop an application capable of analyzing visitor feelings based on ratings and the selection of Indonesian attractions. TripAdvisor compiled the data from ten reviews of Indonesian tourism spots written in Indonesian. Manual categorization of the text into positive and negative emotions was performed to use as training data for the SVM. The raw text data was prepared and converted to a basic text format, and the noise was removed during the preprocessing stage. Additionally, features such as derivation, encoding, and stop word removal were retrieved. Additionally, (TF-IDF): The features were extracted using inverse Document Frequency term weighting and n-gram, unigram. The data was analyzed using the SVM algorithm, which classified it into two categories (positive and negative) and assigned positive emotions the symbol "1" and negative emotions the symbol "-1". The software is written in PHP and includes the Libsvm library, allowing SVM training and compilation. The supporting vector machine is fed data in the form of a matrix. The prediction of negative and pleasant emotions was 85 percent accurate. Positive emotion recall is 80%, while negative emotion recall is 100%. They must improve training data to increase accuracy. Additionally, it would be preferable if they classified the reviews into many categories based on the traveler's demands, such as hospitality, weather, etc.

The research is based on [13]. The objective is to collect data on international tourists' perceptions of Bangkok to enhance and promote tourism in the city. Tourism was grouped into five purposes: travel, business, relative visits, education, and health. The text was evaluated as hidden shapes, patterns, and trends in the source text using statistics and mathematics during the preprocessing step. Natural language processing (NLP) was utilized to eliminate redundant content while maintaining the document's key ideas. Then, the document and categories were categorized to streamline the further administration and sorting operation. Documents with similar contents are grouped to facilitate information extraction, retrieval, and filtering. They categorized emotions into positive and negative categories and assigned positive emotions a numerical value of "1" and bad emotions a "0." We concentrated on four machine learning techniques: ANN, which achieved the best accuracy of 80.32 percent, followed by SVM, which reached an accuracy of 80.11 percent. Then comes the decision tree with an accuracy of 83.79 percent, followed by Naive Bayes with a 55.66 percent accuracy.

The study cited in [14] intends to leverage machine learning techniques for data analysis to affect future improvements in the industry. ML is divided into three phases in tourism: The following destination can be forecasted before embarking on a journey. It employs algorithms such as neuro-fuzzy and discrete hidden Markov NN-based econometric models, neural networks, modeling by adding neural networks into Gray-Markov models, and NN-enhanced hidden Markov models. During the trip: The most critical tools at this stage are the recommendation system based on the user's mobile phone. Numerous machine learning algorithms are employed to improve the accuracy of the model. As a result, more precise advice is made. After the trip: Using machine learning techniques, it is also feasible to examine travelers' feelings and opinions about the areas they visited. This stage involves collecting data via internet networks and social networking sites such as Twitter, followed by the analysis of written content.

To the best of our knowledge, there is no dataset compiled from Arabic and English Twitter library resources connected to Saudi tourism in the literature. The majority of state-of-the-art methods are applied to online datasets. The authors are confident that this research can contribute to establishing the Kingdom's 2030 vision by recognizing and tracking tourist sentiment.

3. PROPLEM MOTIVATION AND GOALS

In this study, Twitter is used as a source of textual data. Tweet data provides businesses to gain a deeper understanding of tourists' activity patterns in Saudi Arabia, enabling them to alter their plans and better meet tourists' needs. The proposed model uses textual data from Saudi Arabian tourists to identify and track tourist sentiment, thereby increasing tourist satisfaction by resolving any problems they encounter, resolving issues affecting the hospitality and tourism sectors due to social media, and measuring tourists' satisfaction with visited places and proposed services in Saudi Arabia. Additionally, we help and build unique, unmatched strategies for promoting tourism in Saudi Arabia based on sentiment research.

The questions of the paper are: 1) can a dataset for tweets related to Saudi tourism improve such industry. 2) are we need the dataset to be in both Arabic and English. The main contribution of this work is the creation of the first dataset for tweets related to Saudi tourism (FDTST) in both Arabic and English. This study is still in its infancy in terms of the Saudi Arabian tourism business. Twitter, one of the most popular microblog platforms, is utilized in the proposed study, which may be an effective method for evaluating Twitter data to determine how individuals feel about popular travel websites. To achieve this, we use the Twitter API to stream tweets, filter relevant tweets using a search query, perform sentiment analysis on Twitter textual data from Saudi Arabia tourism to determine how people feel about each tourist site, and save the tweets for future research. On this basis, we can contribute to the development of the Kingdom's 2030 vision, as tourism is essential to the advancement and prosperity of a nation.

4. PROPOSED METHODOLOGY

The planned strategy was divided into seven phases. The initial stage is a collection. In the subsequent step, the Tweets dataset will be preprocessed for English and Arabic text to remove symbols, perform stemming and lemmatization, and other operations. In the third stage, characteristics will be extracted to allow for the extraction of valuable terms from tweets. In the fourth step, features will be selected to identify the characteristics and improve classification accuracy. The fifth stage involves using several classifiers to predict the mood of a tweet, which could be good or negative (positive or negative). The result will next be analyzed in the sixth stage. Finally, as illustrated in Figure 2, certain regularly used strategies will be employed to represent works to evaluate the automated classification method.

5. TWEETS COLLECTION AND PREPROCESSING

5.1 Data Collection Mechanism

In this research, Twitter was used as a source for collecting data because it is the most used micro-blogging site, and Twitter users vary from ordinary people, influencers, businessmen and various personalities. Moreover, different age groups and religious and social backgrounds from all countries are present on Twitter, so it is an excellent resource for sentiment analysis. Contact with support for Twitter and request a developer account, A valid reason must be conveyed for academic research, and when approved by Twitter, we will be provided with primary keys such as consumer key, access token key, and secret key for data access. then we use the passwords given after opening an account Developer to authenticate with Twitter API. Data collection functions: The verified accounts for tourism were manually selected according to their reliability (that is, all accounts are verified on Twitter and specialized in tourism) and their competence in the most important tourist areas in Saudi Arabia (alulamoments, NEOM, visitsaudinow, saudi_mt, SouthernSaudia, alulaguide, asirtourism, tourism_in_ksa, Scth_Jeddah, Sea, TheRedSeaSA, JED_SEASON, tourisminksa21, SaudiTourismd2d, GEA_SA, traveldiv, infosaudiacom, TirhalSA, till now, ExperienceAlUla, _Saudi_Tourism) implementation of the data collection process using the Python programming language.

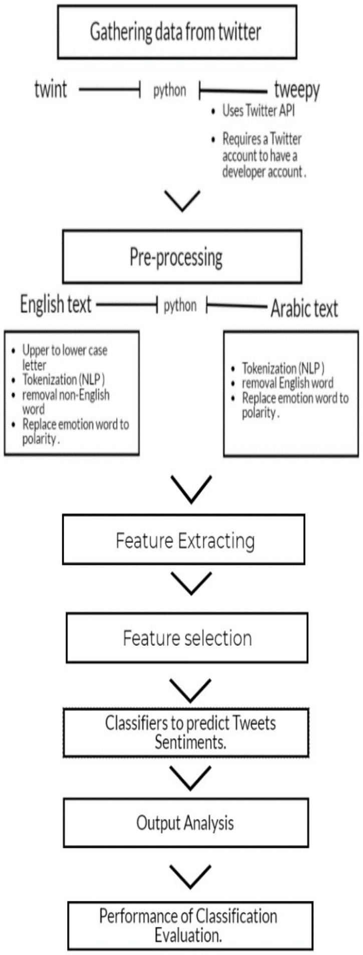

Gathering data from twitter
twint
python
tweepy
• Uses Twitter API
• Requires a Twitter account to have a developer account.
Pre-processing
English text
python
Arabic text
• Upper to lower case letter
• Tokenization (NLP)
• removal non-English word
• Replace emotion word to polarity.
• Tokenization (NLP)
• removal English word
• Replace emotion word to polarity.
Feature Extracting
Feature selection
Classifiers to predict Tweets Sentiments.
Output Analysis
Performance of Classification Evaluation.

Figure 2: The Proposed Methodology

5.2 Most important libraries

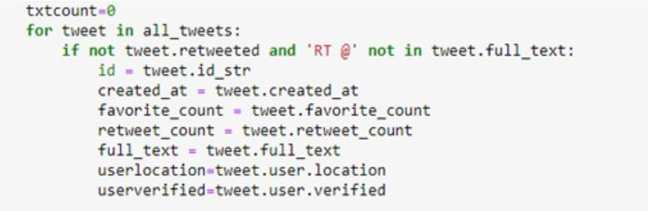
The most important libraries used in this research are the Tweepy library, for using Twitter API in Python language, and the CSV library, which converts files from one format to another. The Time library is used for calculating the time between pulling data from one account to another account, and the Pandas library is for handling the operations on collected data. In addition, The Nltk library helps with preprocessing steps. The matplotlib. pyplot as plt library that is used for visualization purposes. The Stopwords library deletes stopwords words that do not have meaning to the data. The SnowballStemmer library, through which we work as a preprocessing step.

5.3 Detailed explanation:

There were restrictions for the Tweepy library to access the Twitter API. The biggest limitation of Tweepy is that it has different limits depending on the level of the account. One of the advantages of Twitter API is that we can get data for free. Initially, a developer account was created on Twitter, requiring permission from 14 to 21 days. The project was named Twitter API (Saudi tourism). Key and Token were obtained, and then we started working using Jupiter based on Python programming language. After that, we created a folder containing two files, the first selected accounts for tourism in both Arabic and English to pull the tweets Including that.

5.4 Data cleaning

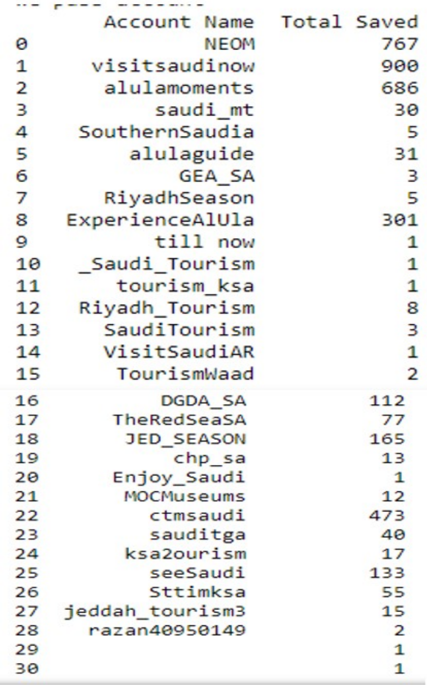
In addition, some of the accounts when running them in Jupiter did not show any results and therefore were removed (note: some accounts are common for Arabic and English), the second file config a Python file containing Consumer_key (Consumer_token, access_token, access_token_secret) after running the codes, a CSV file will appear. And another Excel file, after that, data was extracted in Arabic (23042) and data was extracted in English (19281). A condition has been added to the data to ensure that the data is relevant. The condition collects the original tweets of the verified accounts only. in addition, the attributes have been modified. We added location and verified to ensure that the accounts It is documented, and some of them, such as the URL, have been deleted because they are of no use in sentiment analysis. After adding the condition (it only collects the original tweets of the verified accounts, any tag of the account or re-tweet is not added and is not saved in the list of tweets, as shown in Figure 3), tweets in Arabic became (9513), and tweets in English (10457). In the cleaning phase, we made statistics so that more extracted details about the data appeared, as shown in Figure 4. Furthermore, and as depicted in Figure 5, the null values have been verified for all attributes



```
txtcount=0
for tweet in all_tweets:
    if not tweet.retweeted and 'RT @' not in tweet.full_text:
        id = tweet.id_str
        created_at = tweet.created_at
        favorite_count = tweet.favorite_count
        retweet_count = tweet.retweet_count
        full_text = tweet.full_text
        userlocation=tweet.user.location
        userverified=tweet.user.verified
```

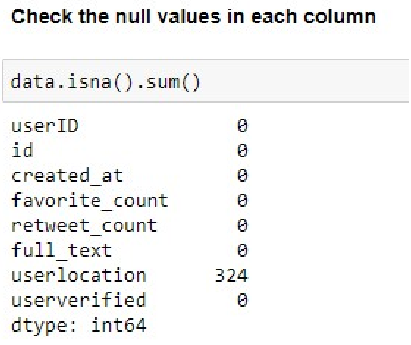
Figure 3: Tweets Collection from the Verified Accounts

By adding a code to show the statistics that appear when the data is withdrawn, the data in Arabic became 8798, and in English became 3837 because every time the code is running, the data is repeating. After adding the statistics code that contained duplicate data, we made sure that the data that appeared in the statistics and compared them with what appeared in the Excel sheet. Many accounts were searched, but only a few numbers appeared. Many accounts were removed because they did not contain any data such as (tourism_in_ksa / SaudiTourism2d / tourisminksa21 / riyadhintourism / STS_KSU / SeeSaudi1 / kiram / MohAlsaid / TourismJeddah / Jeddah_events1). The Twint library in Python collects the data in English, but this library is not stable and has many issues. When I pull the data, it comes with all data regardless of the language or configured filter. Meanwhile, I used as mentioned in this URL https://github.com/twintproject/twint. The data, it's so little and not retrieval with all data for my test, I used about five accounts and hashtags, and the retrieval round 100 records with different Languages. By the guidelines, the lib has a limitation with a 3200 record. The server has not been stable because sometimes it is working and some is not working; during my test, the lib was down about three times.



```
       Account Name  Total Saved
0              NEOM          767
1      visitsaudinow         900
2       alulamoments         686
3           saudi_mt          30
4     SouthernSaudia           5
5         alulaguide          31
6             GEA_SA           3
7       RiyadhSeason           5
8    ExperienceAlUla         301
9           till now           1
10     _Saudi_Tourism          1
11       tourism_ksa           1
12    Riyadh_Tourism           8
13      SaudiTourism           3
14      VisitSaudiAR           1
15       TourismWaad           2
16            DGDA_SA        112
17        TheRedSeaSA         77
18         JED_SEASON        165
19             chp_sa         13
20        Enjoy_Saudi          1
21         MOCMuseums         12
22           ctmsaudi        473
23           sauditga         40
24         ksa2ourism         17
25           seeSaudi        133
26           Sttimksa         55
27    jeddah_tourism3         15
28      razan40950149          2
29                             1
30                             1
```

Figure 4: Statistics of English Tweets



Check the null values in each column

```
data.isna().sum()
```

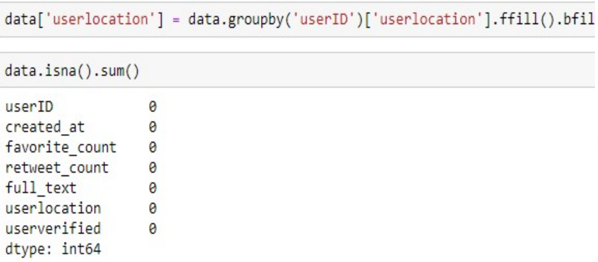
```
userID              0
id                  0
created_at          0
favorite_count      0
retweet_count       0
full_text           0
userlocation      324
userverified        0
dtype: int64
```

Figure 5: Accounts with null English Tweets

The tweets were collected through hashtags for tourist areas and events. After going through them manually, it became clear that many repetitions and tweets have nothing to do with the specific topic and are difficult to clean and purify. There are no specific restrictions in writing within the hashtag on Twitter. It may be within the hashtag tweets marketing for an unrelated topic, invitations to an acquaintance, literary and religious expressive phrases, etc. Those are not related to the hashtag, so choosing to collect data from accounts on Twitter ensured that they were more relevant to tourism. Ultimately, our machine learning model is expected to parse the text of a Tweet about views of Saudi Arabia tourism at the sentence level and categorize them as positive, negative, or neutral. The selected tweets convey some feelings (positive or negative), and the objective of our model is to extract valuable information from such tweets to determine the sentiment orientation of the input text. Based on two human experts, the months-long annotation process of the tweets is manually conducted. If the experts agree on the label of a particular tweet, then the tweet is assigned this label. Otherwise, a third expert is consulted to break the tie.

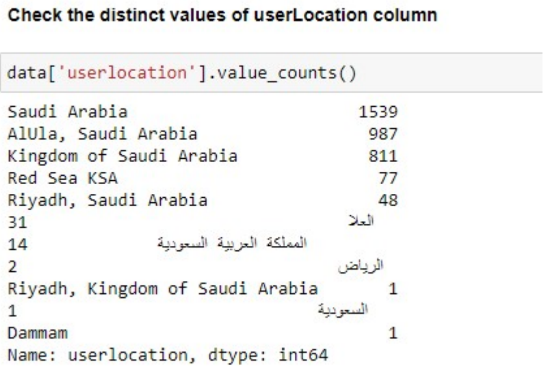
Developing a novel machine learning algorithm for feature selection is the next step in this research—the issue id to how to get suitable features from the collected dataset by the developed algorithm. Also, developing a novel machine-learning algorithm for classification is required to classify the extracted feature from the dataset to get valuable suggestions to improve tourism in Riyadh.

Numerous steps and procedures are involved in the cleaning, preprocessing, and visualization phases to assure stability and suitability of the data for the planned objectives and ambitions, Figures 6 to 21 illustrate examples of such proceedings.



```
data['userlocation'] = data.groupby('userID')['userlocation'].ffill().bfil

data.isna().sum()

userID           0
created_at       0
favorite_count   0
retweet_count    0
full_text        0
userlocation     0
userverified     0
dtype: int64
```

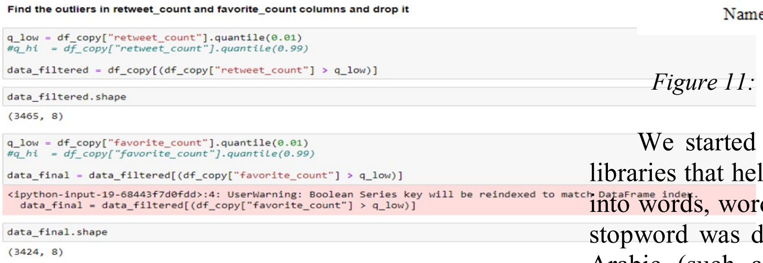
Figure 6: Tweets grouped by user ID and user location of English Tweets



Check the distinct values of userLocation column

```
data['userlocation'].value_counts()

Saudi Arabia                           1539
AlUla, Saudi Arabia                     987
Kingdom of Saudi Arabia                 811
Red Sea KSA                              77
Riyadh, Saudi Arabia                     48
31                                     العلا
14              المملكة العربية السعودية
2                                    الرياض
Riyadh, Kingdom of Saudi Arabia           1
1                                  السعودية
Dammam                                    1
Name: userlocation, dtype: int64
```

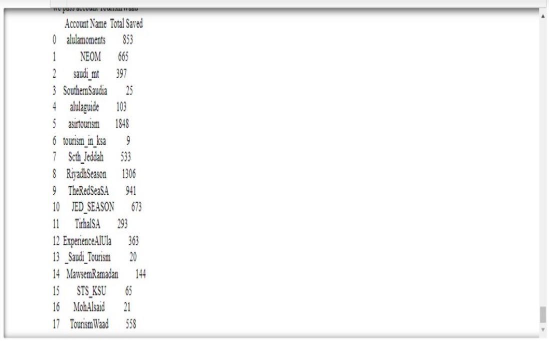
Figure 7: User Location Data of English Tweets



Find the outliers in retweet_count and favorite_count columns and drop it

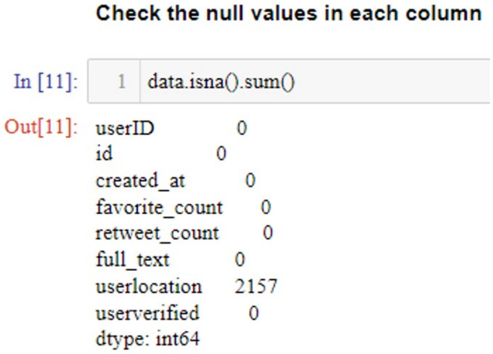
```
q_low = df_copy["retweet_count"].quantile(0.01)
#q_hi  = df_copy["retweet_count"].quantile(0.99)

data_filtered = df_copy[(df_copy["retweet_count"] > q_low)]

data_filtered.shape

(3465, 8)

q_low = df_copy["favorite_count"].quantile(0.01)
#q_hi  = df_copy["favorite_count"].quantile(0.99)

data_final = data_filtered[(df_copy["favorite_count"] > q_low)]

<ipython-input-19-68443f7d0fdd>:4: UserWarning: Boolean Series key will be reindexed to match DataFrame index.
  data_final = data_filtered[(df_copy["favorite_count"] > q_low)]

data_final.shape

(3424, 8)
```

Figure 8: Drop retweet and favorite count data of English Tweets



```
we pass account TourismWaad
      Account Name  Total Saved
0     alulamoments          853
1             NEOM          665
2          saudi_mt         397
3    SouthernSaudia          25
4       alulaguide          103
5       asirtourism        1848
6     tourism_in_ksa          9
7       Scth_Jeddah         533
8      RiyadhSeason        1306
9       TheRedSeaSA         941
10       JED_SEASON         673
11          TirhalSA        293
12  ExperienceAlUla         363
13    _Saudi_Tourism         20
14    MawsemRamadan         144
15           STS_KSU         65
16         MohAlsaid         21
17       TourismWaad        558
```

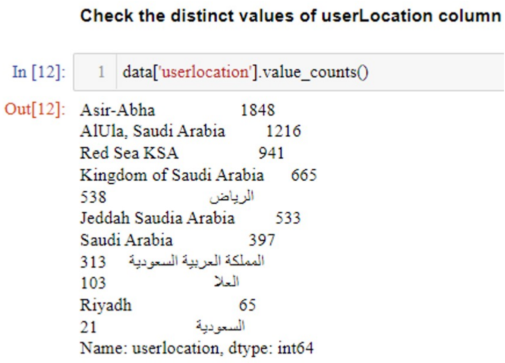
Figure 9: Statistics of Arabic Tweets



Check the null values in each column

```
In [11]:  data.isna().sum()

Out[11]:  userID             0
          id                 0
          created_at         0
          favorite_count     0
          retweet_count      0
          full_text          0
          userlocation    2157
          userverified       0
          dtype: int64
```

Figure 10: Accounts with null Arabic Tweets



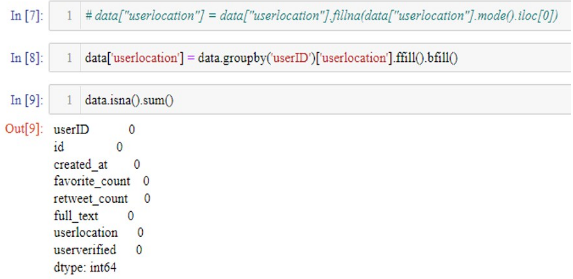
Check the distinct values of userLocation column

```
In [12]:  data['userlocation'].value_counts()

Out[12]:  Asir-Abha                  1848
          AlUla, Saudi Arabia        1216
          Red Sea KSA                 941
          Kingdom of Saudi Arabia     665
          538                       الرياض
          Jeddah Saudia Arabia        533
          Saudi Arabia                397
          313     المملكة العربية السعودية
          103                         العلا
          Riyadh                       65
          21                       السعودية
          Name: userlocation, dtype: int64
```

Figure 11: User Location Data for Arabic Tweets

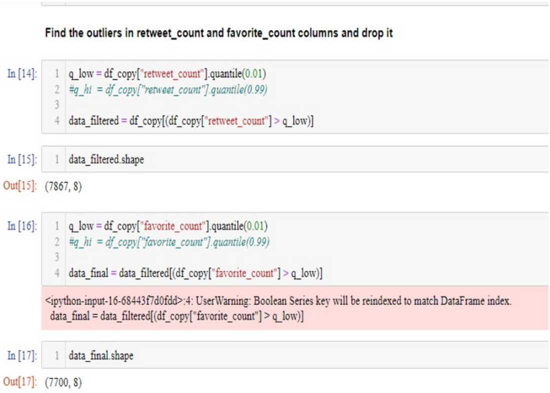
We started the cleaning process using several libraries that help with this. Sentences were divided into words, words were divided into letters, and the stopword was deleted (such as the, is, are) and in Arabic (such as who, who, he, and she). New columns have been added which are tweet_without_stopwords, stemmed_without_stopWords, lemmatized_without_stopWords. It might make sense to delete the stopword because it is words that do not add sentiment information to the tweets. The missing data was extracted and obtained in usreloction in Arabic 2157 and English 340 are all in userloction and will be filled in manually. A code was tried but we failed so it will be manually filled out by userid and we have tried to filter the data from the ads by specifying certain words indicating that this text (ad text) is in English such as (follow the action- stay tuned- racing- watch) And the Arab, such as (register now - share with us - announcement).

Visualization was done for more usrelocation from which tweets appeared and more userid from which tweets appeared. Visualization was done for the outlier for the two columns (retweet count - favorite_count) and the id column was deleted because we see that it is not useful in sentiment analysis. It is a series of numbers that are not useful to us in sentiment analysis and we have not finished the cleaning process. A code was tried to fill in the empty values by userid, but it took the highest repetition account and put it in the empty places, and this is wrong because we will have a bias for one account and one area, so we used another code that is filled with more than one value from the userid to ensure that there is no bias.



```
In [7]:  # data["userlocation"] = data["userlocation"].fillna(data["userlocation"].mode().iloc[0])

In [8]:  data['userlocation'] = data.groupby('userID')['userlocation'].ffill().bfill()

In [9]:  data.isna().sum()

Out[9]:  userID          0
         id              0
         created_at      0
         favorite_count  0
         retweet_count   0
         full_text       0
         userlocation    0
         userverified    0
         dtype: int64
```

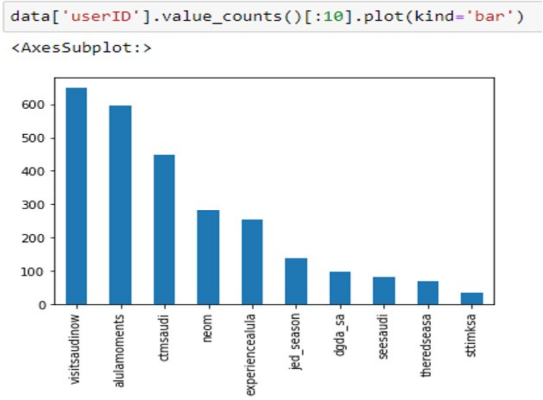
Figure 12: Tweets grouped by user ID and user location of Arabic Tweets



Find the outliers in retweet_count and favorite_count columns and drop it

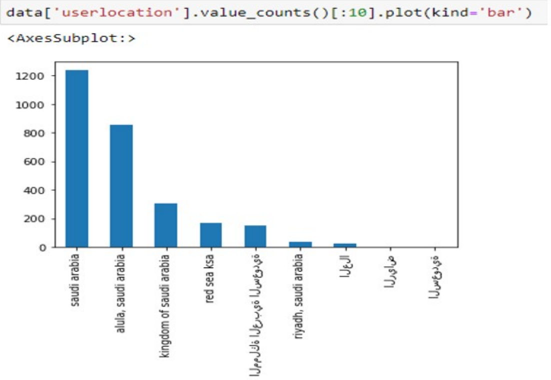
```
In [14]:  q_low = df_copy["retweet_count"].quantile(0.01)
          #q_hi  = df_copy["retweet_count"].quantile(0.99)

          data_filtered = df_copy[(df_copy["retweet_count"] > q_low)]

In [15]:  data_filtered.shape

Out[15]:  (7867, 8)

In [16]:  q_low = df_copy["favorite_count"].quantile(0.01)
          #q_hi  = df_copy["favorite_count"].quantile(0.99)

          data_final = data_filtered[(df_copy["favorite_count"] > q_low)]

          <ipython-input-16-68443f7d0fdd>:4: UserWarning: Boolean Series key will be reindexed to match DataFrame index.
            data_final = data_filtered[(df_copy["favorite_count"] > q_low)]

In [17]:  data_final.shape

Out[17]:  (7700, 8)
```

Figure 13: Drop retweet and favorite count data of Arabic Tweets



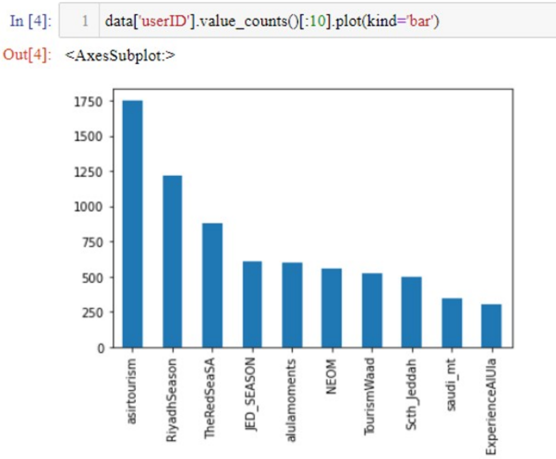
```
data['userID'].value_counts()[:10].plot(kind='bar')
<AxesSubplot:>
```

Figure 14: Counts of English Tweets based on user ID



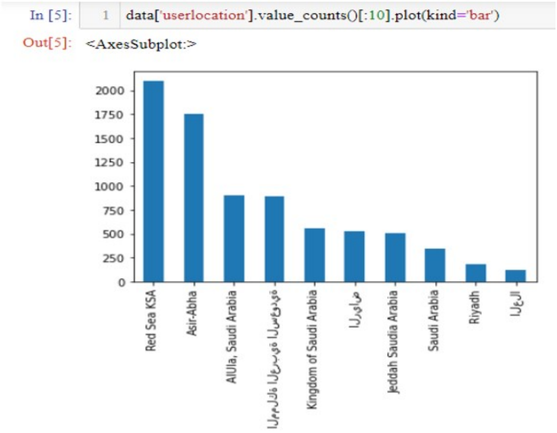
```
data['userlocation'].value_counts()[:10].plot(kind='bar')
<AxesSubplot:>
```

Figure 15: Counts of English Tweets based on user location



```
In [4]:  data['userID'].value_counts()[:10].plot(kind='bar')

Out[4]:  <AxesSubplot:>
```

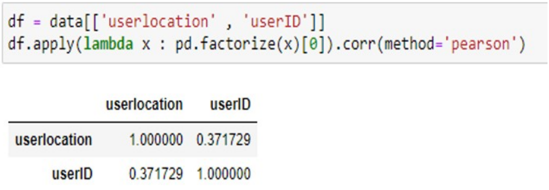
Figure 16: Counts of Arabic Tweets based on user ID



```
In [5]:  data['userlocation'].value_counts()[:10].plot(kind='bar')
Out[5]:  <AxesSubplot:>
```

Figure 17: Counts of Arabic Tweets based on user location

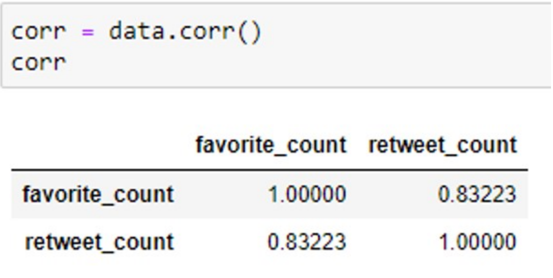
The correlation between "userID" and "userlocation" is analyzed for English and Arabic Tweets to identify the features that would be the best inputs to the machine learning model. The correlation between "retweet_count" and "favorite_count" is also analyzed. A value closer to 1 means a higher correlation. The following shows the correlation between different values.



```
df = data[['userlocation' , 'userID']]
df.apply(lambda x : pd.factorize(x)[0]).corr(method='pearson')
```

```
             userlocation    userID
userlocation     1.000000  0.371729
      userID     0.371729  1.000000
```

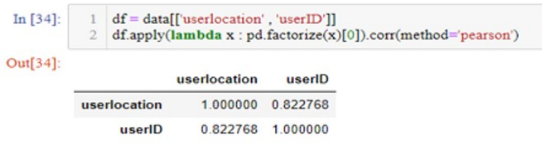
Figure 18: Correlation between "userID" and "userlocation" of English Tweets



```
corr = data.corr()
corr
```

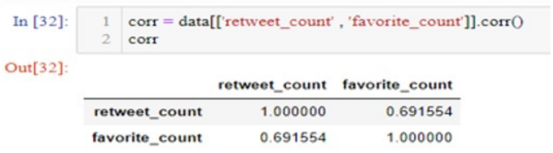
```
                favorite_count  retweet_count
favorite_count         1.00000        0.83223
 retweet_count         0.83223        1.00000
```

Figure 19: Correlation between "retweet_count" and "favorite_count" of English Tweets



```
In [34]:  df = data[['userlocation' , 'userID']]
          df.apply(lambda x : pd.factorize(x)[0]).corr(method='pearson')
Out[34]:
             userlocation    userID
userlocation     1.000000  0.822768
      userID     0.822768  1.000000
```

Figure 20: Correlation between "userID" and "userlocation" of Arabic Tweets



```
In [32]:  corr = data[['retweet_count' , 'favorite_count']].corr()
          corr
Out[32]:
               retweet_count  favorite_count
 retweet_count      1.000000        0.691554
favorite_count      0.691554        1.000000
```

Figure 21: Correlation between "retweet_count" and "favorite_count" of Arabic Tweets

5.5 Data extracted by hashtag

According to the Datasets - Saudi Open Data website, the main cities in the northern region are: Tabuk, Arar, Sakaka, and Hail. In the Eastern Province, they are represented in Dammam, Al-Ahsa, and Al-Khobar. And in the middle region are Riyadh and Al-Qassim. In the western region, it is represented in: Jeddah, Taif, Madinah, and Makkah. In the southern region, it is represented in Al Baha, Jazan, Najran, and Abha. According to the official website of Saudi tourism, Visit Saudi - the official website of Saudi tourism (visitsaudi.com), The tourist destinations in Tabuk are Tabuk Castle, Wadi Al-Disah, and Maghair Shuaib. And in Al-Jawf-Sakaka are: Aamdt AL- Rajajeel and Qal'at Zaabal. In Hail, they are the Nufud desert, the Ayraf castle, the Barzan castle, and the Qishleh castle. This is for the North. As for the Eastern Province, the tourist destinations in Dammam are Half Moon Beach, the Saudi Cultural Center, and the Khobar Waterfront. And in Al-Ahsa are Al-Qarah, Jawatha Mosque, Al-Qaysaria Market, and Al-Asfar Lake. As for the central region, the tourist destinations in Riyadh are Al-Masmak Palace, Al-Zal Market, Najd Village, King Fahd International Stadium, Al-Turaif Palace, and Al-Diriyah. As for the western region, the tourist destinations in Jeddah are Al Balad, Jeddah Corniche, and Al Tayebat City. And in Taif: Jabal al-Hada, Taif Central Market, and rose gardens. In Al-Ula - Medina: Al-Hajar, Dadan and Jabal Ikmah.

As for the southern region, the tourist destinations in Al-Baha are Dhi Ain, Raghadan Forest, and Shada Village. In Jizan: the coast of Jizan and the Farasan Islands. In Abha: The green mountain, Al-Habla hanging village, and the caves of Shada mountain. Also, according to the official Saudi Tourism website, the events and activities held in Riyadh are represented by the Riyadh

Season event, which includes: Riyadh Front, Wonderland, Riyadh City Boulevard, The Groves, Riyadh Pulse, Oasis, Formula, and Diriyah. In Al-Ula, the events include Al-Ula moments, a 4×4 safari in Sharaan Nature Reserve, the Via Ferrata experience, Annabelles Al-Ula, historical performances in the old town of Al-Ula, the arts of Al-Ula, a dinner experience, the balloons experience in Al-Ula, the Al-Ula sky, a constellations show, a music show under the Stars, AlUla Festival for Recreation and Relaxation, The Five Senses Sanctuary, Moments Garden and Eco-Trail.

In the data extracted by hashtag, we need the name of the hashtag and the text of the tweet only. The hashtags were determined based on what came from tourist areas on the official website of tourism in Saudi Arabia and the hashtags found in the documented accounts of events in the Kingdom of Saudi Arabia. We collected the tweets using the Tweepy library through Twitter API on Google Collab. Using the script commands:

```
import pandas as pd
import tweepy

# function to display data of each tweet
def printtweetdata(n, ith_tweet):
        print()
        print(f"Tweet {n}:")
        print(f"Tweet Text:{ith_tweet[0]}")
        print(f"Hashtags Used:{ith_tweet[1]}")

# function to perform data extraction
def scrape(words, date_since, numtweet):

        # Creating DataFrame using pandas
        db = pd.DataFrame(columns=['text',
                                   'hashtags'])

        # We are using .Cursor() to search
        # through Twitter for the required tweets.
        # The number of tweets can be
        # restricted using .items(number of tweets)
        tweets = tweepy.Cursor(api.search,
                  words, lang="ar",
                  since_id=date_since,
                 tweet_mode='extended').items(numtwee)


   # .Cursor() returns an iterable object. Each item in
        # the iterator has various attributes
        # that you can access to
        # get information about each tweet
        list_tweets = [tweet for tweet in tweets]

       # Counter to maintain Tweet Count
        i = 1

      # we will iterate over each tweet in the
     # list for extracting information about each tweet
        for tweet in list_tweets:
                hashtags = tweet.entities['hashtags']

                # Retweets can be distinguished by
                # a retweeted_status attribute,
                # in case it is an invalid reference,
                # except block will be executed
                try:
                      text = tweet.retweeted_status.full_text
                except AttributeError:
                        text = tweet.full_text
                hashtext = list()
                for j in range(0, len(hashtags)):
                        hashtext.append(hashtags[j]['text'])

                # Here we are appending all the
                # extracted information in the DataFrame
                ith_tweet = [ text, hashtext]
                db.loc[len(db)] = ith_tweet

             # Function call to print tweet data on screen
                printtweetdata(i, ith_tweet)
                i = i+1
        filename = 'KSA.csv'

        # we will save our database as a CSV file.
        db.to_csv(filename)

if __name__ == '__main__':

        # Enter your own credentials obtained
        # from your developer account
        consumer_key = "yh3yDf1Kk3sYzEFqkI2prk
5dl"# API KEY
        consumer_secret = "YI7gVqQ7loC5zcxulk4K
L1flQeiL0MA1VfzAQTQkRrwqO1fW2v"# API
SECRET KEY
        access_key = "1003952741708914688-
eyOdK9StnnfFnEpibSyaM1rWK33eWL"# token
KEY
      access_secret = "W4v4MiAaFQQ5TYgSxky3m
Qzb3frcTzHYYCw3jz3tzsklC"# token secret KEY


    auth = tweepy.OAuthHandler(consumer_key, co
nsumer_secret)
   auth.set_access_token(access_key, access_secret)
        api = tweepy.API(auth)
    # Enter Hashtag and initial date
    print("Enter Twitter HashTag to search for")
    words = input()
    print("Enter Date since The Tweets are require
d in yyyy-mm--dd")
    date_since = input()

  # number of tweets you want to extract in one run
    numtweet = 10000
    scrape(words, date_since, numtweet)
    print('Scraping has completed!')
```

The extract was in the name of the hashtag and the period from 01-01-2012 for tourist areas. As for the new events, it was in the period from 1-1-2020. Through observation, tweets containing the hashtag's name are collected and not necessarily the presence of the hashtag within it. 16,044 tweets were collected through approximately 75 hashtags. By going through the data, it became clear that there are many repetitions, as the tweet is repeated with the number of retweets. For example, if there is a tweet that got 200 retweets, it will be repeated 200 times and collect responses to the tweets as well. The data collected by hashtag contains many related tweets and tweets that are not related, such as an advertisement for a product, an invitation to get to know each other, and others. We first started by deleting the duplicates through the text of the tweet using the code:

```
import pandas as pd
import tweepy
import numpy as np
import os
os.getcwd()
df = pd.read_excel('.xlsx')
#Remove duplicates tweets

df=df.drop_duplicates(subset=['text'])
#Number of tweet after remove duplicates
df.value_counts(df['text']).sum()
clean_df= df[['text']]
def remove_hashtag(df, col = 'text'):
  for letter in r'#.][!XR':
    df[col] = df[col].astype(str).str.replace(letter,",
regex=True)

remove_hashtag(df)
clean_df
clean_df.to_csv('Clean_Tweets.csv')
```

The number of tweets reached approximately 4,044. We manually filtered the tweets so that the tweets related to tourism in the Kingdom of Saudi Arabia were personal opinions and not objective facts. It took about 6 days for the number of tweets to reach 925. By cleaning the data extracted from the accounts of tourism in the Kingdom of Saudi Arabia, we noticed that the tweets that contain facts and news about tourism outnumber the tweets that express a personal opinion. This resulted in a significant decrease in the number of tweets. We wrote down the most prominent tourist places and events mentioned in these tweets and extracted them using the hashtag to get a larger number of tweets. We searched through more than 80 hashtags, some of which provided us with several tweets, and some of them did not return any tweets. The number of extracted tweets was recorded in the excel file. Some hashtags require searching through them more than once on separate days to retrieve tweets. The number of hashtag data has increased from 925 tweets to 2,000 tweets. We then manually categorized them into positive, negative, and neutral.

The number of tweets reached approximately English 2774. We manually filtered (cleaned) the tweets so that the tweets related to tourism in Saudi Arabia were personal opinions and not objective facts. It took some days for the number of tweets to reach 649, excluding advertisements and nonimportant tweets. Additional 1096 tweets are then added to increase the number of tweets. The additional tweets are also cleaned/ filtered and classified in the same way. Now, the number of tweets reached 950 tweets in total. The number of tweets reached approximately Arabic7915, we manually filtered (cleaned) the tweets so that the tweets related to tourism in Saudi Arabia were personal opinions and not objective facts. It took some days for the number of tweets to reach 1902, The text of a Tweet about the views of Saudi Arabia tourism at the sentence level is categorized as positive, negative, or neutral. The selected tweets convey some feelings (positive or negative), and the objective of the machine learning model will be to extract valuable information from such tweets to determine the sentiment orientation of the input text. The tweets dataset collected from Twitter libraries in Arabic and English languages is considered the first tourism dataset in Saudi Arabia. Therefore, a vital and important contribution of this project is creating the First Dataset for Tweets Saudi Tourism (FDTST) in both Arabic and English languages collected from Twitter libraries .

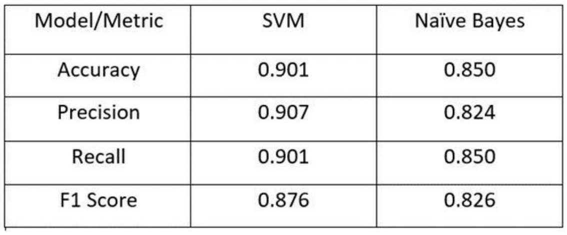
6. EXPERIMENTAL AND RESULT ANALYSIS

The experiments use Intel(R) Core(TM) i7-10510U CPU with 2.30 GHz and 12.0 GB RAM. 64-bit Windows 11 operating system based on the x64-based processor is running.

The cleaning of the dataset was done manually, and the classification of the tweets was also done manually. After trying different codes to classify the initial tweets, the output did not show satisfactory results, so the classification task of the dataset was done manually. The Arabic data reached 1900 tweets, and the English data reached 950 tweets after cleaning/filtering. A python-based code is applied to show the number of tweets from each user, and we used another code to indicate the number of positive, negative, and neutral Tweets. A clear bias for the positive is applied, and to solve this problem, we used a redistribution function, which solved the bias issue.

Several previous works from different years were studied, the works in which classification algorithms implemented in the field of sentiment analysis were used and were the most efficient ones that were used in this research, respectively SVM, Naïve Bayes, logistic regression and random forest classifier. The dataset was divided into test and train data, with 20% for testing and 80% for training. The method worked with Vectorizer, which is famous in the field of sentiment analysis, to convert tweets into numbers (0 and 1) for the model to understand the tweets. Two base classifier models of Support Vector Machine (SVM) and Naïve Bayes (NB) are applied to the collected Arabic and English datasets. Four types of SVM, named linear- poly-rbf-sigmoid, were tried, and the results show that the best accuracy for Arabic and English is achieved when using the rbf type, which achieved high accuracy. For the NB, three types, named bernoulli, and complement, were applied, and the best accuracy appeared in Arabic and English datasets when using the complement type. Table 1 shows the classification accuracy, precision, recall, and F1 score for the English dataset based on SVM and NB classifiers. Table 2 shows the classification accuracy, precision, recall, and F1 score for the Arabic dataset based on SVM and NB classifiers.

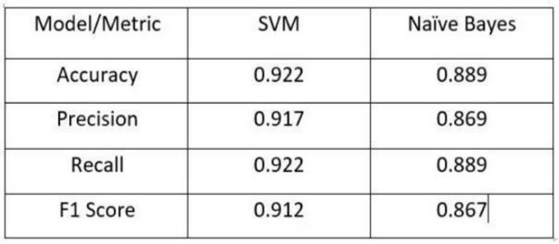
Table 1: Classification Accuracy, Precision, Recall, and F1 Score for the English Dataset



Model/Metric	SVM	Naïve Bayes
Accuracy	0.901	0.850
Precision	0.907	0.824
Recall	0.901	0.850
F1 Score	0.876	0.826

We first used the support vector machine with different types of kernels but the RBF kernel gives us the best results as follows:
It managed to reach 0.901 accuracy, 0.907 for precision, 0.901 for recall, and finally 0.876 for f1-score which are acceptable results. The second model is the Naïve Bayes and for sure we tried many types of it including Complement NB, Bernoulli NB. The Complement NB had the best performance. It managed to get 0.850 accuracy, 0.824 for precision, 0.850 for recall, and finally 0.826 for f1-score. As we saw in the table above: Support Vector Machine has a better performance than Naïve Bayes regarding all the evaluation matrices that we used including Accuracy, Precision, Recall, and F1-Score.

Table 2: Classification Accuracy, Precision, Recall, and F1 Score for the Arabic Dataset



Model/Metric	SVM	Naïve Bayes
Accuracy	0.922	0.889
Precision	0.917	0.869
Recall	0.922	0.889
F1 Score	0.912	0.867

Again, in the Arabic dataset, we implemented the 2 types of models which are SVM and Naïve Bayes. We first used the support vector machine with different types of kernels including poly, RBF, and linear. the RBF gives us the best results as follows:

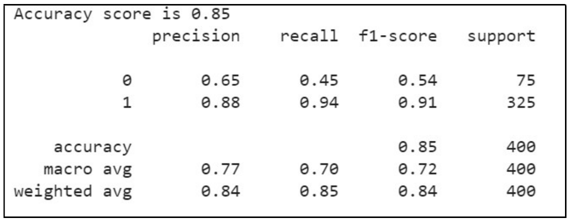
It managed to reach 0.922 accuracy, 0.917 for precision, 0.922 for recall, and finally 0.912 for f1-score which are acceptable results.

The second model is the Naïve Bayes and for sure we tried many types of it including Complement NB, Bernoulli NB . The Complement NB had the best performance. It managed to get 0.889 accuracy, 0.869 for precision, 0.889 for recall, and finally 0.867 for f1-score.
As we saw in the table above:
Support Vector Machine has a better performance than Naïve Bayes regarding all the evaluation

matrices that we used including Accuracy, Precision, Recall, and F1-Score.

The number of data extracted from the hashtag into 2000 rows, representing the positive classification with the number (1) and the negative with the number (0), and dividing it through the code into 20% test and 80% training, and several machine learning algorithms were applied, and the accuracy calculated for them. Accordingly, Tables 3 to 6 illustrate and summarize the evaluation result of the developed classification model

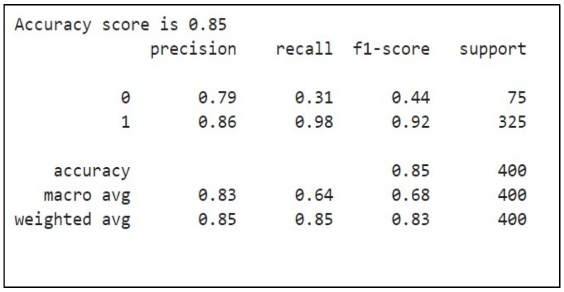
Table 3: Classification Accuracy, Precision, Recall, F1 Score, and Support for the English Dataset Using Logistic Regression



```
Accuracy score is 0.85
              precision    recall  f1-score   support

           0       0.65      0.45      0.54        75
           1       0.88      0.94      0.91       325

    accuracy                           0.85       400
   macro avg       0.77      0.70      0.72       400
weighted avg       0.84      0.85      0.84       400
```

As shown above, this is the classification for applying the logistic regression model to the Arabic dataset. It managed to reach accuracy 0.85 for accuracy, 0.85 for recall, 0.84 for precision, and finally 0.84 for f1-score.

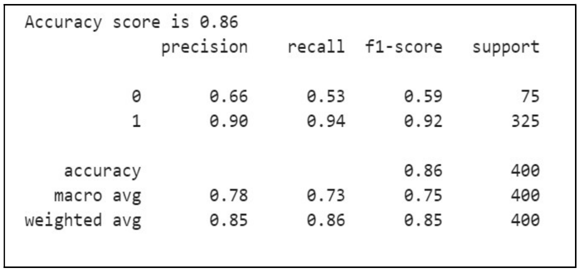
Table 4: Classification Accuracy, Precision, Recall, F1 Score, and Support for the English Dataset Using Random Forest Classifier



```
Accuracy score is 0.85
              precision    recall  f1-score   support

           0       0.79      0.31      0.44        75
           1       0.86      0.98      0.92       325

    accuracy                           0.85       400
   macro avg       0.83      0.64      0.68       400
weighted avg       0.85      0.85      0.83       400
```

As shown above, this is the classification for applying the Random Forest Classifier to the Arabic dataset. It managed to reach accuracy 0.85 for accuracy, 0.85 for recall, 0.85 for precision, and finally 0.83 for f1-score.

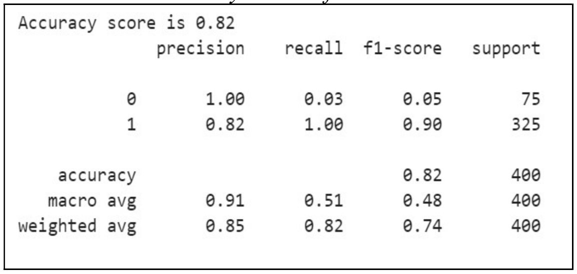
Table 5: Classification Accuracy, Precision, Recall, F1 Score, and Support for the English Dataset Using Support Vector Machine



```
Accuracy score is 0.86
              precision    recall  f1-score   support

           0       0.66      0.53      0.59        75
           1       0.90      0.94      0.92       325

    accuracy                           0.86       400
   macro avg       0.78      0.73      0.75       400
weighted avg       0.85      0.86      0.85       400
```

As shown above, this is the classification for applying the Support Vector Machine to the Arabic dataset. It managed to reach accuracy 0.86 for accuracy, 0.86 for recall, 0.85 for precision, and finally 0.85 for f1-score.

Table 6: Classification Accuracy, Precision, Recall, F1 Score, and Support for the English Dataset Using Naïve Bayes Classifier



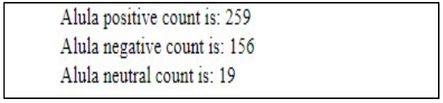
```
Accuracy score is 0.82
              precision    recall  f1-score   support

           0       1.00      0.03      0.05        75
           1       0.82      1.00      0.90       325

    accuracy                           0.82       400
   macro avg       0.91      0.51      0.48       400
weighted avg       0.85      0.82      0.74       400
```

As shown above, this is the classification for applying the Naïve Bayes to the Arabic dataset. It managed to reach accuracy 0.82 for accuracy, 0.82 for recall, 0.85 for precision, and finally 0.74 for f1-score.

7. VALUABLE INSIGHTS AND RECOMMENDATIONS

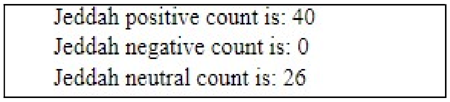
Over and above, the developed predictive models aid to appoint and specify several valuable recommendations and insights for continuous improvements and sustainable growth in the Saudi tourism industry. Indeed, such insights and advice are inferred based on the types and weight of the associated classifiers added to the values of some attributes such as the number of tweets and the number of favorites, followings are some brief examples.

- Alula has a majority of positive reviews and we encourage them to continue their amazing work (keep on).



```
Alula positive count is: 259
Alula negative count is: 156
Alula neutral count is: 19
```

- Jeddah: The results were excellent and acceptable.



```
Jeddah positive count is: 40
Jeddah negative count is: 0
Jeddah neutral count is: 26
```

8. CONCLUSION

The paper demonstrated machine learning-based sentiment analysis models for tweets on Saudi tourism. It examined and analyzed tweets related to tourism collected about six touristic places in Saudi Arabia from twenty accounts and nineteen touristic places from 144 hashtags. Two base classifier models of Support Vector Machine (SVM) and Naïve Bayes (NB) are applied. The First Dataset for Tweets Saudi Tourism (FDTST) in both Arabic and English languages collected from Twitter libraries is created. The developed classification models are evaluated based on their performance computation, and the experimental results show that the developed models have achieved righteous and reliable upshots.

REFERENCES:

[1] K. Ravi, And V. Ravi, A Survey On Opinion Mining And Sentiment Analysis: Tasks, Approaches And Applications, Knowledge-Based Systems, Vol. 89, 2015.

[2] R. Varghese And M. Jayasree, A Survey On Sentiment Analysis And Opinion Mining, International Journal Of Research In Engineering

[3] S. Hemamalini, Literature Review On Sentiment Analysis, International Journal Of Scientific & Technology Research, Vol. 9, Issue 04, April 2020.

[4] G. Vinodhini, And R. M. Chandrasekaran, Sentiment Analysis And Opinion Mining: A Survey, International Journal, Vol. 2, No. 6, Pp. 282-292, 2012.

[5] S. Pandya And P. Mehta, A Review On Sentiment Analysis Methodologies, Practices And Applications, 2020.

[6] Wenjie Xiao And Changguo Xiang, "Overview Of Tourism Data Mining In Big Data Environment, "7th International Conference On Education, Management, Computer And Medicine (Emcm 2016), Vol. 59, 2017.

[7] S. Mukherjee And P. Bhattacharyya, Sentiment Analysis: A Literature Survey, 2013.

[8] M. Hajmohammadi, R. Ibrahim, And Z. A. Othman, Opinion Mining And Sentiment Analysis: A Survey, International Journal Of Computers & Technology, Vol. 2, No. 3, Pp.171-178, 2012.

[9] S. Saad, And B. Saberi, Sentiment Analysis Or Opinion Mining: A Review, International Journal On Advanced Science, Engineering And Information Technology, Vol. 7, No. 5, Pp.1660, 2017.

[10] A. Adl, And A. K. Elfergany, Tracking How A Change In A Telecom Service Affects Its Customers Using Sentiment Analysis And Personality Insight, International Journal Of Service Science, Management, Engineering, And Technology (Ijssmet), Vol. 11, No. 3, Pp. 33-46. Http://Doi.Org/10.4018/Ijssmet.2020070103.

[11] A. Feizollah, M. M. Mostafa, A. Sulaiman, Et Al., Exploring Halal Tourism Tweets On Social Media. J Big Data, Vol. 8, No. 72, 2021. Https://Doi.Org/10.1186/S40537-021-00463-5

[12] I. P. Windasari And D. Eridani, Sentiment Analysis On Travel Destination In Indonesia, 2017.

[13] T. Kuhamanee, N. Talmongkol, K. Chaisuriyakul, W. San-Um, N. Pongpisuttinun, And S. Pongyupinpanich, Sentiment Analysis Of Foreign Tourists To Bangkok Using Data Mining Through Online Social Network, 2017.

[14] Y. A. Al-Mulla, Machine Learning In Tourism, 2020.