TWITTER SENTIMENT DATA ANALYSIS

A PROJECT REPORT

Submitted by

RAHSHITHA K S (21ADR037)
SOUMIYA K (21ADR049)
MANIKANDAN T (21ADR026)
THAMILARASON K B (21ADR055)

for

20ADC33 DATA ANALYSIS

DEPARTMENT OF ARTIFICIAL INTELLIGENCE

KONGU ENGINEERING COLLEGE

(Autonomous)

PERUNDURAI, ERODE – 638 060

DECEMBER 2022

DEPARTMENT OF ARTIFICIAL INTELLIGENCE

KONGU ENGINEERING COLLEGE

(Autonomous)

PERUNDURAI, ERODE – 638 060

DECEMBER 2022

DEPARTMENT OF ARTIFICIAL INTELLIGENCE

20ADC33 – DATA ANALYSIS PROJECT REPORT

Signature of course in-charge

Signature of the HOD

Submitted for the continuous Assessment viva voice examination held on _____

EXAMINER I

EXAMINER II

ABSTRACT

Nowadays, social media tend to be where people are more interactive and express their opinions more openly and confidently compared to real life. Twitter is one of the commonly used platforms where people express their views in the form of tweets. On Twitter, people sharing their opinions can be divided into many categories such as Education, Technology, News, Politics, Media, Filmography, Art, Culture, and many more. Thus, among the different social media platforms that are currently available, Twitter can be considered one of the most significant arenas which play a major role in determining crucial factors such as business product reviews, public opinion, and emotions of a text.

This mini project mainly focuses on analyzing the sentiment of the text available in the tweets posted by the public. The emotions behind any text or sentence are to be analyzed using Natural Language Processing or using an algorithm. The purpose of performing sentiment analysis can be categorized under various purposes. One of the main purposes is to analyze the reviews of any new business products or brands or even any newly launched technology. By doing so the businessmen can understand if not their product has received a positive response. Similarly, it can be used to know the popularity of a brand over the world.

TABLE OF CONTENTS

CHAPTER - 1

INTRODUCTION

1.1. INTRODUCTION

This mini project mainly focuses on analyzing the sentiment of the text available in the tweets posted by the public. The emotions behind any text or sentence are to be analyzed using some techniques- which may be Natural Language Processing or by generating some standard machine learning algorithm. The sentiment is generally given using three numerals – 0 / 1 / -1. The number 0 represents a neutral tweet, number 1 represents a positive tweet and number -1 represents a negative tweet.

Natural Language Processing

This methodology includes the mechanism of analyzing the emotions of the text conveyed by a person to understand if the particular person is happy or sad or tense or is just feeling neutral. It includes the usage of many python libraries such as TextBlob and Tweepy respectively. In general, the process of NLP includes two steps to be followed namely – Data preprocessing and developing an algorithm for analyzing the emotions behind the text. NLP can also be described as a process using which we are trying to enable the understanding ability of human language by the computers or machines which are prevalent everywhere.

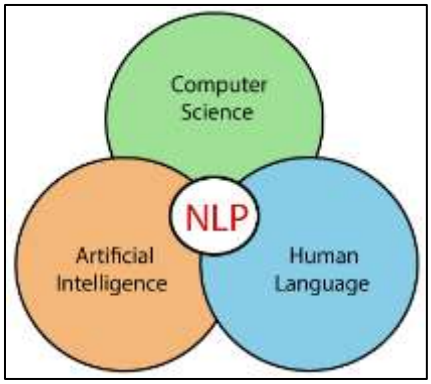


Figure – 1.1 Process of Natural Language Processing

Twitter Application Program Interface

The sentiment analysis process of Twitter is generally not done by using any existing datasets instead, the datasets are gathered from the Twitter app in the form of live tweets. The purpose of gathering love tweets is to get a more realistic analysis of the opinion of the public on different themes all around the world. This is possible only by using the Application Program Interface. To access the live tweets a person is in needs a Developer Twitter account respectively.

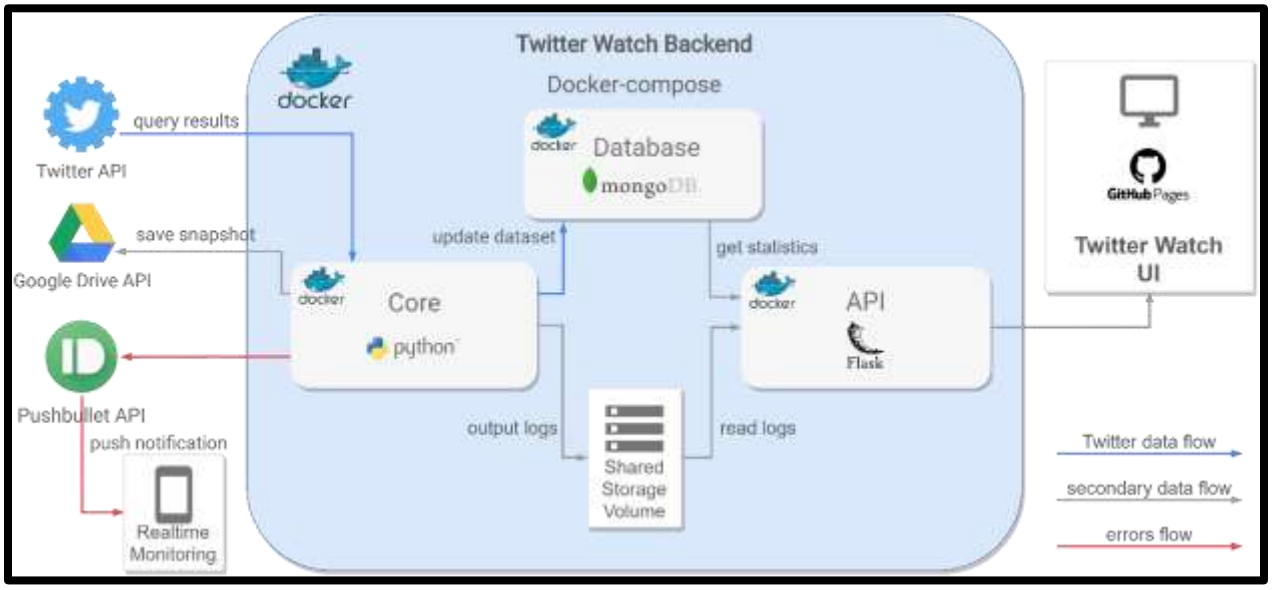


Figure – 1.2 Workflow of data extraction using Twitter API

Analyzing data with Power BI tools

The collected data is then pre-processed after which the final dataset is then analyzed using the software application termed POWER BI. It is an application that can be used to analyze the data available over a particular set of events to learn the insights of the provided data. The analysis can be made over comparisons such as gradual increase or decrease, or percentage of sales over some time

1.2. DATA COLLECTION

The dataset is collected from the Twitter Application by using the Twitter Developer Account which implements the usage of Twitter API. The dataset is collected under five different categories to know the understanding of emotions behind a person's tweet under different domains.

Twitter API

```
# Authentication
consumerKey = 'K4Z1gyHuZLbGhx69bPWGElZd6'
consumerSecret = 'vZNawwg74RaIeSMq819gnDWcpAMLcojdpTO4wxRGCArWO95waV'
accessToken = '1597617657075097601-CEmXV9PHIquXUA7ScjgemPmWjaU6YW'
accessTokenSecret = '2gjUsH3YMTT9kBP5VqHt1lU7njCl22F0mdibaMeH7r3h1'
auth = tweepy.OAuthHandler(consumerKey, consumerSecret)
auth.set_access_token(accessToken, accessTokenSecret)
api = tweepy.API(auth)
```

Figure – 1.3 Key Tokens used for live tweet authorization

Process Of Sentiment Analysis

Every time we would like to gather a set of live tweets, we are supposed to specify the required terms and date, and time of the tweet. This is achieved by using the Advanced Search option available in the Twitter Application. Here, the key term, hashtags, and language, from the date and to the date of the tweets are specified.

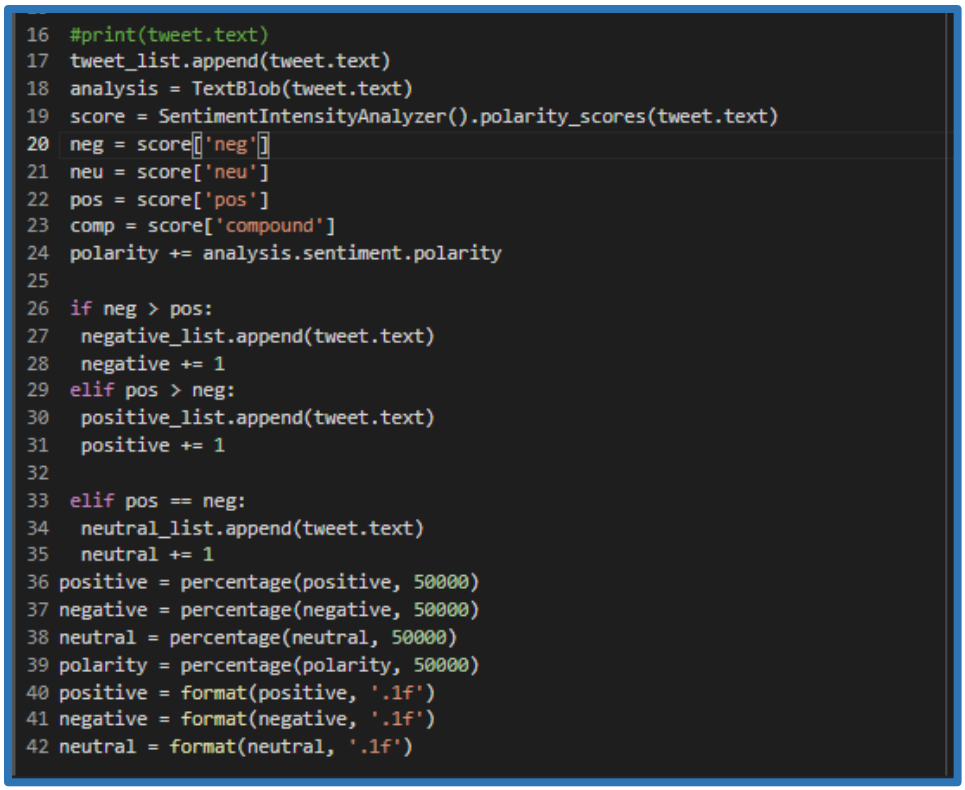
```
#Sentiment Analysis
def percentage(part,whole):
 return 100 * float(part)/float(whole)

tweets = tweepy.Cursor(api.search, q="US Presidential Election (#DonaldTrump) lang:en until:2021-01-20 since:2021-01-19").items(50000)
positive = 0
negative = 0
neutral = 0
polarity = 0
tweet_list = []
neutral_list = []
negative_list = []
positive_list = []
for tweet in tweets:

```

Figure – 1.4 Example query specifying keyword, language, and date of the tweet

The below figure shows the code implemented for the finding of the percentage of positive, negative, and neutral tweets among the total collected tweets.



```
#print(tweet.text)
tweet_list.append(tweet.text)
analysis = TextBlob(tweet.text)
score = SentimentIntensityAnalyzer().polarity_scores(tweet.text)
neg = score['neg']
neu = score['neu']
pos = score['pos']
comp = score['compound']
polarity += analysis.sentiment.polarity

if neg > pos:
 negative_list.append(tweet.text)
 negative += 1
elif pos > neg:
 positive_list.append(tweet.text)
 positive += 1

elif pos == neg:
 neutral_list.append(tweet.text)
 neutral += 1
positive = percentage(positive, 50000)
negative = percentage(negative, 50000)
neutral = percentage(neutral, 50000)
polarity = percentage(polarity, 50000)
positive = format(positive, '.1f')
negative = format(negative, '.1f')
neutral = format(neutral, '.1f')
```

Figure – 1.5 Extracting the sentiment of a sample of 50,000 live tweets over a particular topic

Categories

The five categories included are

- Mobile – sales and popularity
- Western Music – three genres namely pop, jazz and hip-hop
- Covid19 – trend over activities preferred
- Movie – knowing the favoritism of different genres of movies
- FIFA World Cup – comparison of fans over different countries

Description of Dataset

The attributes available in the datasets are as follows

▪ TWEET_COUNT	INT	index of the tweet
▪ TWEET_USERNAME	STR	name of the user account
▪ TWEET_DATE	INT	date of the tweet posted
▪ TWEET_TIME	INT	time of the tweet posted
▪ TWEET_TEXT	LONG	content of the tweet
▪ TWEET_LOCATION	STR	location of the user
▪ TWEET_POLARITY	INT	sentiment of the tweet

The same set of attributes is used for all 5 tables included in the dataset for performing the sentiment analysis of Twitter data.

1.3.PROBLEM STATEMENT

The understanding of any statement gives the lead to know the mindset as well as the emotion of the person that gave the corresponding sentence. The project also focuses on a similar aspect since it is important to analyze the public review on any topics over social media. The range of people available over any platform does not have any limit and has a very wide range. Thus, by knowing the emotions of the opinions placed by a commoner it is possible to make several conclusions such as the popularity of celebrity, important decisive news, reach of any available new gadgets as well.

1.4 BUSINESS OBJECTIVE

1. To analyze public opinion or simply to know the response among people on the newly released product in the market – to know if the product is welcomed with great responses or just some mild opinions or is hated by many.
2. To know which leader has more followers and more support over the netizens compared to other competing leaders over the Election period.
3. To understand which career path shall be more suitable by knowing the public opinion on different gadgets or technologies that are newly generated every day.
4. To know which theme of entertainment or culture is better preferred – can be over any type of celebrities or dance or music.
5. To know the trend of sales over different brands of the same product overseas.

CHAPTER – 2

DATA PREPARATION AND MODELLING

2.1. DATA CLEANING

The final dataset loaded in Power BI must be processed through the following discussed steps. The reason to do so is that the raw dataset may contain many null values or may contain missing values that must be either removed or replaced. All the more the data gathered by using python libraries implementing the Twitter API contains tweets that contain many unwanted unnecessary elements such as ‘emojis’, ‘hashtags’, ‘links to websites’, ‘unidentifiable words’, and many more. All these elements must be removed so as to proceed with the sentiment analysis of the complete text message conveyed in the tweet.

Removing Of Unwanted Elements

For A Single Tweet

The following Python libraries are installed, namely – Transformers and Scipy respectively. The purpose of installing these libraries is to perform the process of cleaning and removing unwanted elements from each of the collected tweets.

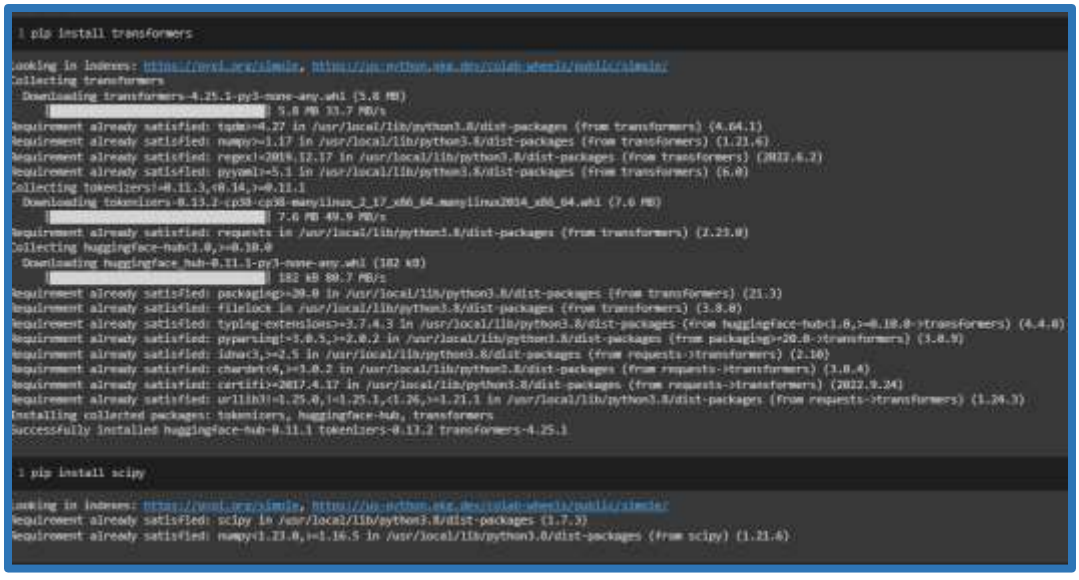


Figure – 2.1 Installation of required python libraries

Here the processing of a single tweet is performed. A sample tweet is taken from Twitter which contains defective elements such as '@', website link, and emoji. All these elements are then removed.

```
from transformers import AutoTokenizer, AutoModelForSequenceClassification
from scipy.special import softmax
```

```
tweet = "@rahshi_21 let's start working @ home 😁 https://www.google.com/"
```

rocess tweet

```
tweet_words = []
for word in tweet.split(' '):
  if word.startswith('@') and len(word) > 1:
    word = '@user'
  elif word.startswith('http'):
    word = "http"
  tweet_words.append(word)
```

```
print(tweet_words)
```

```
['@user', "let's", 'start', 'working', '@', 'home', '😁', 'http']
```

```
tweet_processed = " ".join(tweet_words)
```

```
print(tweet_processed)
```

```
@user let's start working @ home 😁 http
```

```
roberta = "cardiffnlp/twitter-roberta-base-sentiment"
```

```
model = AutoModelForSequenceClassification.from_pretrained(roberta)
tokenizer = AutoTokenizer.from_pretrained(roberta)
```

```
Downloading: 100% 747/747 [00:00<00:00, 8.34kB/s]
Downloading: 100% 499M/499M [00:13<00:00, 66.0MB/s]
Downloading: 100% 899k/899k [00:00<00:00, 1.45MB/s]
Downloading: 100% 456k/456k [00:00<00:00, 1.40MB/s]
Downloading: 100% 150/150 [00:00<00:00, 4.46kB/s]
```

Figure – 2.2 Sentiment analysis for a sample tweet

The final score of the single sample tweet is categorized as

```
for i in range(len(scores)):
  l = labels[i]
  s = scores[i]
  print(l,s)
```

```
Negative 0.0025944852
Neutral 0.24713646
Positive 0.750269
```

Figure – 2.3 Final score of the tweet (after preprocessing)

For a Large Number of Live Tweets

The preprocessing of a large number of tweets is explained here by implementing Python code and the mechanism of Natural Language Processing.

```
[ ] tweet_list.drop_duplicates(inplace = True)
```

```
[ ] #Cleaning Text (RT, Punctuation etc)
    #Creating new dataframe and new features
    tw_list = pd.DataFrame(tweet_list)
    tw_list['text'] = tw_list[0]
    #Removing RT, Punctuation etc
    remove_rt = lambda x: re.sub('RT @\w+: ',' ',x)
    rt = lambda x: re.sub('(@[A-Za-z0-9]+)|([0-9A-Za-z \t])|(\w+:\/\/\S+)',' ',x)
    tw_list['text'] = tw_list.text.map(remove_rt).map(rt)
    tw_list['text'] = tw_list.text.str.lower()
    tw_list.head(10)
```

	0	text
0	RT @nojumper: #DonaldTrump calls #JoeBiden's p...	# # ' ...
1	RT @themagacancer: Trump Bashes Jewish Leaders...	' ...
2	RT @foxnewsradio: In a #TruthSocial post last ...	# , ...
4	@johnmccain09 @Snarfaleptic @FoxesWi @SmellyCa...	...
18	@johnmccain09 @Snarfaleptic @FoxesWi @SmellyCa...	' ...
21	Random Trump https://t.co/Xgtzda8SGY - #random...	:// . / - # ...
22	December 08, 2018 US President Donald Trump an...	, ...
24	RT @DaAnsahonSports: 6/22/94 @HoustonRockets 1...	/ / 💍 - . # ...
43	RT @carolmswain: I have never seen a man so ha...	...
49	Random Trump https://t.co/1urWrgDF9j - #random...	:// . / - # ...

Figure – 2.4 Text before pre-processing

After the completion of pre-processing the polarity, subjectivity, and sentiment of the tweet is extracted which is exported and saved as a Comma Separated Value File that can then be imported into a Power BI file for performing the analysis of the formed Dataset.

```
#Calculating Negative, Positive, Neutral and Compound values
tw_list[['polarity', 'subjectivity']] = tw_list['text'].apply(lambda Text: pd.Series(TextBlob(Text).sentiment))
for index, row in tw_list['text'].iteritems():
 score = SentimentIntensityAnalyzer().polarity_scores(row)
 neg = score['neg']
 neu = score['neu']
 pos = score['pos']
 comp = score['compound']
 if neg > pos:
  tw_list.loc[index, 'sentiment'] = 'negative'
 elif pos > neg:
  tw_list.loc[index, 'sentiment'] = 'positive'
 else:
  tw_list.loc[index, 'sentiment'] = 'neutral'
  tw_list.loc[index, 'neg'] = neg
  tw_list.loc[index, 'neu'] = neu
  tw_list.loc[index, 'pos'] = pos
  tw_list.loc[index, 'compound'] = comp
tw_list.head(10)
```

	0	text	polarity	subjectivity	sentiment	neg	neu	pos	compound
0	RT @nojumper: #DonaldTrump calls #JoeBiden's p...	# # ' ...	0.0	0.0	neutral	0.0	1.0	0.0	0.0
1	RT @themagacancer: Trump Bashes Jewish Leaders...	' ...	0.0	0.0	neutral	0.0	1.0	0.0	0.0
2	RT @foxnewsradio: In a #TruthSocial post last ...	# , ...	0.0	0.0	neutral	0.0	0.0	0.0	0.0
4	@johnmccain09 @Snarfaleptic @FoxesWi @SmellyCa...	...	0.0	0.0	neutral	0.0	1.0	0.0	0.0
18	@johnmccain09 @Snarfaleptic @FoxesWi @SmellyCa...	' ...	0.0	0.0	neutral	0.0	1.0	0.0	0.0
21	Random Trump https://t.co/Xgtzda8SGY - #random...	:// . / - # ...	0.0	0.0	neutral	0.0	1.0	0.0	0.0
22	December 08, 2018 US President Donald Trump an...	, ...	0.0	0.0	neutral	0.0	1.0	0.0	0.0
24	RT @DaAnsahonSports: 6/22/94 @HoustonRockets 1...	/ / ♁ - . # ...	0.0	0.0	neutral	0.0	0.0	0.0	0.0
43	RT @carolmswain: I have never seen a man so ha...	...	0.0	0.0	neutral	0.0	0.0	0.0	0.0
49	Random Trump https://t.co/1urWrgDF9j - #random...	:// . / - # ...	0.0	0.0	neutral	0.0	1.0	0.0	0.0

Figure – 2.5 Text after pre-processing and polarity generation

2.2 DATA TRANSFORMING

The collected data is now transformed accordingly based on some features. The values of each attribute in the dataset may not belong to the correct domain and may vary from the actual domain. Similarly, the columns had to split at times or merged at other times according to the needs of the insights performed. Also, the insights can be better explained using the DAX measures and calculated column function provided in the 'Transform Data Tool'. This also involves the implementation of the three steps – Extract, Transform and Load respectively.

Steps in ETL (Extract, Transform, Load)

1. **In the Power BI desktop, the dataset is loaded.**

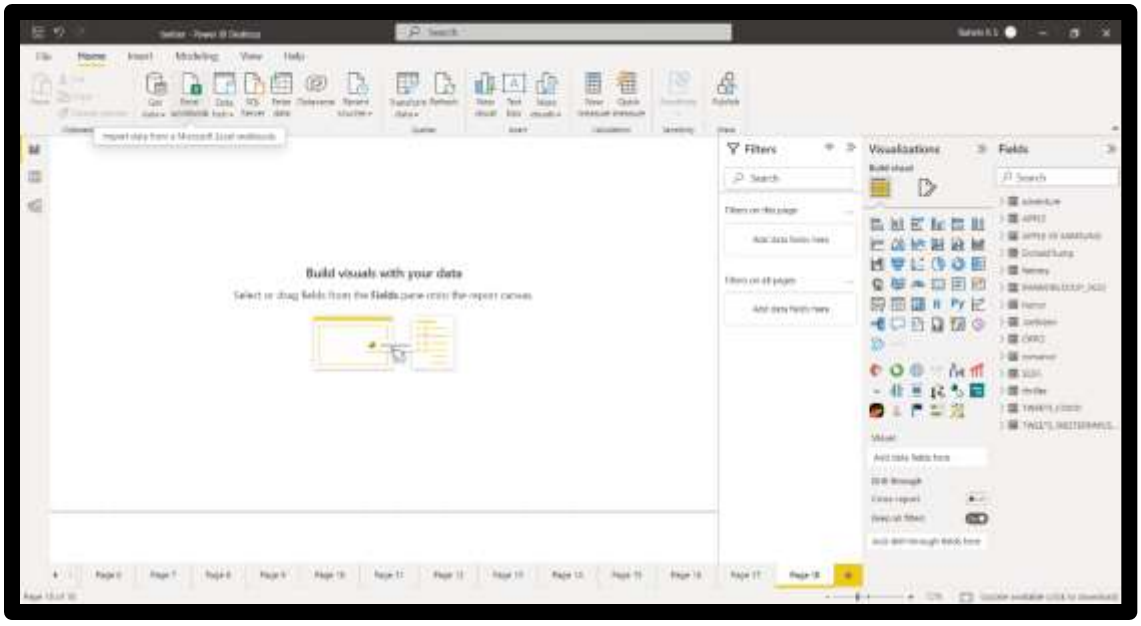


Figure – 2.6 The dataset is loaded in Power BI

2. **To transform the data, click on the Transform Data icon present in the home tab.**

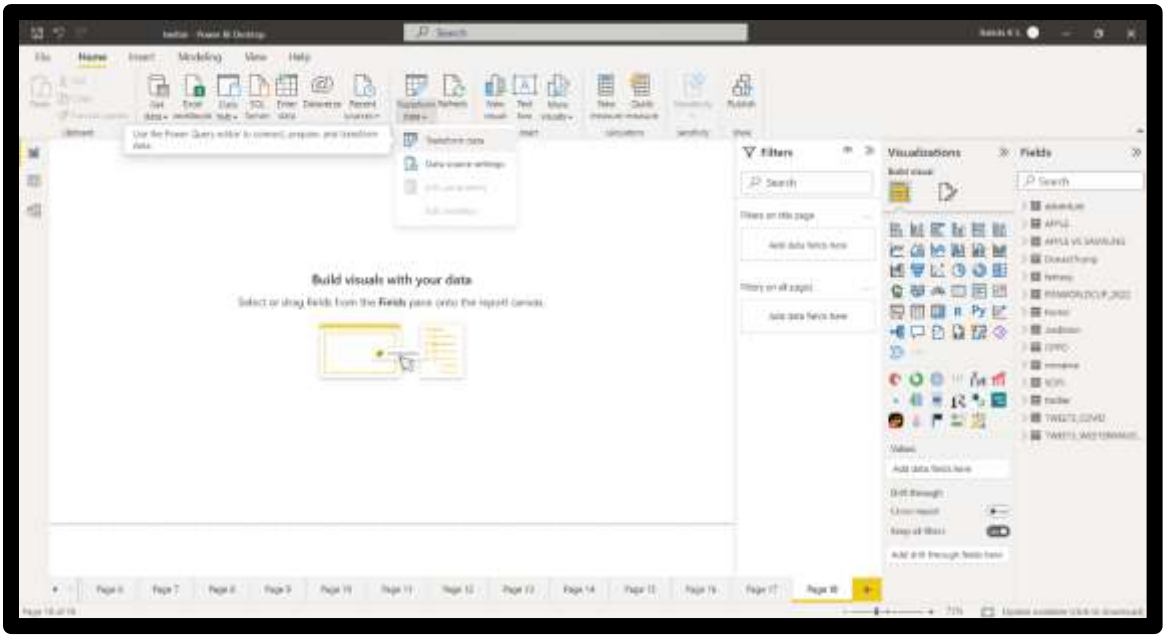


Figure - 2.7 Opening Power Query Editor

3. **In the table named ‘OPPO’, the column with date and time is split into two columns using a column splitter**

   = Table.SplitColumn(Table.TransformColumnTypes(#"Changed Type", {{"Date", type text}}, "en-US"), "Date", Splitter.SplitTextByPositions({0, 10}, false), {"Date.1", "Date.2"})

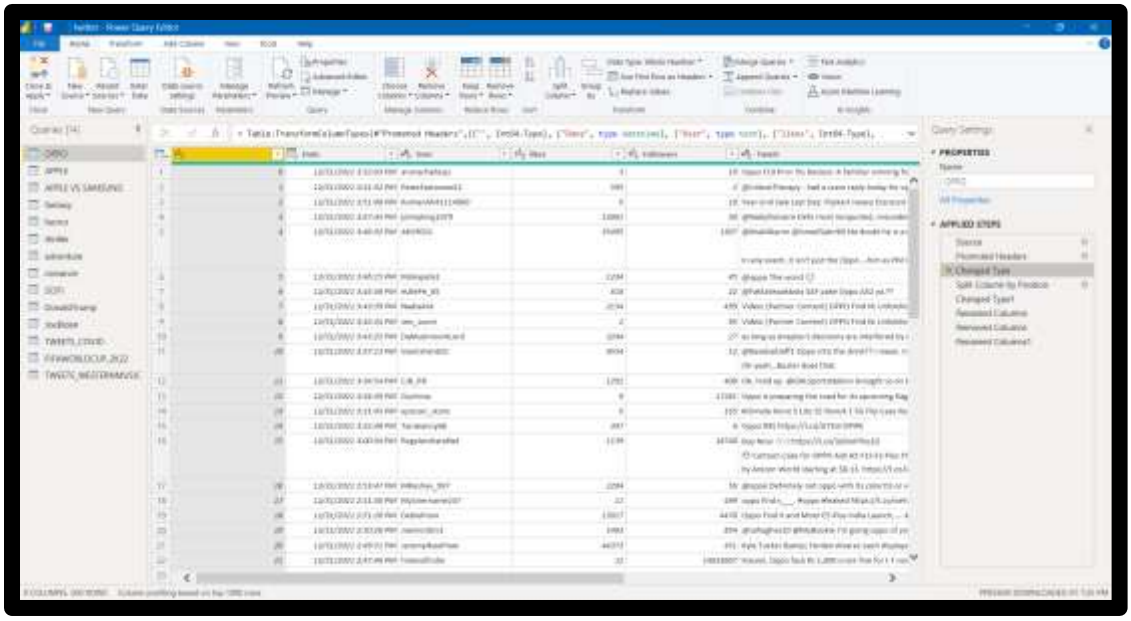


Figure - 2.8 Splitting the Date and Time column

4. **In all the tables the column names are similar. Thus, in table ‘OPPO’ the columns are renamed accordingly**

   = Table.RenameColumns(#"Removed Columns",{{"", "OPPO_SNO"}, {"Date", "OPPO_DATE"}, {"Time", "OPPO_TIME"}, {"User", "OPPO_USERID"}, {"likes", "OPPO_TWEET_LIKES"}, {"Followers", "OPPO_TWEET_USERFOLLOWERS"}, {"Tweet", "OPPO_TWEET"}, {"Location", "OPPO_TWEET_LOCATION"}, {"Polarity", "OPPO_TWEET_POLARITY"}})

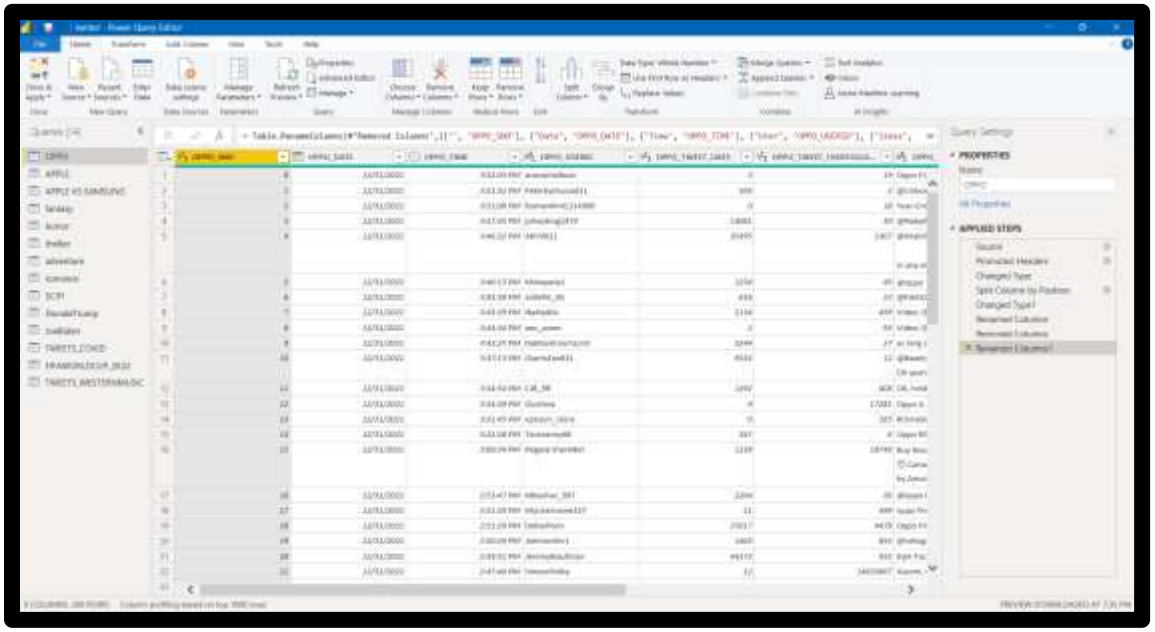


Figure - 2.9 Renaming the columns of the ‘OPPO’ table

5. **In all the tables the column names are similar. Thus in table ‘APPLE’ the columns are renamed accordingly**

   = Table.RenameColumns(#"Changed Type1",{{"Date.1", "APPLE_DATE_OF_TWEET"}, {"Date.2", "APPLE_TIME OF TWEET"}, {"sno", "APPLE_sno"}, {"User", "APPLE_USERNAME_OF TWEET"}, {"likes", "APPLE_LIKES_OF_TWEET"}, {"Followers", "APPLE_FOLLOWERS_OF USERNAME"}, {"Tweet", "APPLE_TWEET_TEXT"}})

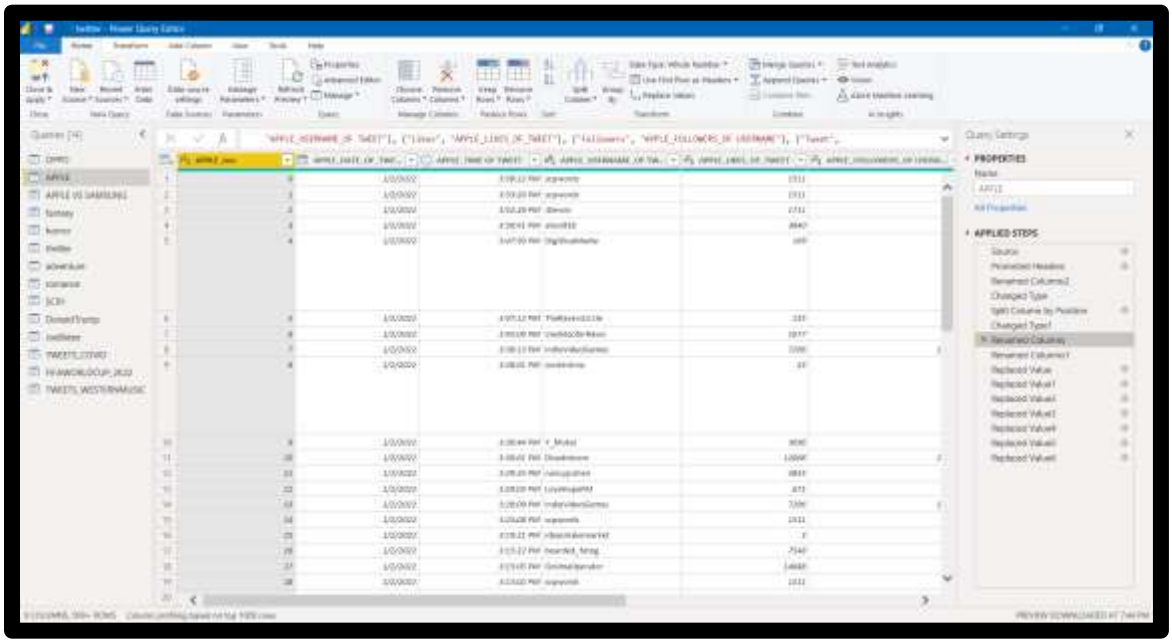


Figure - 2.10 Renaming the columns of the ‘APPLE’ table

6. **In all the tables the column names are similar. Thus, in table ‘TWEET_COVID’ the columns are renamed accordingly**

   = Table.RenameColumns(#"Changed Type1",{{"Date.1", "DATE_TWEET[COVID]"}, {"Date.2", "TIME_TWEET[COVID]"}, {"User", "TWEET_USERID[COVID]"}, {"Likes", "TWEET_LIKES[COVID]"}, {"Tweet", "TWEET[COVID]"}, {"Location", "TWEET_LOCATION[COVID]"}, {"Activity", "ACTIVITY[COVID]"}, {"Polarity", "POLARITY_TWEET[COVID]"}})

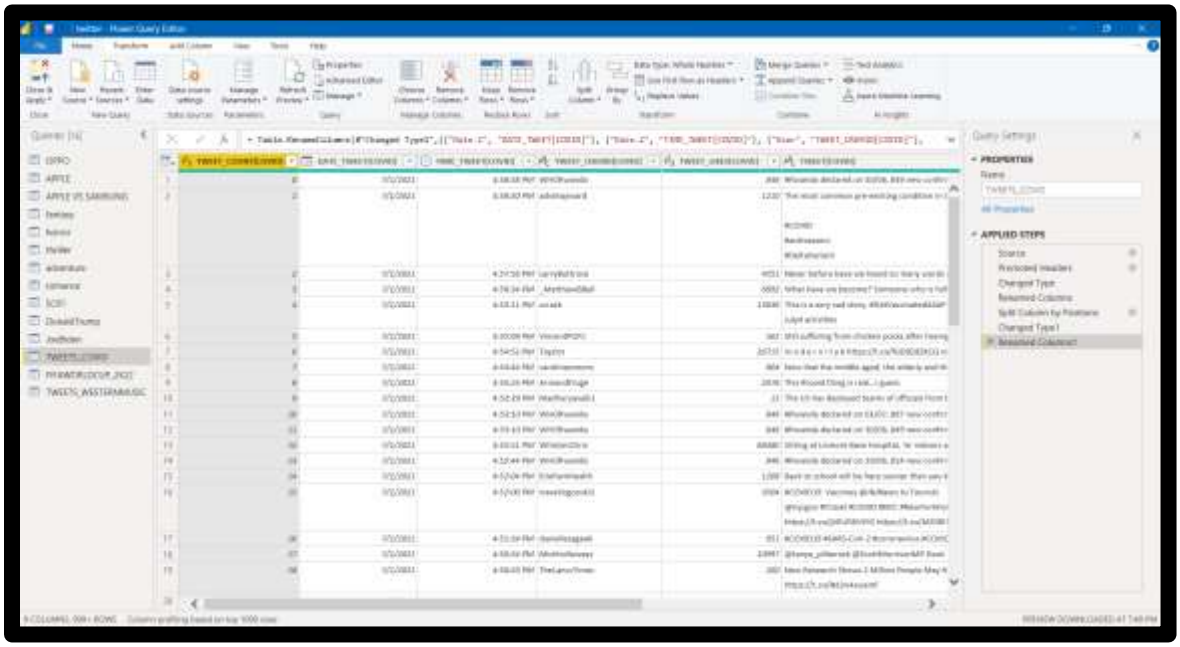


Figure - 2.11 Renaming the columns of the ‘COVID’ table

7. **To create a DAX measure to find the total average movie viewers for the fantasy genre.**

   DAX :=adventure_avg_views=CALCULATE(AVERAGE(adventure[ADVENTURE]))

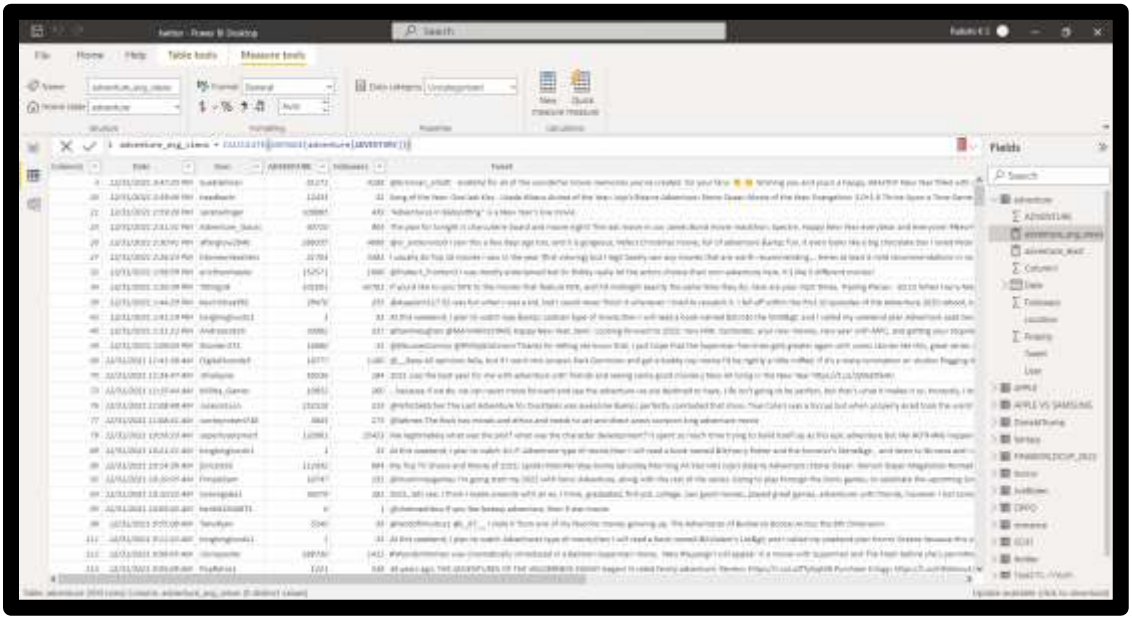


Figure – 2.12 DAX Measure to find the average of likes on the fantasy genre

8. **To create a DAX measure to find the total number of likes received by the tweets quoted on the domain of APPLE brand phones.**

DAX = APPLE_SUM_OF_LIKES =

CALCULATE(SUM(APPLE[APPLE_LIKES_OF_TWEET]))

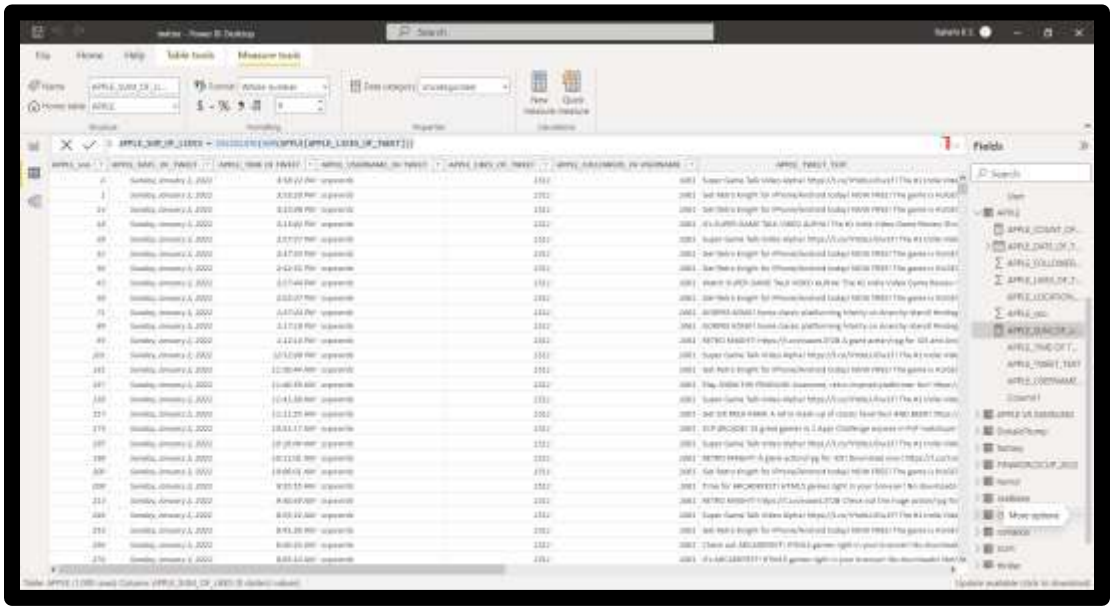


Figure - 2.13 DAX Measure to calculate total likes received for the ‘APPLE’ brand

9. **To calculate the total likes received for all the tweets posted for each location that is chosen to be compared for the sales of APPLE and SAMSUNG.**

BRAZIL_APPLE =

CALCULATE(SUM('APPLE VS SAMSUNG'[likes_apple]),FILTER('APPLE VS SAMSUNG','APPLE VS SAMSUNG'[Location] = "BRAZIL"))

BRAZIL_SAMSUNG =

CALCULATE(SUM('APPLE VS SAMSUNG'[likes_samsung]),FILTER('APPLE VS SAMSUNG','APPLE VS SAMSUNG'[Location] = "BRAZIL"))

```
GERMANY_APPLE =

CALCULATE(SUM('APPLE VS SAMSUNG'[likes_apple]),FILTER('APPLE VS SAMSUNG','APPLE VS SAMSUNG'[Location] = "GERMANY"))

GERMANY_SAMSUNG =

CALCULATE(SUM('APPLE VS SAMSUNG'[likes_samsung]),FILTER('APPLE VS SAMSUNG','APPLE VS SAMSUNG'[Location] = "GERMANY"))

SKOREA_APPLE =

CALCULATE(SUM('APPLE VS SAMSUNG'[likes_apple]),FILTER('APPLE VS SAMSUNG','APPLE VS SAMSUNG'[Location] = "SOUTH KOREA"))

SKOREA_SAMSUNG =

CALCULATE(SUM('APPLE VS SAMSUNG'[likes_samsung]),FILTER('APPLE VS SAMSUNG','APPLE VS SAMSUNG'[Location] = "SOUTH KOREA"))

UK_APPLE =

CALCULATE(SUM('APPLE VS SAMSUNG'[likes_apple]),FILTER('APPLE VS SAMSUNG','APPLE VS SAMSUNG'[Location] = "UK"))

UK_SAMSUNG =

CALCULATE(SUM('APPLE VS SAMSUNG'[likes_samsung]),FILTER('APPLE VS SAMSUNG','APPLE VS SAMSUNG'[Location] = "UK"))

USA_APPLE =

CALCULATE(SUM('APPLE VS SAMSUNG'[likes_apple]),FILTER('APPLE VS SAMSUNG','APPLE VS SAMSUNG'[Location] = "USA"))

USA_SAMSUNG =

CALCULATE(SUM('APPLE VS SAMSUNG'[likes_samsung]),FILTER('APPLE VS SAMSUNG','APPLE VS SAMSUNG'[Location] = "USA"))
```

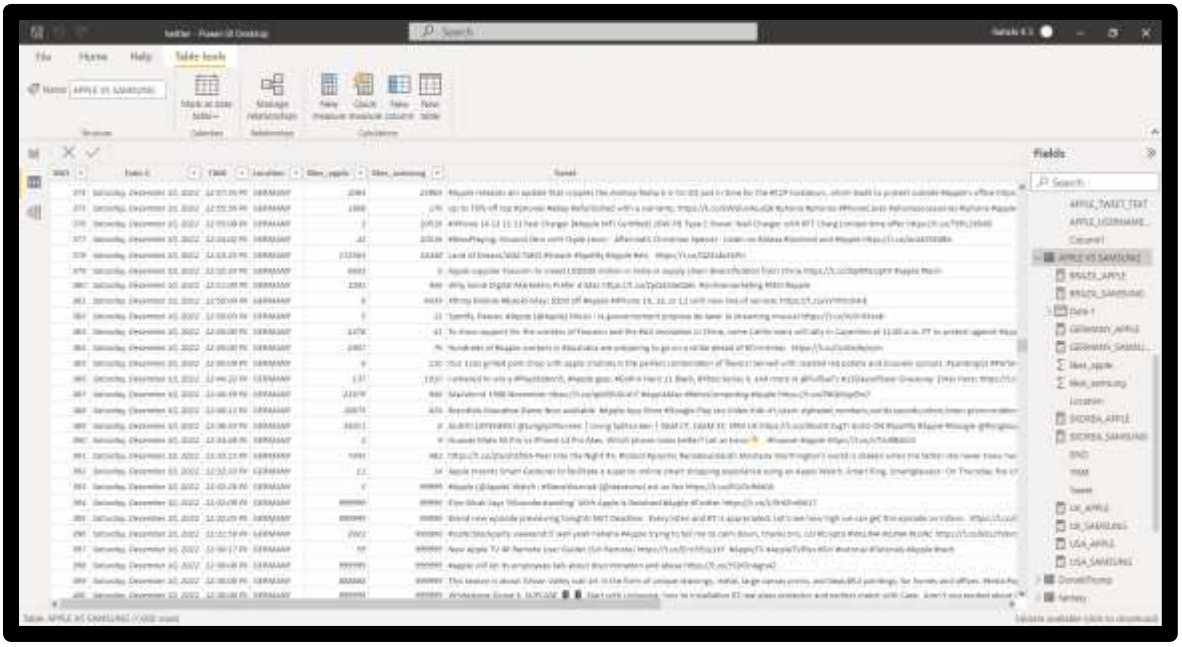


Figure 2.14 - Using various DAX measures for different locations to compare sales

10. To calculate the count of tweets based on each type of activity included in the COVID table.

BOARDGAMES_COUNT

=CALCULATE(COUNT('TWEETS_COVID'[TWEET_COUNT[COVID]]]),FILTER('TWEETS_COVID','TWEETS_COVID'[ACTIVITY[COVID]]] = "BoardGames"))

EXERCISE_COUNT

=CALCULATE(COUNT('TWEETS_COVID'[TWEET_COUNT[COVID]]]),FILTER('TWEETS_COVID','TWEETS_COVID'[ACTIVITY[COVID]]] = "Exercise"))

FACETIME_COUNT

=CALCULATE(COUNT('TWEETS_COVID'[TWEET_COUNT[COVID]]]),FILTER('TWEETS_COVID','TWEETS_COVID'[ACTIVITY[COVID]]] = "Facetime"))

HOUSEHOLD_COUNT
=CALCULATE(COUNT('TWEETS_COVID'[TWEET_COUNT[COVID]]]),FILTER('TWEETS_COVID','TWEETS_COVID'[ACTIVITY[COVID]]] = "Household chores"))

OLDHOBBY_COUNT
=CALCULATE(COUNT('TWEETS_COVID'[TWEET_COUNT[COVID]]]),FILTER('TWEETS_COVID','TWEETS_COVID'[ACTIVITY[COVID]]] = "OldHobbies"))

OTT_COUNT
=CALCULATE(COUNT('TWEETS_COVID'[TWEET_COUNT[COVID]]]),FILTER('TWEETS_COVID','TWEETS_COVID'[ACTIVITY[COVID]]] = "#OTT"))

VIDEO_COUNT
=CALCULATE(COUNT('TWEETS_COVID'[TWEET_COUNT[COVID]]]),FILTER('TWEETS_COVID','TWEETS_COVID'[ACTIVITY[COVID]]] = "Videogames"))

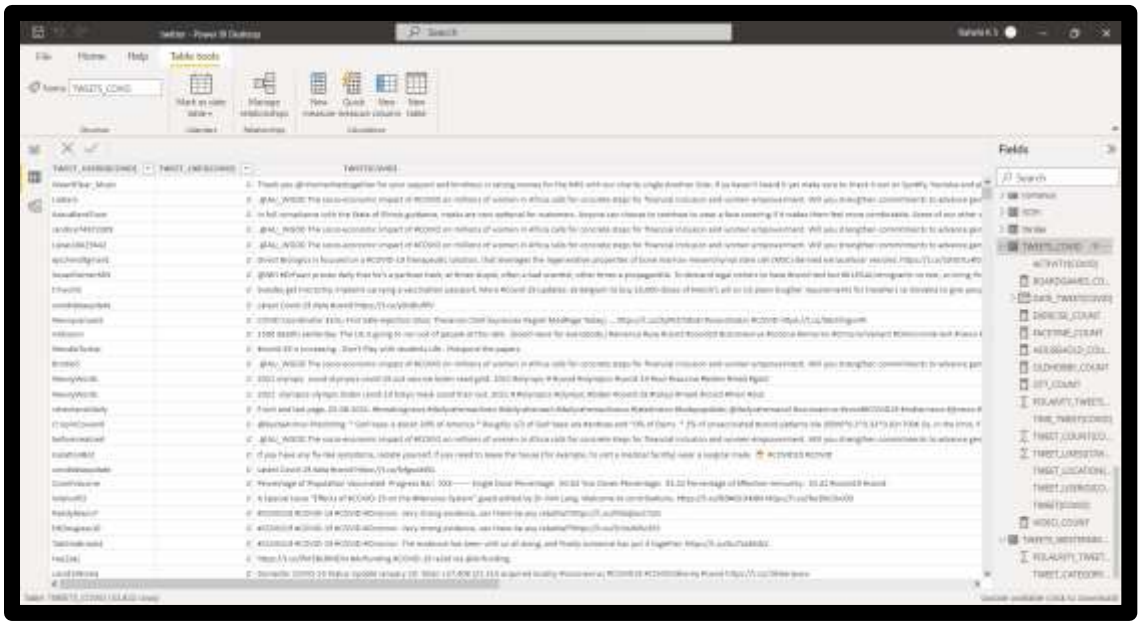


Figure - 2.15 Using various DAX measures for a different activity to compute likes

11. **After completing all the necessary transformations click on the 'Close & Apply' icon in the top left corner of the Power Query Editor to save all the changes performed.**

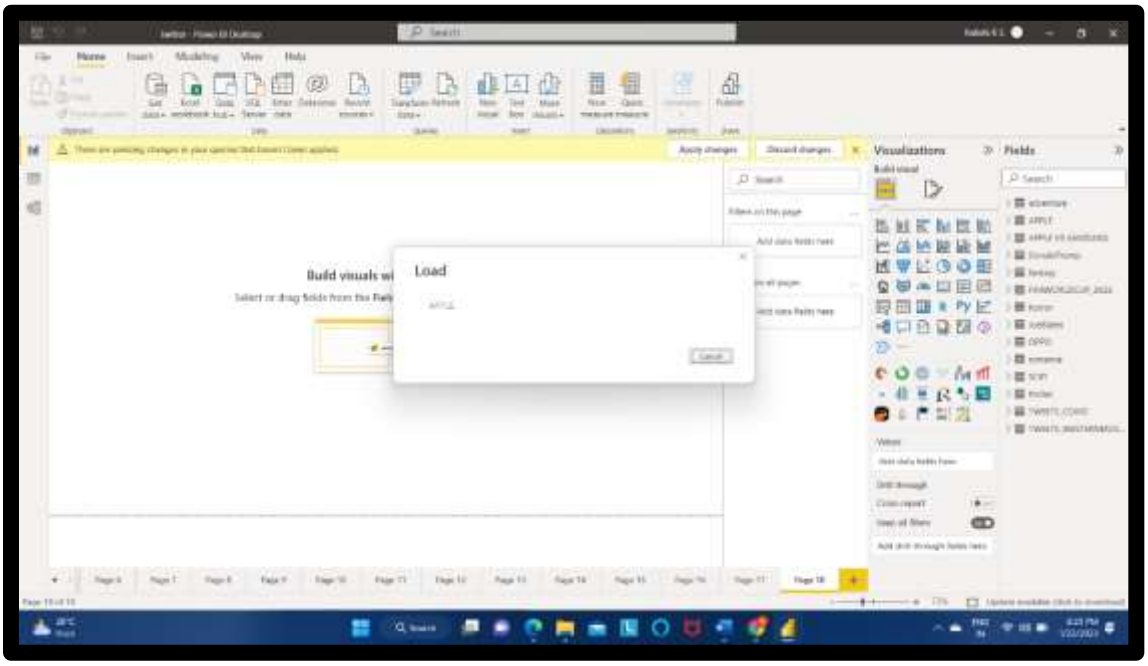


Figure - 2.16 Final loading of the dataset

2.3 DATA MODELING

Data modeling is the process in which the relationships present between different tables included in the dataset are represented visually to get a better understanding of which attributes are related to each other in which manner or in what kind of ways are analyzed in this section of Power BI.

Figure - 2.17 Relationships are visualized using Data Model

The relationships found in the above data model are

1) A total of five “One to One” relationships were formed between the following Tables:
    - ‘ROMANCE’ and ‘FANTASY’
    - ‘FANTASY’ and ‘SCIFI’
    - ‘ADVENTURE’ and ‘FANTASY’
    - ‘FANTASY’ and ‘THRILLER’
    - ‘FANTASY’ and ‘HORROR’ respectively.

The common attributes that lead to the formation of the above relationships are ‘TWEET_COUNT’ for the first two relations and ‘TWEET_LIKES’ for the succeeding three relations.

2) A “One to Many” relationship formed between the Tables – THRILLER’ and ‘ADVENTURE’ based on the common attribute namely – ‘TWEET_POLARITY’ and ‘LOCATION’ respectively.
3) A “Many to One” relationship formed between the Tables – ‘APPLE’ and ‘APPLE_VS_SAMSUNG’ based on the common attributes namely – ‘TOTAL_LIKES’ and ‘LOCATION’ respectively.
4) A “Many to Many” relationship formed between the Tables – ‘TWEETS_COVID’ and ‘TWEETS_WESTERNMUSIC’ based on the common attribute namely – ‘TWEET_CATEGORY” respectively.

CHAPTER – 3

DATA ANALYSIS AND INTERPRETATION

3.1. DATA ANALYSIS

The project is explained under five categories based on which different datasets have been collected to perform the data analysis of the collected data which is nothing but the extracted and pre-processed live tweets. The arrow marks present beside each topic take you to the requested page accordingly.

An Overview of the Content Page

Figure – 3.1 Content page of the Powe BI file

Westernized music – Pop, Jazz, Hip-Hop

1. **Which type of westernized music is most popular?**

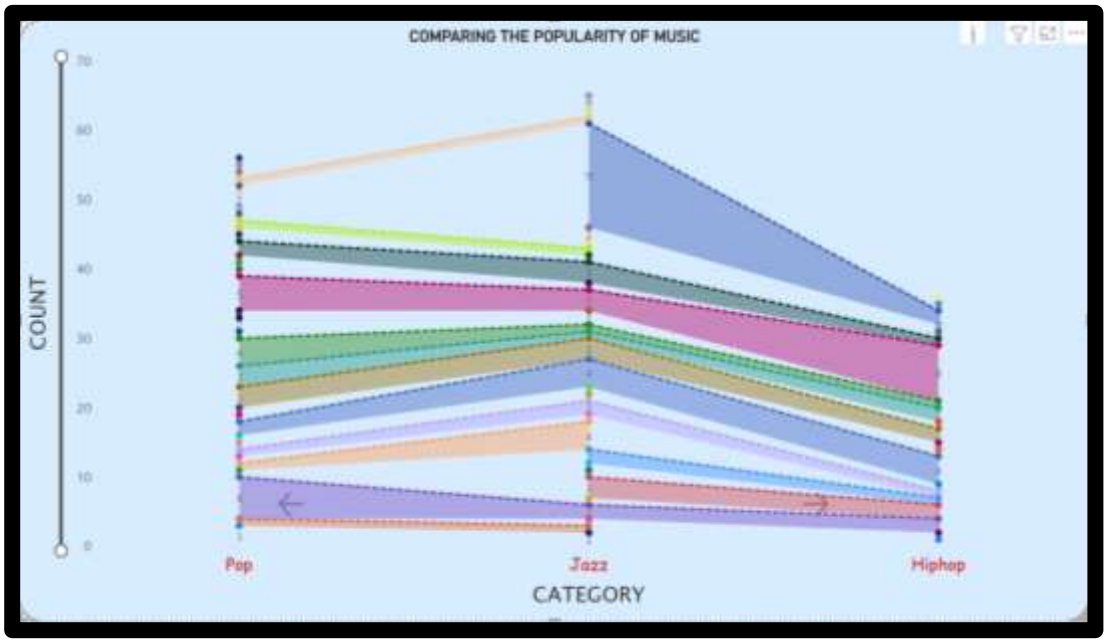


Figure – 3.2 Comparing the popularity of music

2. **To compare and analyze people's opinions on all three categories of music using a slicer.**

Here, the slicer is set to the category called 'Jazz' respectively.

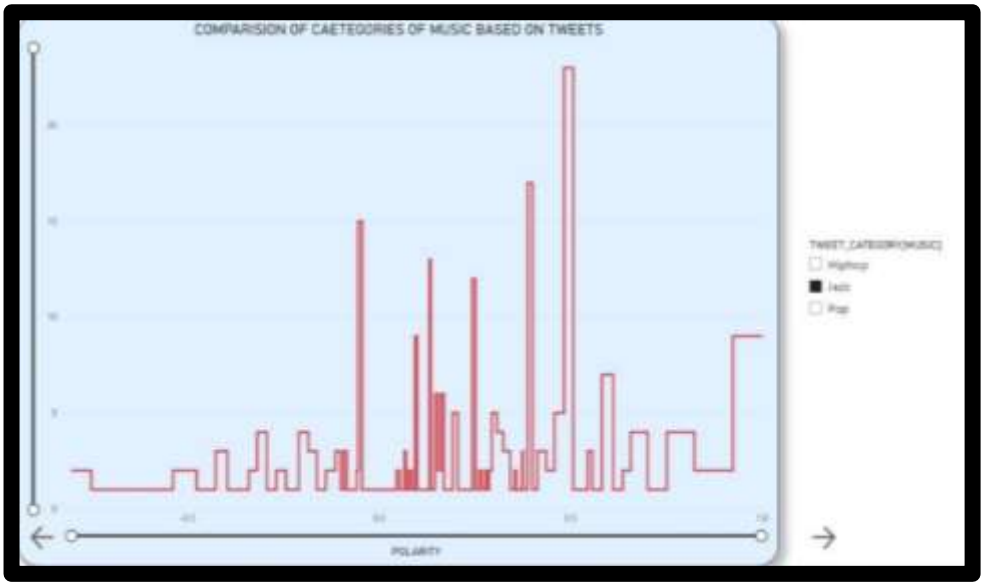


Figure – 3.3 Jazz music polarity variation

Here, the visualization called Slicer is used. In figure 3.4 the slicer is set to the category called ‘Hip-hop’ respectively.

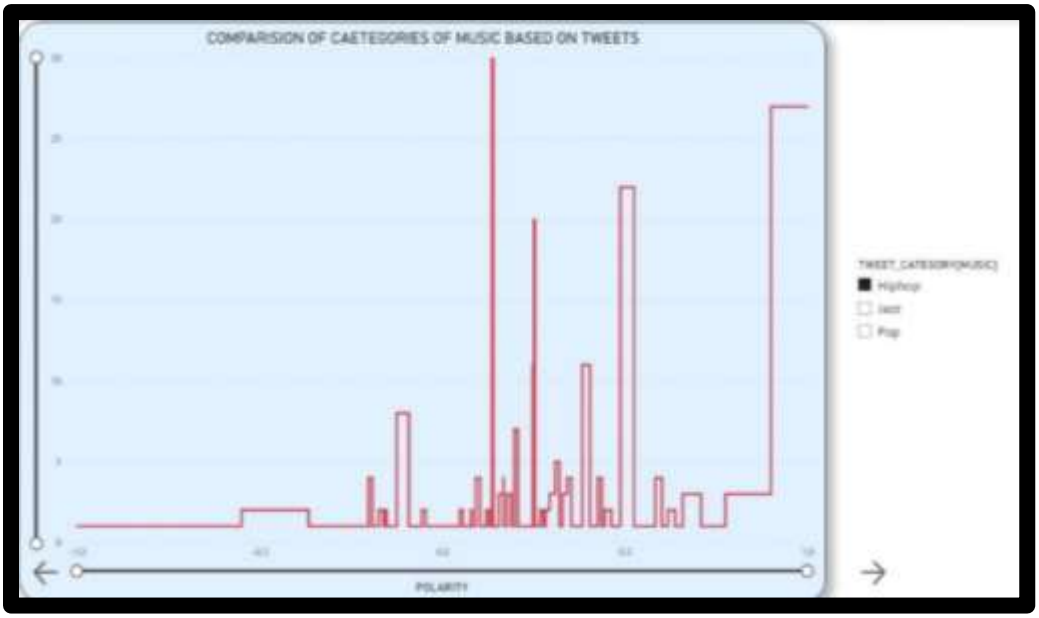


Figure – 3.4 Hip-hop music polarity variation

Here, the visualization called Slicer is used. In figure 3.5 the slicer is set to the category called ‘Pop’ respectively.

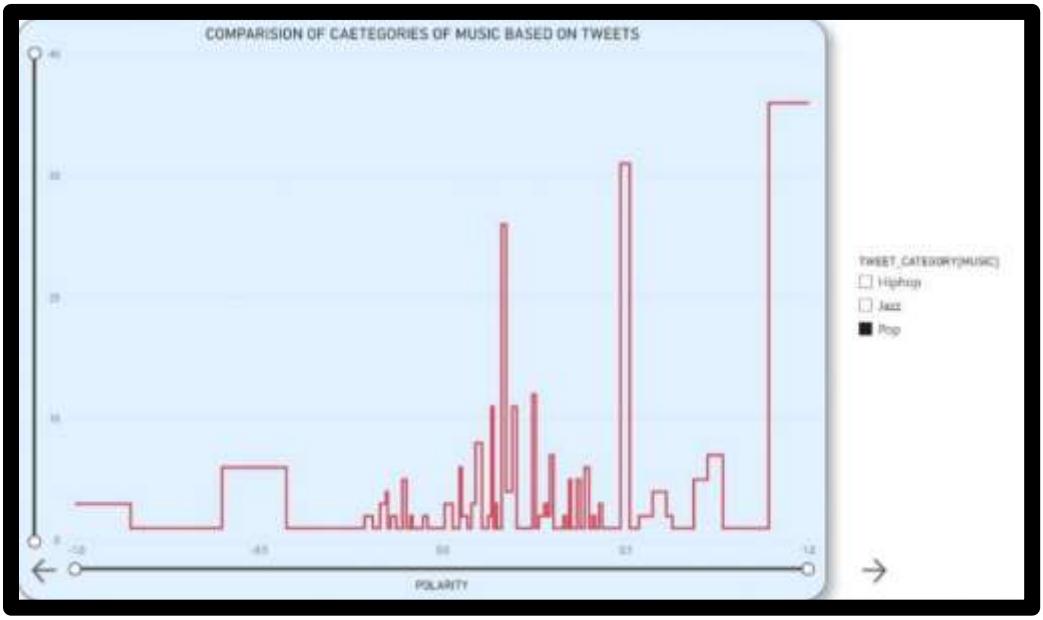


Figure – 3.5 Pop music polarity variation

FIFA World Cup 2022

1. **Depiction of countries involved in the FIFA match.**

Figure – 3.6 Top FIFA viewer's country highlighted in the World Map

2. **Distribution of fans over different countries.**

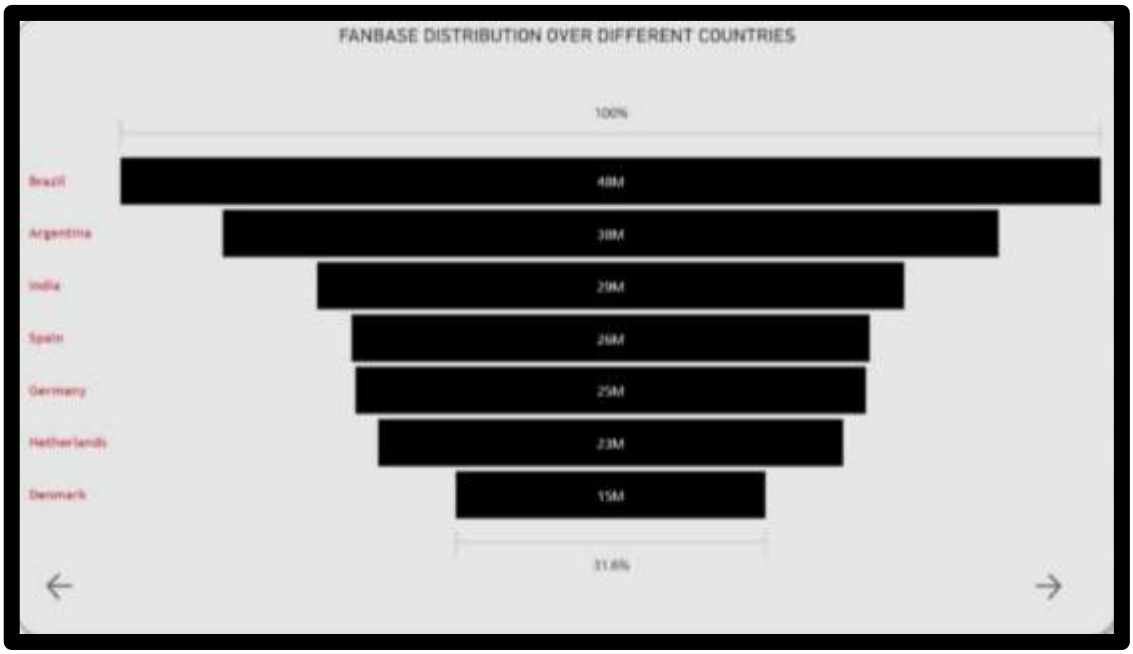


Figure – 3.7 Depiction of fanbase distribution over six different countries

3. **Which country tweets were comprised of the maximum count of positive tweets?**

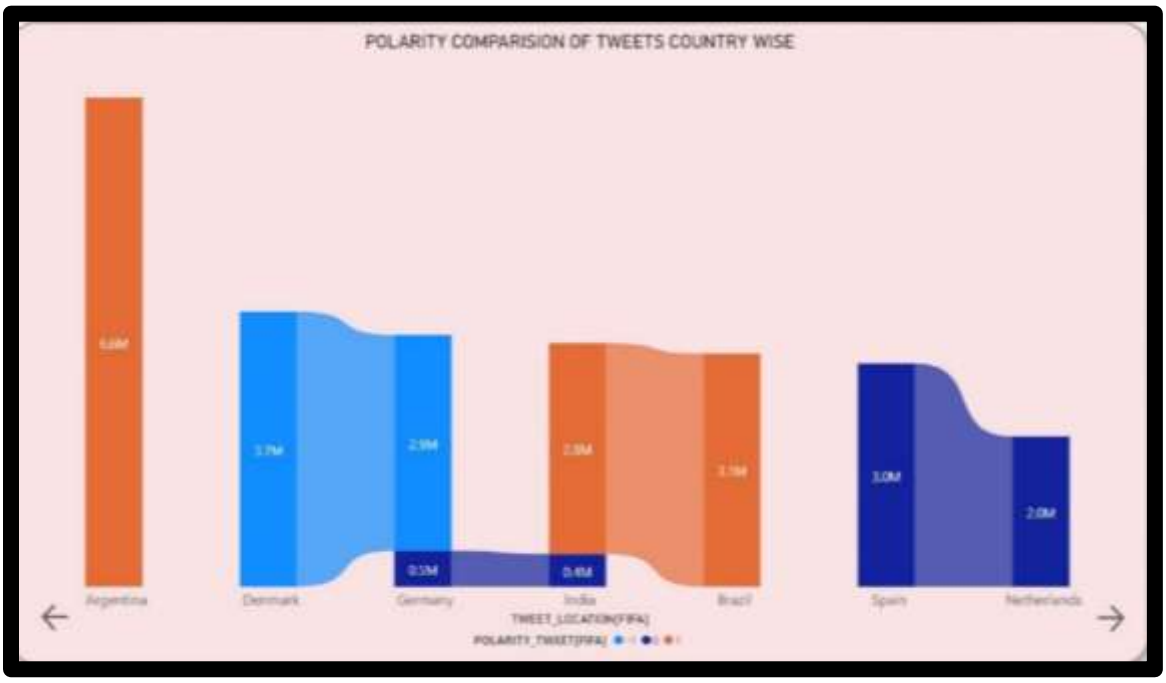


Figure – 3.8 Comparison of the polarity of tweets for different countries

Covid-19 and activities

1. **Analyze the significance of different kinds of activities preferred at covid times on all three levels of polarity.**

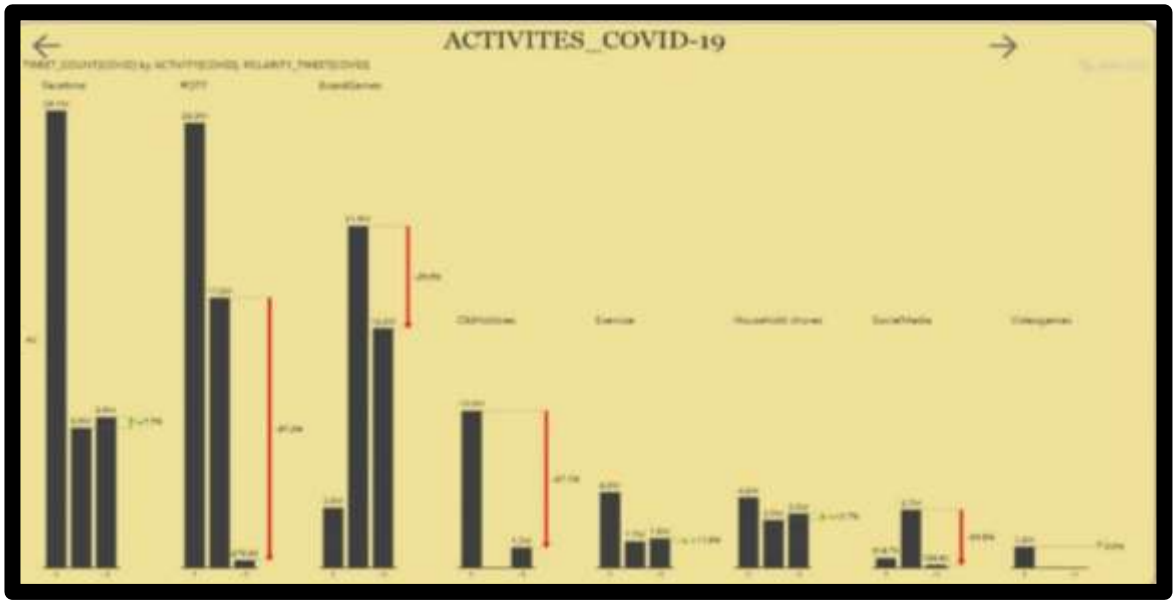


Figure – 3.9 Different activities that were preferred during Covid-19

2. **To compare and find which of the 8 activities has received maximum and minimum polarity of positive tweets.**

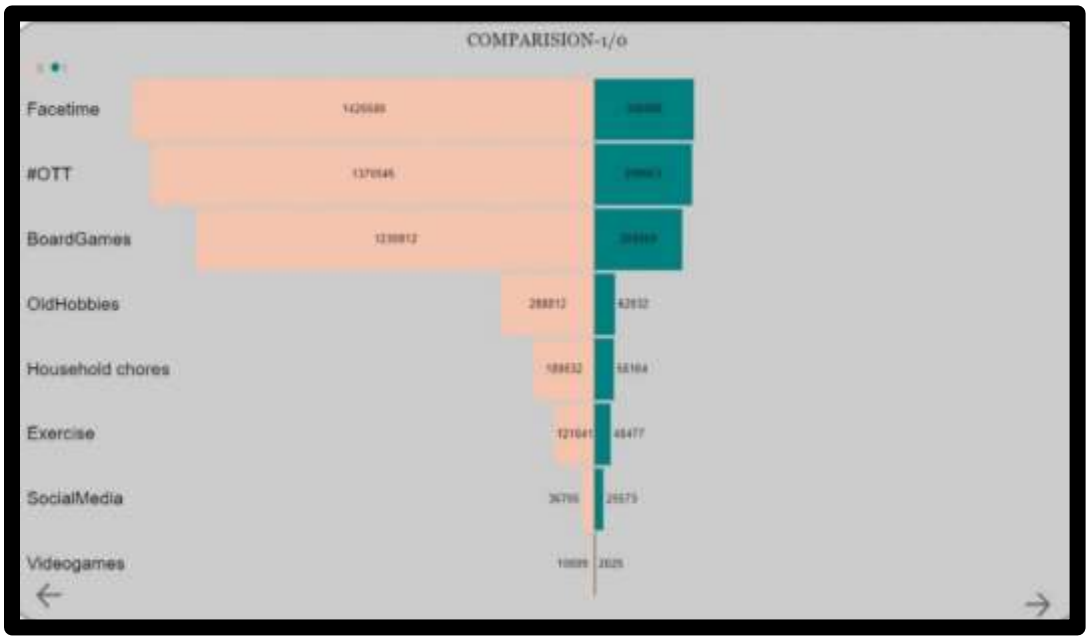


Figure – 3.10 Polarity comparison for different activities in Covid-19

3. **Which activity is most preferred during Covid 2020?**

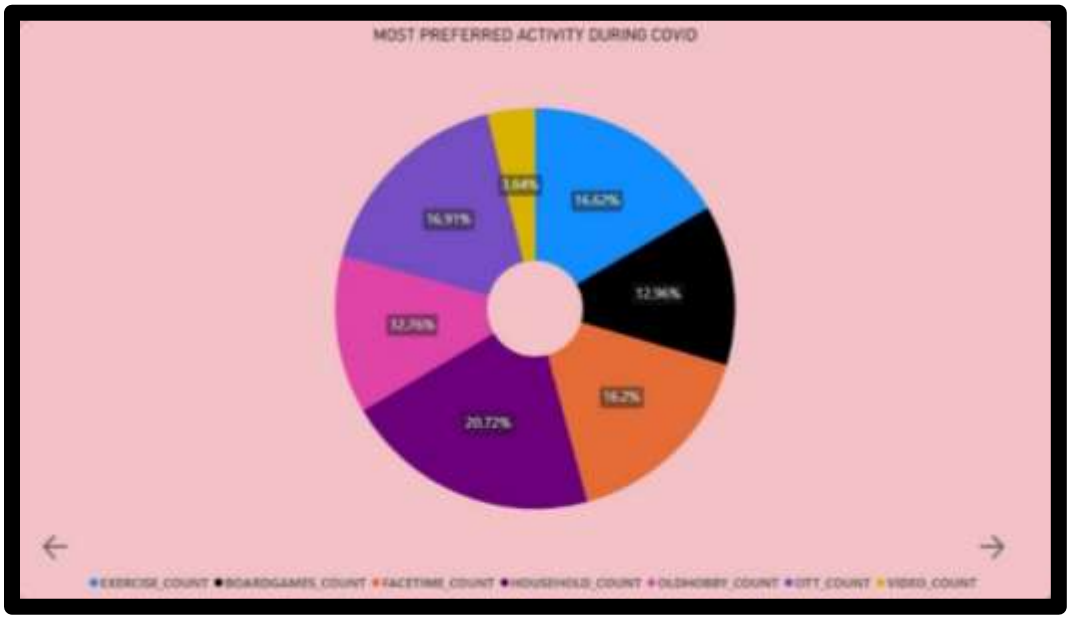


Figure – 3.11 Most preferred activity during Covid-19

Movies in 6 genres – Adventure, Fantasy, Horror, Sci-fi, Thriller, Romance

1. **Which genre of movie is most liked or favored by netizens?**

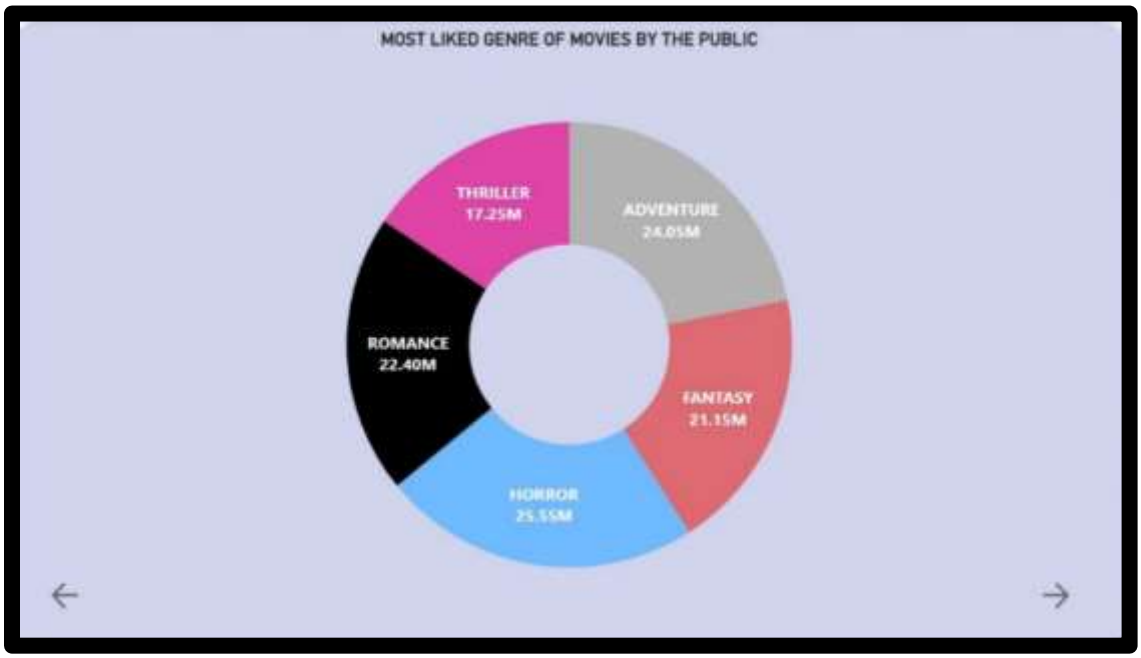


Figure – 3.12 Most liked genre of movie

2. **Which country has the most viewers that prefer sci-fi-themed movies?**

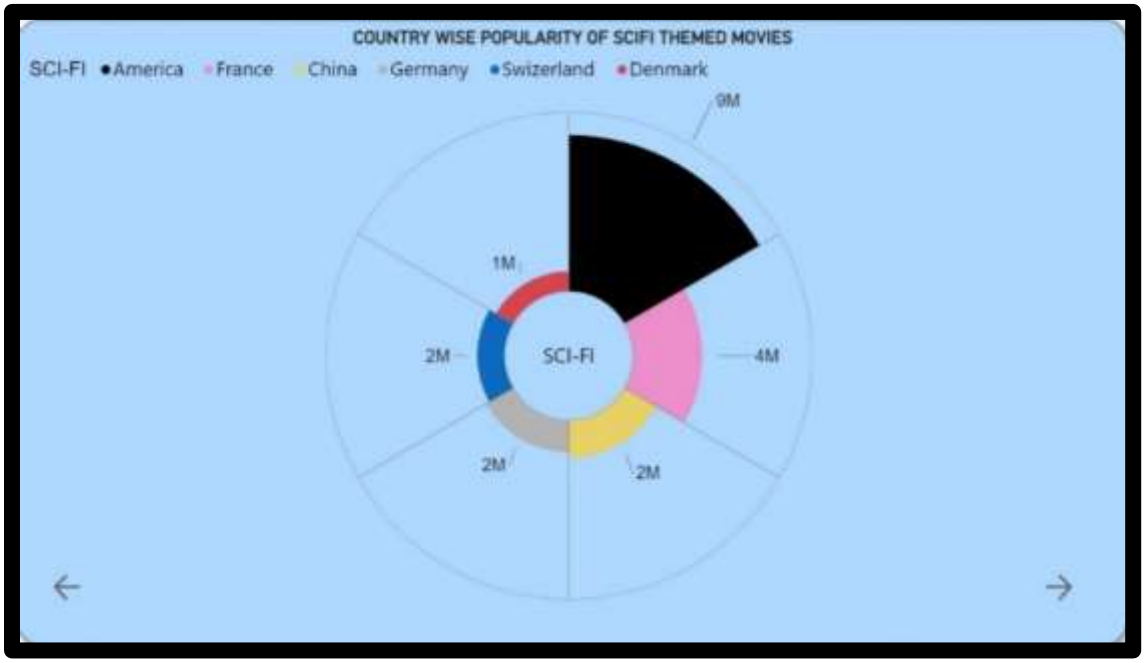


Figure – 3.13 Country-wise popularity of Sci-Fi themed movie

3. **Which genre of movie has the maximum count of average viewers?**

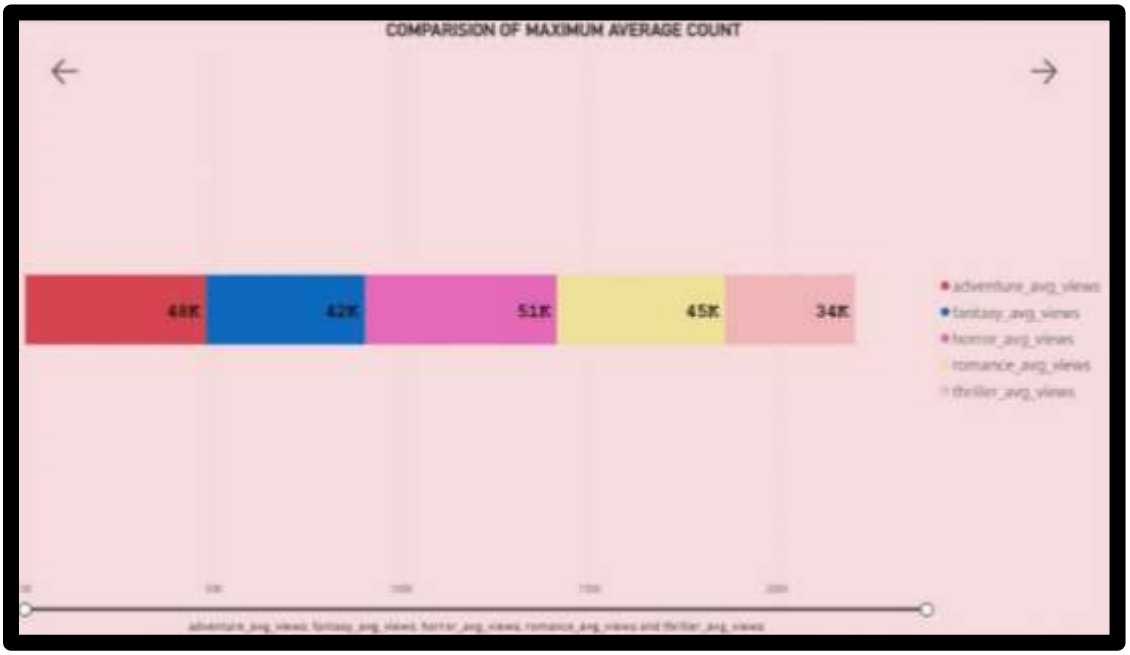


Figure – 3.14 Comparison of maximum average count for different themed movies

4. **What type of polarity is seen the most in horror-themed movies – positive/negative/neutral?**

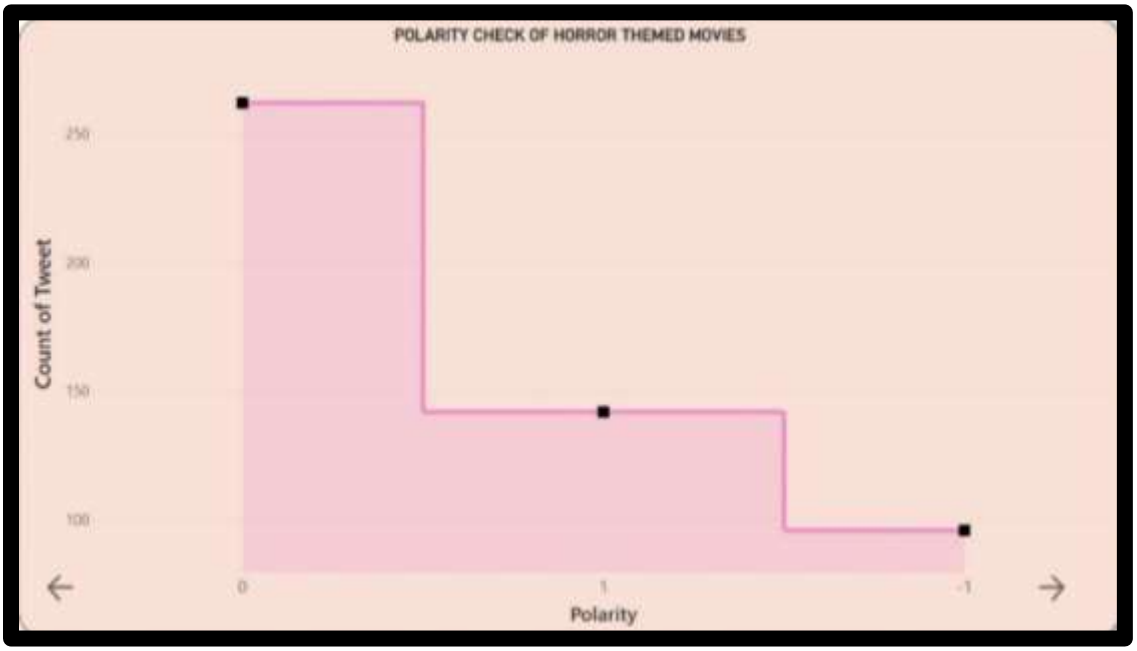


Figure – 3.15 Polarity comparison of Horror themed movies

Comparison of Sales of Mobile Overseas

1. **Comparison of sales of APPLE branded phones in six major countries to find the country with maximum sales.**

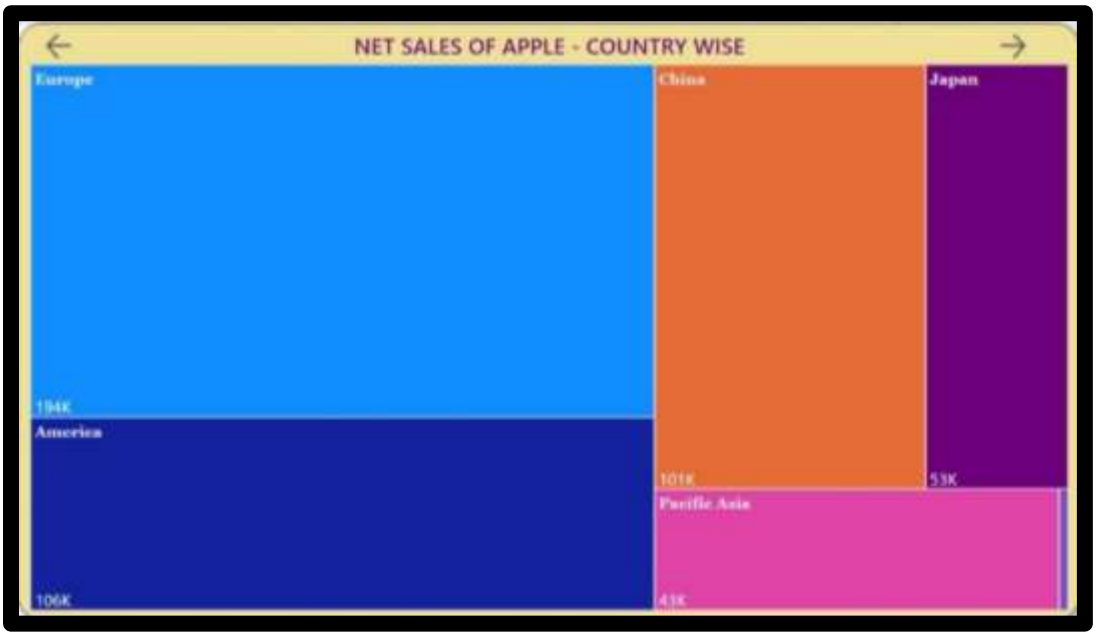


Figure – 3.16 Net sales of APPLE phones are compared over countries

2. **Comparison of the popularity of brands – 'APPLE' and 'SAMSUNG' in five different countries.**

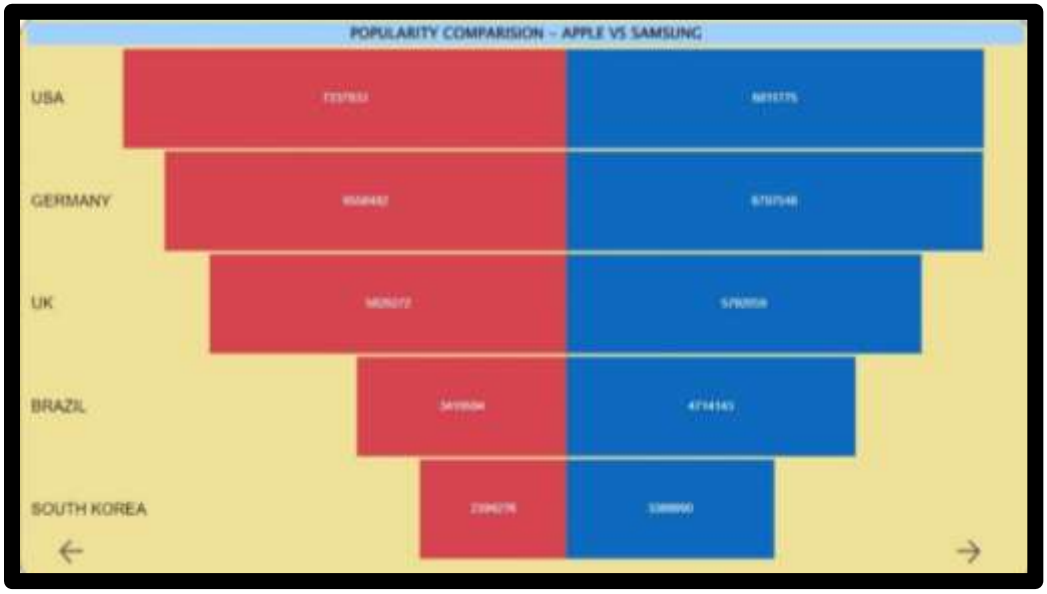


Figure – 3.17 Comparing the popularity of two brands over five countries

3.2. PUBLISHING DASHBOARDS

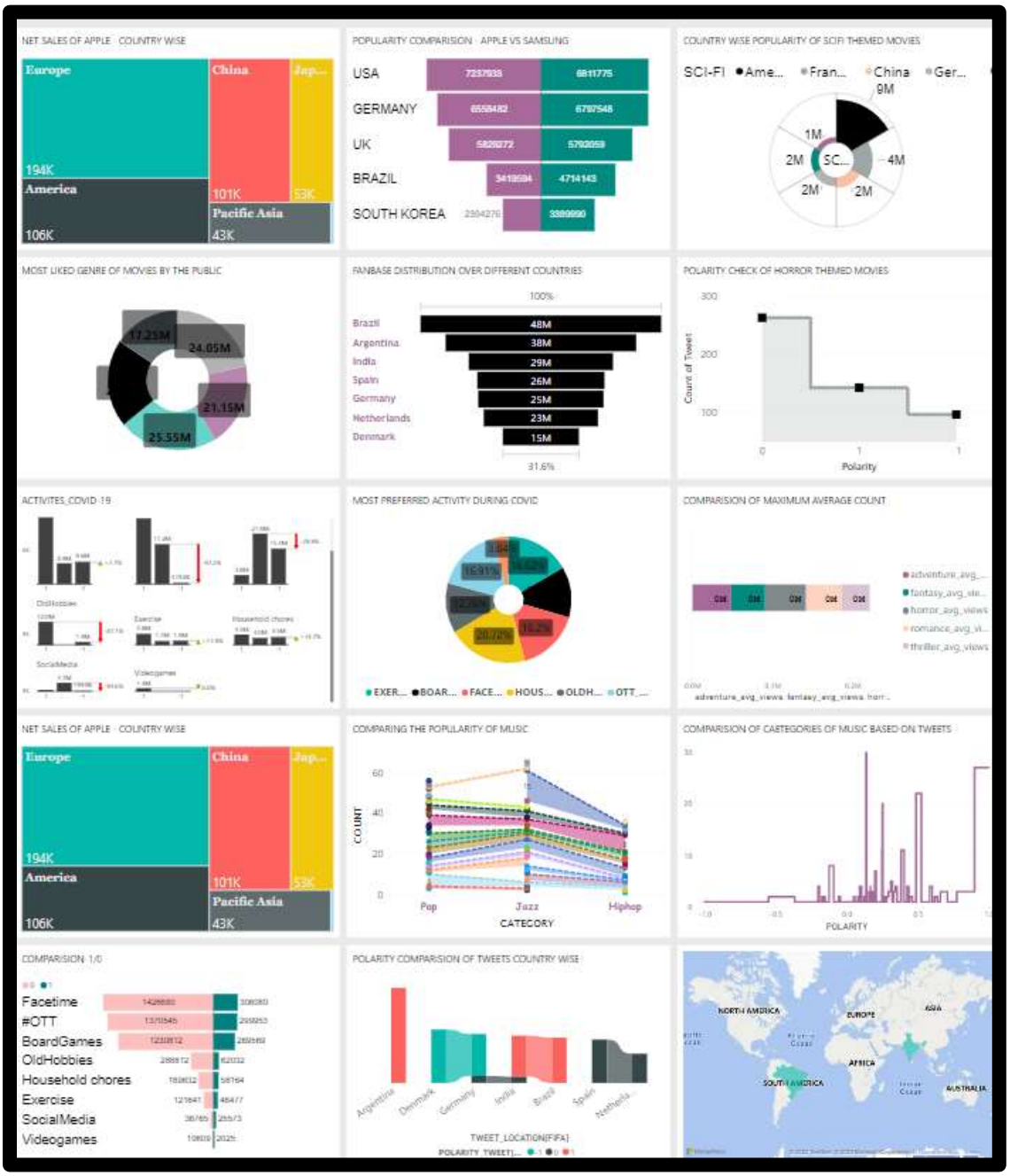


Figure - 3.18 Twitter Sentiment Analysis Dashboard

3.3. INFERENCES

[1] The country Europe shows the maximum sales of APPLE branded mobile phones.

[2] The brand APPLE is more popular as compared to SAMSUNG in the USA.

[3] The citizens of INDIA best prefer horror-themed movies as compared to others themes according to live tweets.

[4] The country America has the maximum number of sci-fi viewers.

[5] The neutral polarity is seen the most in horror-themed movies.

[6] Leader Joe Biden has received the maximum count of positive responses constituting nearly 21 million likes.

[7] The following activities namely – video calls and watching movies on OTT platforms were most preferred in covid times of the year 2020 respectively.

[8] Among the chosen three different themes jazz-based music is most preferred by people.

[9] Brazil had the maximum number of fans and Denmark had the least number of fans during the FIFA World Cup 2022.

CHAPTER – 4
CONCLUSION AND FUTURE WORK

4.1. RECOMMENDATIONS

The analysis performed on the text message of each tweet has given a clear insight into how people place different kinds of opinions on different domains according to the situation and trends respectively. We have been able to implement Natural Language Processing technology using Python codes and pre-defined libraries to analyze the sentiment or emotion behind the text or tweet of an individual. The obtained sentiment analysis of the live tweets has been subjected to different domains of tweets and the results were considered that showed which brand is most popular and which celebrity is more welcomed and has received positive reviews. This also talks about the different kinds of culture and music respectively. The main purpose of such sentiment analysis of the tweets is to know or have a greater hand over the opinion made by people online to identify both the merits and demerits of any newly launched technology or the other extreme that includes the marketing of newly developed products which could be both liked or disked as well. Based on these opinions the businessman can decide what to do next to improve their product.

REFERENCES

[1] Go, A., Huang, L. and Bhayani, R., 2009. Twitter sentiment analysis. *Entropy*, *17*, p.252.

[2] Kouloumpis, Efthymios, Theresa Wilson, and Johanna Moore. "Twitter sentiment analysis: The good the bad and the omg!." In *Proceedings of the international AAAI conference on web and social media*, vol. 5, no. 1, pp. 538-541. 2011.

[3] Sarlan, A., Nadam, C., & Basri, S. (2014, November). Twitter sentiment analysis. In *Proceedings of the 6th International conference on Information Technology and Multimedia* (pp. 212-216). IEEE.

[4] Zimbra D, Abbasi A, Zeng D, Chen H. The state-of-the-art in Twitter sentiment analysis: A review and benchmark evaluation. ACM Transactions on Management Information Systems (TMIS). 2018 Aug 24;9(2):1-29.

[5] Mittal, A. and Goel, A., 2012. Stock prediction using twitter sentiment analysis. *Standford University, CS229 (2011 http://cs229. stanford. edu/proj2011/GoelMittal-StockMarketPredictionUsingTwitterSentimentAnalysis. pdf)*, *15*, p.2352.

[6] Mittal, Anshul, and Arpit Goel. "Stock prediction using twitter sentiment analysis." *Standford University, CS229 (2011 http://cs229. stanford. edu/proj2011/GoelMittal-StockMarketPredictionUsingTwitterSentimentAnalysis. pdf)* 15 (2012): 2352.