INDUSTRIAL TRAINING REPORT

TWITTER SENTIMENT ANALYSIS

Submitted in partial fulfillment of the

Requirements for the award of

Degree of Bachelor of Technology in Computer Science and Engineering

Submitted By:

Name: **Tanya Goel**

University Roll Number: **36614802718**

SUBMITTED TO:

Department of Computer Science and Engineering

MAHARAJA AGRASEN INSTITUTE OF TECHNOLOGY

GGSIPU, DELHI

AH Infotech LLC
Regd Off: 116 Village Blvd, Suite 200, Princeton, NJ 08540
Email: viwaswat.singh@ahinfotechusa.com

AH INFOTECH

CERTIFICATE

This is to certify that Ms. Tanya Goel has completed 6 weeks industrial training during the period from 23/03/2020 to 04/05/2020 in our Organization / Industry as a Partial Fulfillment of Degree of Bachelor of Technology in Computer Science & Engineering. She was trained in the field of Data Analysis using Python ('Twitter Sentiment Analysis').

Viwaswat Singh

Director - Global Operations
Viwaswat.Singh@ahinfotechusa.com
mobile 1(609)865-4949

DECLARATION

I hereby declare that the Training Report entitled "TWITTER SENTIMENT ANALYSIS" is an authentic record of my own work as requirements of 6 weeks Training during the period from 23/03/2020 to 04/05/2020 for the award of degree of Bachelor of Technology (Computer Science & Engineering), GGSIPU, under the guidance of Ms. Viha Gupta (Software Developer at AH Infotech, USA).

Tanya Goel

36614802718

Date: _____

Certified that the above statement made by the student is correct to the best of our knowledge and belief.

Signatures:

Examined by:

1. Viha Gupta (AH Infotech) 2.

(Guide/Trainer) **(Faculty Coordinator)**

Ms. Namita

Head of Department

ACKNOWLEDGEMENT

First and foremost, I thank the Almighty God for sustaining the enthusiasm with which I plunged into this endeavor.

I avail this opportunity to express my profound sense of sincere and deep gratitude to the many people who are responsible for the knowledge and experience I have gained during this Industrial Training.

I have great pleasure in expressing my deep sense of gratitude to **Mr. Vivaswat Singh**, Global Director at AH Infotech, USA for giving me this opportunity to work with his esteemed organization, and **Ms. Viha Gupta**, Software Developer at AH Infotech, USA for guiding me and mentoring me through this whole process. I also express my sincere gratitude and thankfulness to the AH Infotech staff for their cooperation and support.

I express my sincere gratitude to Computer Science Engineering Department, Maharaja Agrasen Institute of Technology worthy principal for providing me this golden opportunity to undergo summer industrial training at **AH Infotech, USA**.

My hearty and inevitable thanks to all the respondents who helped me to bring out the project in a successful manner. Last but not the least, I extend my gratitude towards my parents, faculty and friends who extended their whole-hearted support towards the successful completion of this industrial training.

Thank you

Tanya Goel

Bachelor of Technology (CSE Department)

3rd Year

36614802718

Maharaja Agrasen Institute of Technology, GGSIPU

TABLE OF CONTENTS

ABOUT THE COMPANY

Figure 1: Company logo

INTRODUCTION

AH Infotech is a blooming company in the IT industry. This company was started in 2015 and now has its branches spread over various countries, head-office being in New York, and branches being in Munich and Bangalore, India.

1.1. We are a cloud first company

We help our customers make the move to the cloud in ways that meet security, affordability and scalability concerns.

1.2. We are change agents

We are a consulting services company that uses emerging technologies to power your business into profitability.

1.3. Our people make all the difference

Our people are our partners in driving customer satisfaction.

WHO WE ARE

We are a company focused on delivering high quality cloud solutions for business. We are a group of professionals that has detailed understanding of today's fast-evolving cloud platforms, connected devices and utility computing. We love working with business to solve challenges.

2.1. We are a technology powerhouse

We at AH take our technology prowess seriously. We also understand how to apply technology to business problems and deliver value.

2.2. We take big challenges

No matter what the scale of your business and how complex your processes are, our team will understand and execute.

2.3. Custom enterprise service design

We create custom enterprise services that deliver on the scale and complexity required by the customer. We create custom enterprise services by integrating a wide range of application and development platforms.

OUR SERVICES

We build cloud infrastructure and applications based on business objectives and stage of cloud-readiness. Our multi-disciplinary approach helps customers take informed decisions and achieve desired outcomes.

3.1. Enterprise Services

We understand that there is a lot to optimize and better in the traditional IT landscape and our team helps you run and manage this efficiently. AH provides application design, development and testing services across a broad range of technology platforms. Our architects and engineers comprise a team that can take requirements of complex, heterogeneous technical components and create a solution that you can host in your data center or deploy to the cloud.

3.2.Cloud Migrate

We consult with Business and IT in connecting the dots and draw the roadmap to cloud-enable your business applications and transition them to cloud based services. We offer a comprehensive and diverse range of cloud solutions; be it modeling data in cloud for analytics, business rules engine based on cloud services and even securing and isolating personal and corporate data with user identity in cloud.

3.3.Cloud Manage

Moving applications to cloud has its benefits but the enterprises encounter many challenges. Finding the required skills and resources, adopting cloud concepts and practices such as automation of infrastructure and operations, managing change of underlying platforms (leading to cyclical development) and testing of cloud-deployed applications are some of these challenges. We at AH are focused to meet the demands of the enterprise and help cloud transitions smoother and to ensure your lights always stay on.

VISION

They help their customers make the move to the cloud in ways that meets:

4.1. Security

4.2. Affordability

4.3. Scalability concerns

LIST OF TABLES

LIST OF FIGURES

INTRODUCTION CHAPTER

The internship at AH Infotech comprised of performing Twitter Sentiment Analysis on the tweets collected manually by writing a script in python using API keys authorized by Twitter using Twitter Developers Account.

I performed the analysis on the tweets made on the artists featured by the Sony Music International (parent company being Sony Music), a client of AH Infotech.

Sony Music International features many artists every month, in the month of April, one of the featured artists was Doja Cat, who is an American singer and rapper. I performed a sentiment analysis on the tweets that were posted mentioning her ('@DojaCat', '#DojaCat', '#dojacat', etc.).

The motivation behind doing this project was to analyze the social media presence of singers featured by the company. It'll help them in analyzing the company's representors.

This project aimed to help company know the social media influence and sentiment displayed by the audience of their artists.

I wrote scripts in Python to collect data and to perform the sentiment analysis on the data that was collected. I used various libraries such as tweepy, pandas, numpy, dtextblob, googletrans, re, matplotlib, csv, time, datetime and timedelta throughout this whole project.

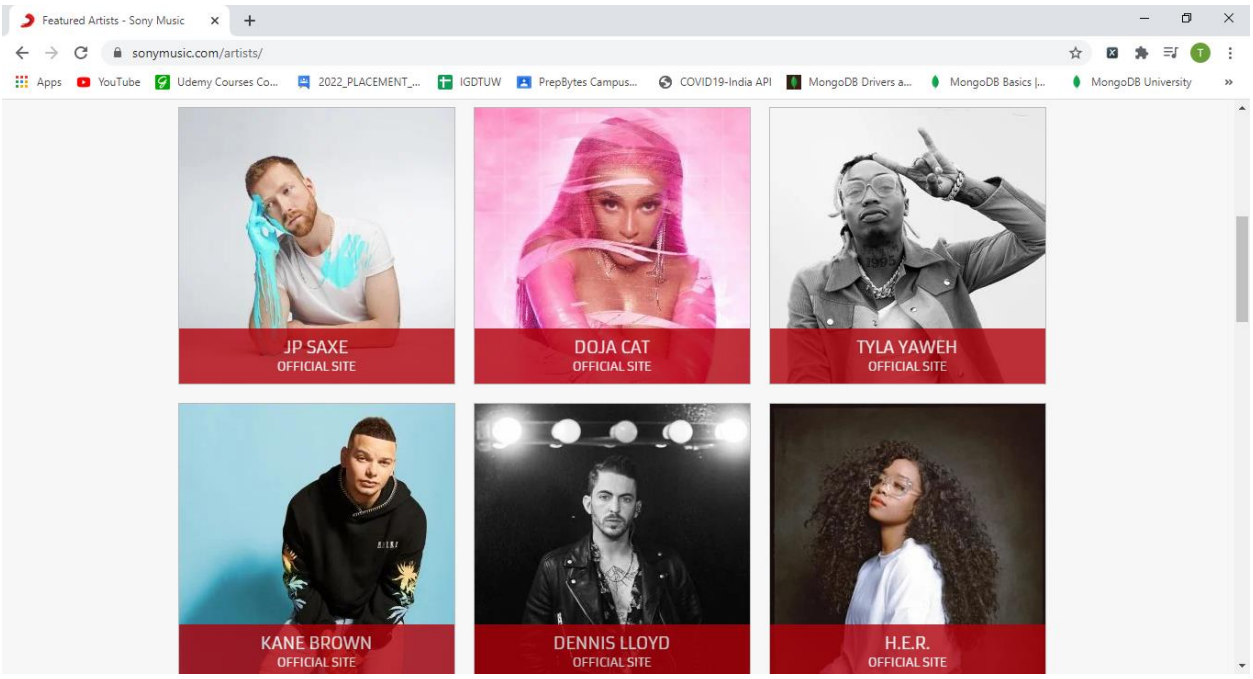

Figure 2: Screenshot of featured artists of Sony Music International in the month of April

TOOLS AND TECHNOLOGIES USED

2.1. LANGUAGE USED:

2.1.1. Python programming language.

2.2. LIBRARIES USED:

2.2.1. Pandas library

2.2.2. Numpy library

2.2.3. Tweepy library

2.2.4. CSV library

2.2.5. Matplotlib library

2.2.6. GoogleTrans library

2.2.7. RegEx library

2.2.8. TextBlob library

2.2.9. Time and DateTime libraries

2.2.10. Seaborn library

2.2.11. Plotly library

2.2.12. WordCloud library

2.3. PLATFORMS USED:

2.3.1. Google Colab

2.3.2. Jupyter Notebook

2.4. DATA COLLECTION:

2.4.1. Twitter APIs, authenticated keys provided by Twitter

TECHNICAL CONTENT

3.1. PROGRAMMING LANGUAGE LEARNED AND IMPLEMENTED:

3.1.1. Python

Python is a high-level, interpreted, general-purpose, interactive and object-oriented scripting language, Created by Guido van Rossum and first released in 1991. Python is designed to be highly readable. It uses English keywords frequently where as other languages use punctuation, and it has fewer syntactical constructions than other languages.

Some of the key advantages of learning Python are:

- Python is Interpreted − Python is processed at runtime by the interpreter. We do not need to compile our program before executing it. This is similar to PERL and PHP.
- Python is Interactive − we can actually sit at a Python prompt and interact with the interpreter directly to write your programs.
- Python is Object-Oriented − Python supports Object-Oriented style or technique of programming that encapsulates code within objects.
- Python is a Beginner's Language − Python is a great language for the beginner-level programmers and supports the development of a wide range of applications from simple text processing to WWW browsers to games.

Python is one of the most widely used language over the web for various reasons:

- Easy-to-learn − Python has few keywords, simple structure, and a clearly defined syntax. This allows the student to pick up the language quickly.
- Easy-to-read − Python code is more clearly defined and visible to the eyes.
- Easy-to-maintain − Python's source code is fairly easy-to-maintain.
- A broad standard library − Python's bulk of the library is very portable and cross-platform compatible on UNIX, Windows, and Macintosh.
- Interactive Mode − Python has support for an interactive mode which allows interactive testing and debugging of snippets of code.

- Portable − Python can run on a wide variety of hardware platforms and has the same interface on all platforms.
- Extendable − We can add low-level modules to the Python interpreter. These modules enable programmers to add to or customize their tools to be more efficient.
- Databases − Python provides interfaces to all major commercial databases.
- GUI Programming − Python supports GUI applications that can be created and ported to many system calls, libraries and windows systems, such as Windows MFC, Macintosh, and the X Window system of Unix.
- Scalable − Python provides a better structure and support for large programs than shell scripting.

Following are important characteristics of Python Programming −

- It supports functional and structured programming methods as well as OOP.
- It can be used as a scripting language or can be compiled to byte-code for building large applications.
- It provides very high-level dynamic data types and supports dynamic type checking.
- It supports automatic garbage collection.
- It can be easily integrated with C, C++, COM, ActiveX, CORBA, and Java.

Figure 3: Python logo

3.2. LIBRARIES USED:

3.2.1. Pandas

In computer programming, pandas is a software library written for the Python programming language for data manipulation and analysis. In particular, it offers data structures and operations for manipulating numerical tables and time series. The name is derived from the term "panel data", an econometrics term for data sets that include observations over multiple time periods for the same individuals.

It is used to analyze and tackle datasets easily. It has many functions pre-defined in it which makes it easier to use and for the user to handle huge datasets easily. It makes operating on datasets easier. It also allows user to read files such as a csv file (comma separated values file) and convert them into data-frames and then perform operations on them.

The library is highly optimized for performance, with critical code paths written in Cython or C. This library has many features, such as:

- DataFrame object for data manipulation with integrated indexing.
- Tools for reading and writing data between in-memory data structures and different file formats.
- Data alignment and integrated handling of missing data.
- Reshaping and pivoting of data sets.
- Label-based slicing, fancy indexing, and sub setting of large data sets.
- Data structure column insertion and deletion.
- Group by engine allowing split-apply-combine operations on data sets.
- Data set merging and joining.
- Hierarchical axis indexing to work with high-dimensional data in a lower-dimensional data structure.
- Time series-functionality: Date range generation and frequency conversion, moving window statistics, moving window linear regressions, date shifting and lagging.
- Provides data filtration.

Figure 4: Pandas

3.2.2. Numpy

NumPy is a python library used for working with arrays. It also has functions for working in domain of linear algebra, fourier transform, and matrices. It was created in 2005 by Travis Oliphant. It is an open source project and anyone can use it freely. 'NumPy' stands for Numerical Python.

This library aims to provide an array object that is up to 50x faster than traditional Python lists. NumPy arrays are stored at one continuous place in memory unlike lists, so processes can access and manipulate them very efficiently.

It is a Python library and is written partially in Python, but most of the parts that require fast computation are written in C or C++. It makes it easier to tackle multi-dimensional arrays, datasets and matrices.

Figure 5: NumPy

3.2.3. Tweepy

It is the official software library written to access the Twitter APIs. It makes it very easy to collect the required data from Twitter using the official authenticated APIs provided by Twitter itself for researchers to analyze the data. To use APIs, a user must have a Twitter Developers Account first, where he/she has to create an app to generate API key, Secret API key, Access token and Secret access token, all four of which are essential in order to collect data.

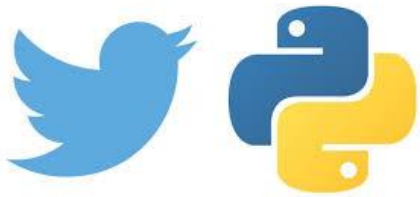

Figure 6: Tweepy

3.2.4. CSV

CSV (Comma Separated Values) is a simple file **format** used to store tabular data, such as a spreadsheet or database. A CSV file stores tabular data (numbers and text) in plain text. Each line of the file is a data record. Each record consists of one or more fields, separated by commas. The use of the comma as a field separator is the source of the name for this file format.

For working with CSV files in python, there is an inbuilt module called csv. It is the library that has been written for Python programming language to handle spreadsheet data, i.e. data having multiple rows and columns. It allows us to easily read a csv file and convert it into a dataframe, and once we are done working and analyzing the file, we can again save the dataframe into a csv file.

Figure 7: CSV

3.2.5. Matplotlib

Matplotlib is a comprehensive library written to create static, animated, and interactive visualizations in Python so that the researchers can analyze the data better and can come to better conclusions. It very easily plots large datasets into graphs and plots very easily and quickly.

Matplotlib produces publication-quality figures in a variety of hardcopy formats and interactive environments across platforms. Matplotlib can be used in Python scripts, the Python and IPython shell, web application servers, and various graphical user interface toolkits.

Some examples of plot that can be plotted using matplotlib library:

- Line plot
- Histogram
- Bar graphs
- Scatter plot
- 3D plot
- Image plot
- Contour plot
- Polar plot

Figure 8: Matplotlib

3.2.6. GoogleTrans

Googletrans is a **free** and **unlimited** python library that implemented Google Translate API. This uses the Google Translate Ajax API to make calls to such methods as detect and translate.

Features of GoogleTrans:

- Fast and reliable - it uses the same servers that translate.google.com uses
- Auto language detection
- Bulk translations
- Customizable service URL
- Connection pooling (the advantage of using requests.Session)
- HTTP/2 support

It is an official Google library written for Python users to translate any content that they encounter during their work or research. The maximum character limit on a single text is 15k. Due to limitations of the web version of google translate, this API does not guarantee that the library would work properly at all times. (so please use this library if you don't care about stability.) If you want to use a stable API, I highly recommend you to use Google's official translate API. If you get HTTP 5xx error or errors like #6, it's probably because Google has banned your client IP address

Figure 9: GoogleTrans

3.2.7. RegEx

A RegEx, or Regular Expression, is a sequence of characters that forms a search pattern. RegEx can be used to check if a string contains the specified search pattern. Python has a built-in package called `re`, which can be used to work with Regular Expressions.

It is the software library written to execute regular expression feature in Python scripts. They work by finding a specific string of character(s) within another string as per the users' requirement.

Built-in methods in RegEx:

Function	Description
findall	Returns a list containing all matches
search	Returns a Match object if there is a match anywhere in the string
split	Returns a list where the string has been split at each match
sub	Replaces one or many matches with a string

Table 1: Built-in methods in RegEx

Figure 10: RegEx

3.2.8. TextBlob

It is a Python library that helps in processing textual data. It provides access to the user to common text-processing operations and tasks. It provides a simple API to perform Natural Language Programming (NLP) tasks and is a very helpful library when working with textual data and performing NLP tasks such as sentiment analysis, classification, etc.

Features of TextBlob:

- Noun phrase extraction
- Part-of-speech tagging
- Sentiment analysis
- Classification (Naive Bayes, Decision Tree)
- Tokenization (splitting text into words and sentences)
- Word and phrase frequencies
- Parsing
- **n**-grams
- Word inflection (pluralization and singularization) and lemmatization
- Spelling correction
- Add new models or languages through extensions
- WordNet integration

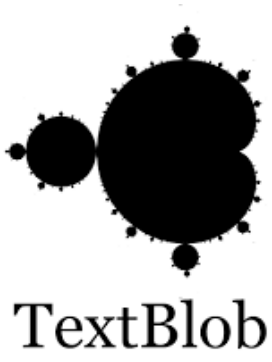

Figure 11: TextBlob

3.2.9. Time and DateTime

These two are the libraries for Python to handle time and date format. They allow multiple features such as finding time difference, converting dates as per our convenient format, etc. These libraries are very useful for someone working with time and dates data. Python timedelta object is used to perform datetime manipulations in an easy way. The timedelta class is part of datetime module. The timedelta object supports mathematical operations such as addition, subtraction, multiplication etc. using basic operators, so it's very easy to use it. It's mostly used to get a datetime object with some delta date and time. Python timedelta object is very useful for datetime manipulations. The support for basic arithmetic operators makes it very easy to use.

Figure 12: Time, DateTime and TimeDelta

3.2.10. Seaborn

Seaborn is a library for making statistical graphics in Python. It is built on top of matplotlib and closely integrated with pandas data structures. Seaborn aims to make visualization a central part of exploring and understanding data. Its dataset-oriented plotting functions operate on dataframes and arrays containing whole datasets and internally perform the necessary semantic mapping and statistical aggregation to produce informative plots.

Here is some of the functionality that seaborn offers:

- A dataset-oriented API for examining relationships between multiple variables
- Specialized support for using categorical variables to show observations or aggregate statistics
- Options for visualizing univariate or bivariate distributions and for comparing them between subsets of data
- Automatic estimation and plotting of linear regression models for different kinds dependent variables
- Convenient views onto the overall structure of complex datasets
- High-level abstractions for structuring multi-plot grids that let you easily build complex visualizations
- Concise control over matplotlib figure styling with several built-in themes
- Tools for choosing color palettes that faithfully reveal patterns in your data

Figure 13: Seaborn logo

3.2.11. Plotly

The plotly Python library is an interactive, open-source plotting library that supports over 40 unique chart types covering a wide range of statistical, financial, geographic, scientific, and 3-dimensional use-cases.

Figure 14: Plotly logo

3.2.12. WordCloud

WordCould library is used to generate a word-cloud of the most used words in a text block, where the word having highest frequency is biggest in size and the word having least frequency is smallest in size.

Figure 15: WordCloud example

3.3. SOFTWARES USED:

3.3.1. Google Colab

It is the official Google Code compiler and executer available for free-of-cost online. Colaboratory is a Google research project created to help disseminate machine learning education and research. It's a Jupyter notebook environment that requires no setup to use and runs entirely in the cloud. It allows the user to write and execute Python in your browser, with

- Zero configuration required
- Free access to GPUs
- Easy sharing

Whether the user is a student, a data scientist or an AI researcher, Colab can makes their work easier. Colab notebooks allow to combine executable code and rich text in a single document, along with images, HTML, LaTeX and more. It makes it easier to download various python libraries in a virtual cloud environment. When the user creates their own Colab notebooks, they are stored in their Google Drive account. They can easily share Their Colab notebooks with co-workers or friends, allowing them to comment on their notebooks or even edit them. The user is not even required to download the data files onto their systems, but can rather just easily mount their drives onto the notebook and access all the data and files present there.

Figure 16: Google Colab

3.3.2. Jupyter Notebook

Project Jupter exists to develop open-source softwares, open-standards, and services for interactive computing across dozens of programming languages. The Jupter Notebook is an open-source web application that allows the user to create and share documents that contain live code, equations, visualizations and narrative text. Its some uses can be:

- Data Cleaning
- Data Transformation
- Numerical Simulation
- Statistical Modeling
- Data Visualization
- Machine Learning

It can be downloaded along with the Anaconda environment and can be easily accessed using Anaconda prompt. We can create virtual environments using it, in which each environment is different than the other as per the users' requirement and has the required separate libraries with different versions as required.

Figure 17: Jupyter Notebook

3.4. UNIQUE DATA TYPES USED:

3.4.1. Lists

They are ordered and mutable datatypes that are easy to access and work upon. In it when we store data, each and every data can be accessed and analyzed using indexing. It also allows the user to use negative indexing, i.e. reverse indexing which starts from the last element present in the list. The user can even specify a range of indexes by specifying where to start and where to end the range in a list. We can store multiple datatypes with a list, i.e. we can store dictionaries, strings, integers, float values and even lists within lists. They allow repetition of values stored in them. They are very easy to work with and to be operated upon. They are declared using square brackets '[ ]'.

Python has a set of built-in methods that users can use on lists:

Method	**Description**
append()	Adds an element at the end of the list
clear()	Removes all the elements from the list
copy()	Returns a copy of the list
count()	Returns the number of elements with the specified value
extend()	Add the elements of a list (or any iterable), to the end of the current list
index()	Returns the index of the first element with the specified value

insert()	Adds an element at the specified position
pop()	Removes the element at the specified position
remove()	Removes the item with the specified value
reverse()	Reverses the order of the list
sort()	Sorts the list

Table 2: Built-in methods for lists

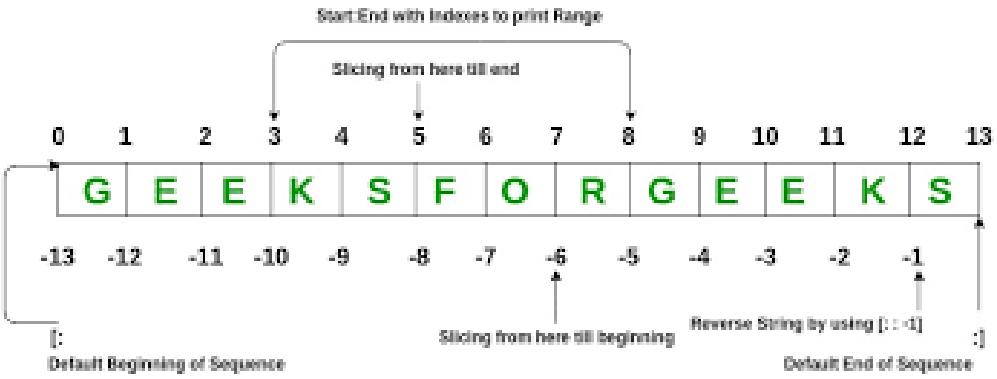

Figure 18: Lists

3.4.2. Dictionaries

They are the unordered, mutable and indexed datatypes consisting of key, value pair in which we can access all the values just by inputting the key values or using indexing. They make it easier to access large data's. While using them, the users does not have to remember indexing or the length of data, but can simply access values using keys and can easily operate on them. Dictionaries can store multiple datatypes within themselves, i.e. it can store dictionaries, strings, integers, float values and even dictionaries within dictionaries. They also allow repetition of values stored in them. They are declared using curly braces '{}'.

Python has a set of built-in methods that users can use on dictionaries:

Method	**Description**
clear()	Removes all the elements from the dictionary
copy()	Returns a copy of the dictionary
fromkeys()	Returns a dictionary with the specified keys and value
get()	Returns the value of the specified key
items()	Returns a list containing a tuple for each key value pair
keys()	Returns a list containing the dictionary's keys

pop()	Removes the element with the specified key
popitem()	Removes the last inserted key-value pair
setdefault()	Returns the value of the specified key. If the key does not exist: insert the key, with the specified value
update()	Updates the dictionary with the specified key-value pairs
values()	Returns a list of all the values in the dictionary

Table 3: Built-in methods for dictionaries

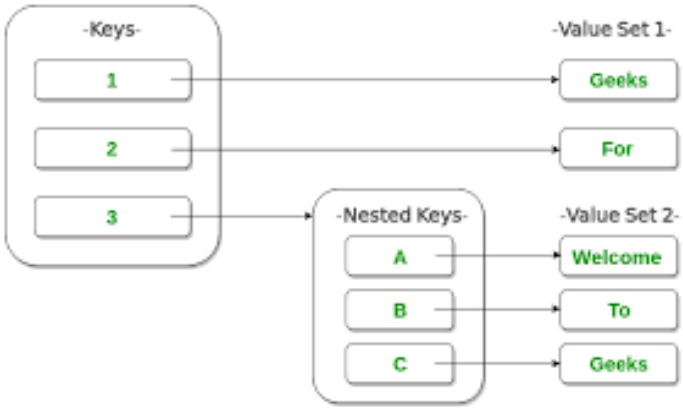

Figure 19: Dictionaries

3.4.3. Sets

They are the unordered, unindexed and unique valued datatypes which do not allow repetition of values among it. It stores a specific value only once. They are declared using the keyword ‘set’ followed by a pair of curly brackets ‘{}’. The user cannot access items in a set by referring to an index, since sets are unordered the items has no index.

Method	**Description**
add()	Adds an element to the set
clear()	Removes all the elements from the set
copy()	Returns a copy of the set
difference()	Returns a set containing the difference between two or more sets
difference_update()	Removes the items in this set that are also included in another, specified set
discard()	Remove the specified item
intersection()	Returns a set, that is the intersection of two other sets

intersection_update()	Removes the items in this set that are not present in other, specified set(s)
isdisjoint()	Returns whether two sets have a intersection or not
issubset()	Returns whether another set contains this set or not
issuperset()	Returns whether this set contains another set or not
pop()	Removes an element from the set
remove()	Removes the specified element
symmetric_difference()	Returns a set with the symmetric differences of two sets
symmetric_difference_update()	inserts the symmetric differences from this set and another
union()	Return a set containing the union of sets

update()	Update the set with the union of this set and others

Table 4: Built-in methods for sets

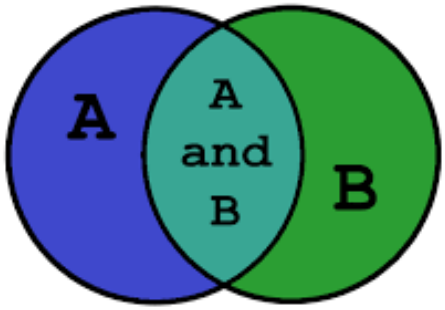

Figure 20: Sets

3.4.4. Tuples

They are ordered but immutable datatype available in Python. They store values in a pair having two values. Since they are ordered, its elements can be accessed using indexing in square brackets '[]'. They also support negative indexing. They are similar to lists apart from the fact that they are declared in parenthesis '( )' and are immutable.

Python has two built-in methods that users can use on tuples:

Method	Description
count()	Returns the number of times a specified value occurs in a tuple
index()	Searches the tuple for a specified value and returns the position of where it was found

Table 5: Built-in methods for tuples

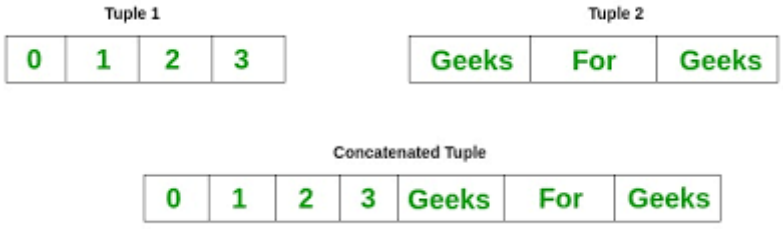

Figure 21: Tuples

3.5. DATA COLLECTION:

3.5.1. Twitter API

Twitter provides API keys to collect data from their website. In order to collect data, a user must have Twitter Developers Account and should have created an app in their account. Once an app is created, Twitter provides API key, API Secret key, Access token and Secret Access Token to the user to authorize the user to collect data legally and formally. A script is written in Python to access those keys and to use them to collect our required dataset for some specific queries, which are also decided by the user.

Figure 22: Twitter API

3.6. ANALYSIS EXECUTION:

3.6.1. Data Preprocessing

3.6.1.1. The data collected had a lot of redundancy and duplicate tweets, so the data was sorted into ascending order and then duplicate rows of data were removed and only the required dataset was kept and worked upon.

3.6.1.2. Furthermore, the data had multiple columns (thirty-four columns) which were not required in our task, so we dropped those columns and our column count was reduced to six columns.

3.6.2. Table Processing

3.6.2.1. Calculated max followers count of each user and added that column to the dataframe.

3.6.2.2. Calculated max retweet count of each user and added that column to the dataframe.

3.6.2.3. Added the columns of sentiment and sentiment score of each tweet as per the words used in it (positive, negative, neutral).

3.6.2.4. Removed unnecessary columns and added new user made columns in the dataframe.

```
Index(['created_at', 'id', 'id_str', 'full_text', 'truncated',
       'display_text_range', 'entities', 'extended_entities', 'metadata',
       'source', 'in_reply_to_status_id', 'in_reply_to_status_id_str',
       'in_reply_to_user_id', 'in_reply_to_user_id_str',
       'in_reply_to_screen_name', 'user', 'geo', 'coordinates', 'place',
       'contributors', 'is_quote_status', 'retweet_count', 'favorite_count',
       'favorited', 'retweeted', 'possibly_sensitive', 'lang',
       'retweeted_status', 'quoted_status_id_str', 'quoted_status_id',
       'quoted_status', 'withheld_in_countries', 'withheld_scope',
       'Unnamed: 0.1', 'Unnamed: 0.1.1'],
      dtype='object')
```

Figure 23: Screenshot of columns before table processing

	user_id	followers_count	retweet_count	sentiment_score	sentiment
0	48871957	180	0	0	neutral
1	92458611	396	0	0	neutral
2	147218661	2361	76	1	positive
3	220192865	16553	0	0	neutral
4	289856832	2909	0	1	positive

Figure 24: Screenshot of columns after table processing

3.6.3. Data Analysis

3.6.3.1. Once our data was cleaned and pre-processed, we had to analyze our data. For that we used Re library.

3.6.3.2. From each tweet special characters such as '@', '#', '$', etc. were removed and all the hyper-links, images and videos were removed to analyze the text of the tweet.

3.6.3.3. Once these special characters were removed, they were passed through a user defined function which analyzed the text and classified it as a 'positive', 'negative' or 'neutral' text.

The results were stored in another dataframe and csv file having only the tweet and the sentiment of that tweet from the analysis performed above.

3.7. EXPLORATORY DATA ANALYSIS (EDA)

Plotted various graphs to analyze the data using matplotlib, seaborn and plotly libraries.

3.8. WORD CLOUD

Generated a word cloud using wordcloud, regex and googletrans library.

SNAPSHOT(S)

```
Neutral tweets percentage: 49.01960784313725 %
Positive tweets percentage: 38.56209150326798 %
Negative tweets percentage: 12.418300653594772 %
```

Figure 25: Screenshot of result percentage

	user_id	followers_count	retweet_count	sentiment_score	sentiment
0	48871957	180	0	0	neutral
1	92458611	396	0	0	neutral
2	147218661	2361	76	1	positive
3	220192865	16553	0	0	neutral
4	289856832	2909	0	1	positive

Figure 26: Screenshot of final dataset

RESULTS AND DISCUSSIONS

5.1. Sentiment Analysis Result

This project analyzes the social media (Twitter) presence of the artist, namely Doja Cat:

5.1.1. Neutral tweets percentage = 49.0196 %

5.1.2. Positive tweets percentage = 38.5621 %

5.1.3. Negative tweets percentage = 12.4183 %

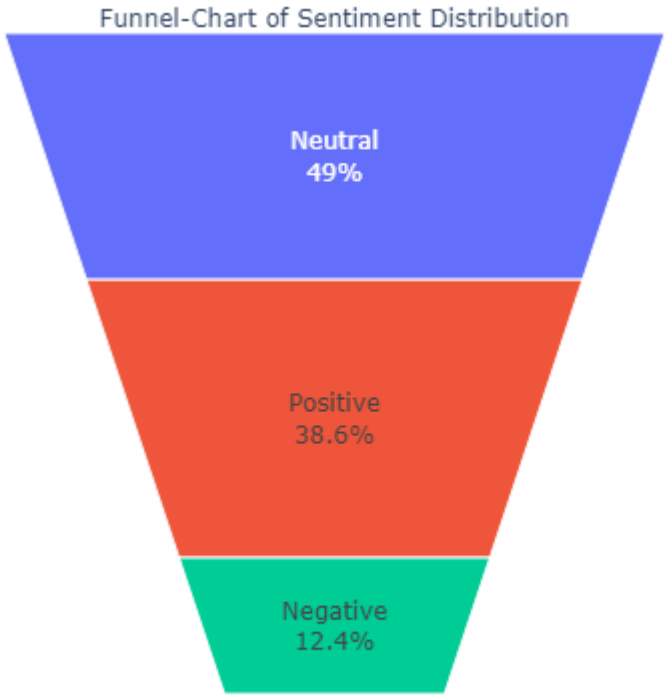

Figure 27: Funnel graph of sentiment analysis

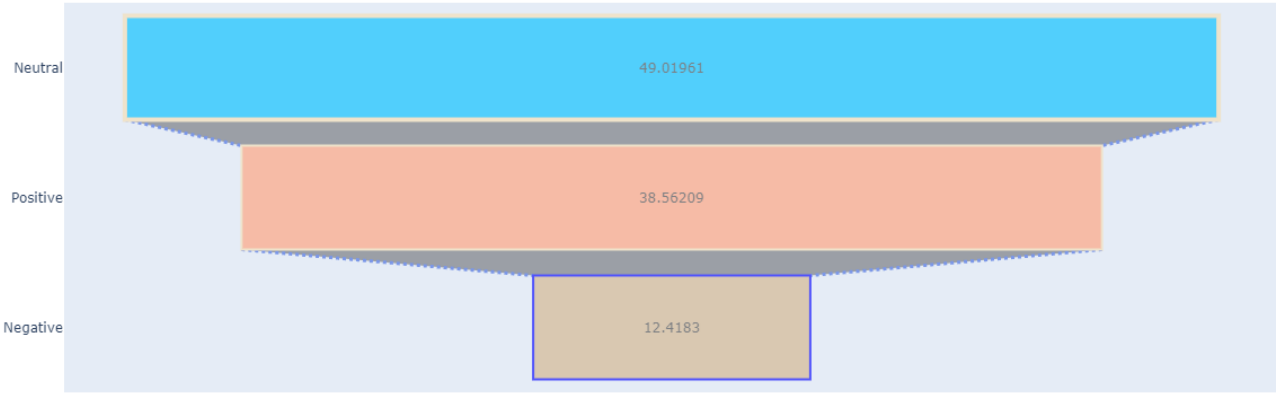

Figure 28: Comparative funnel graph of sentiment analysis

5.2. Exploratory Data Analysis

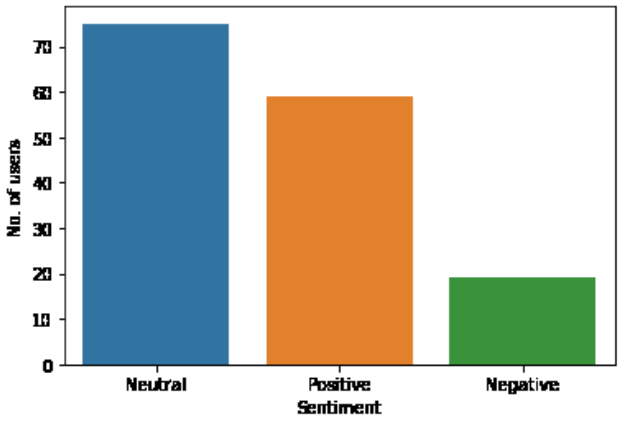

Figure 29: Sentiment vs Number of users graph

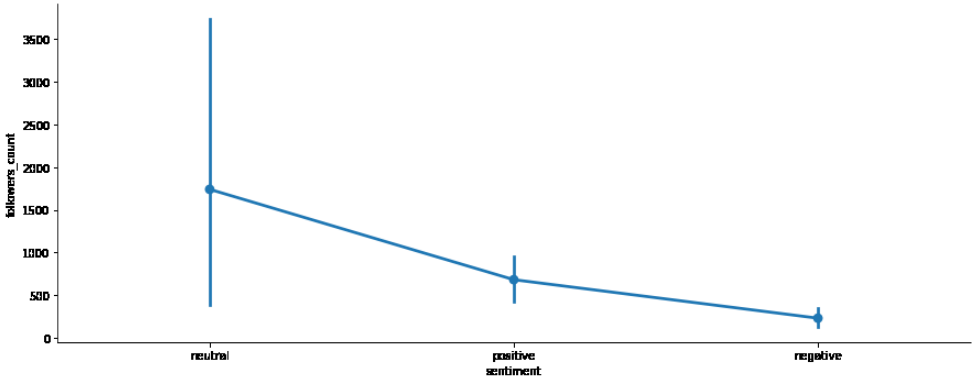

Figure 30: Sentiment vs Followers count

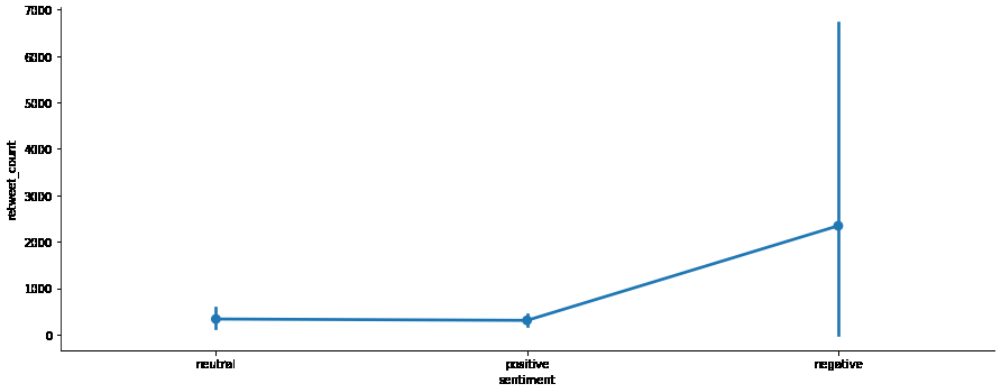

Figure 31: Sentiment vs Retweet count

5.3. Word Cloud

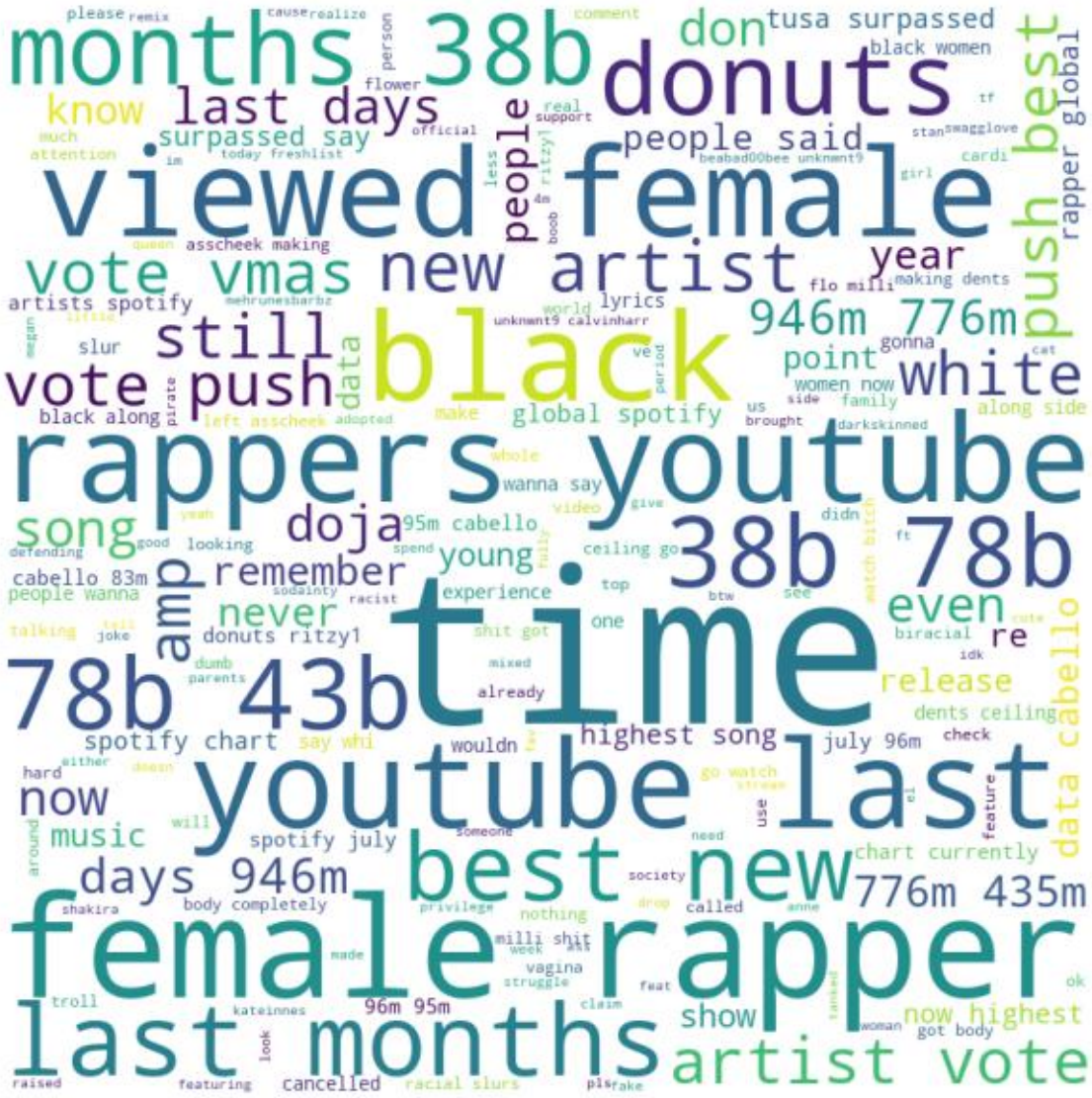

Figure 32: Word cloud generated

CONCLUSIONS AND FUTURE SCOPE

Through this analysis we could find out the social media presence of Doja Cat and how her audience perceives her thoughts. We can also see from the graphs that there are more of neutral tweets than positive tweets, and the negative tweets are the least. We can also see from the graphs that the users who have a neutral emotion have the largest followers count whereas, interestingly, the users who have posted negative tweets have the largest retweet count.

FUTURE SCOPE(S):

- Such analysis can be performed on other featured artists of the Sony Music Entertainment company.
- We can perform a similar Twitter Sentiment Analysis on the songs released by artists featured by Sony Music Entertainment to analyze their reach and peoples reaction towards those songs.

WEEKLY JOBS SUMMARY

WEEK 1:

Description of activity, task, duty, or responsibility	Performed with Team	Performed alone	Time Spent
Studied the history about the company and what the company does		X	2 days
Studied the client website to decide onto a topic		X	1 day
Researched about the featured artists to be chosen for the project		X	2 days

List one thing that went particularly well this week: My project topic got decided and it was a super exciting one.
List one thing that was the most challenging this week: Deciding the artist on whom analysis was to be done.

WEEK 2:

Description of activity, task, duty, or responsibility	Performed with Team	Performed alone	Time Spent
Studied and practiced implementing Python basics and how to write functions in it		X	4 days
Studied and practiced implementing the various libraries required for this project		X	2 days

List one thing that went particularly well this week: Learned Python and various libraries it provides.

List one thing that was the most challenging this week: Implementing what was learned because of being absolutely new to Python.

WEEK 3:

Description of activity, task, duty, or responsibility	Performed with Team	Performed alone	Time Spent
Studied about Twitter APIs and how to use them		X	1 day
Created a Twitter Developer Account (got accepted after 1 day)		X	1 day
Created an app in the said Twitter account and generated keys for the script		X	1 hour
Worked on writing a script to collect tweets using various libraries		X	3 days

List one thing that went particularly well this week: Learned about Twitter, APIs and wrote my first script in Python.
List one thing that was the most challenging this week: Writing the said script without any prior experience.

WEEK 4:

Description of activity, task, duty, or responsibility	Performed with Team	Performed alone	Time Spent
Collected data rigorously on all the five artists (script ran whole days)		X	5 days
Researched about the project Twitter Sentiment Analysis		X	2 days

List one thing that went particularly well this week: Collected data comparatively quickly and learned about drawbacks of Google colab.
List one thing that was the most challenging this week: Collecting data.

WEEK 5:

Description of activity, task, duty, or responsibility	**Performed with Team**	**Performed alone**	**Time Spent**
Worked on writing a script to preprocess my data into workable form		X	1 day
Preprocessed the data		X	1 day
Worked on writing a script to perform sentiment analysis		X	2 days
Performed sentiment analysis on the selected artist		X	1 day

List one thing that went particularly well this week: Wrote my second script, preprocessed data and performed sentiment analysis.
List one thing that was the most challenging this week: Writing the script to perform sentiment analysis.

WEEK 6:

Description of activity, task, duty, or responsibility	Performed with Team	Performed alone	Time Spent
Performed EDA on the final dataset		X	1 day
Generated word cloud		X	1 day
Wrote report on my project to be submitted in the company		X	3 days

List one thing that went particularly well this week: Completing the analysis and plotting graphs and wordcloud.
List one thing that was the most challenging this week: Writing my first report for the company.

Self-Evaluation: (Circle one)

A+ (A) A- B+ B B- C+ C C- D+ D D- F

List one way you can improve your performance: I believe by gaining more skills I can speed up the process and work on more complex problems that could benefit the company more.

MAHARAJA AGRASEN INSTITUTE OF TECHNOLOGY

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

SUMMER TRAINING/ INDUSTRIAL WORKSHOP/ CERTIFICATION FEEDBACK FORM

Name of the student: Semester:

Name of the Industry: Duration with Dates:

Type of Industry (PSU/ Semi Govt/ Private/ online course): Whether report has been submitted:

Following are the ratings

1 – Unsatisfactory **2 – Satisfactory** **3 – Good** **4 – Very Good** **5 – Excellent**

	Questions	Rating (1 to 5)
1	Relevance of the training with respect to B.Tech CSE curriculum	3
2	Did you work as team member, team leader or as an individual during the training?	Individual
3	Have you done research, implementation, analysis, data interpretation synthesis of the information?	All
4	Are you able to apply the knowledge of mathematics, science, engineering fundamentals, and an engineering specialization during the training?	Yes
5	Are you able to identify any specific technical problems (bugs) related to software or hardware during the training?	No
6	Are you able to design solutions for problems related to public health, safety, cultural, societal, and environmental and the impact on sustainable development?	No
7	Have you worked on real time problem/ specific task or any day to day assignment?	Yes
8	Does the training guides you to publish your work?	No
9	Have you used modern tools or Software technologies during the training?	Yes
10	Are you able to apply ethical principles and commit to professional ethics and responsibilities and norms of the engineering practice?	Yes
11	Does training guides you to become entrepreneur?	No
12	Did you get any pre placement offer from the industry or does training helps in the pre placement?	No
13	Does it help to Improve your oral and written communication skills?	Yes
14	Your recommendation for considering this organization for training (or industry institute interaction) in future	5

Signature of the Student

MAHARAJA AGRASEN INSTITUTE OF TECHNOLOGY

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

INDUSTRIAL TRAINING
FEEDBACK FORM

Thank you for supporting our Programme by offering an Industrial Training placement. It represents an important part of our students' professional training. To complete the process, we'd appreciate it if you would complete this short assessment of the student.

Company/division name: AH INFOTECH LLC

Student name: Tanya Goel Period of employment/internship: 23/03/2020 to 04/05/2020

Student's responsibilities: To implement a project on sentiment analyis via twitter for various SONY Music Artists using ML algos

1. **Technical Ability:** ***(Please tick the appropriate box)***

	Excellent	Good	Acceptable	Poor	Very Poor
Knowledge of the field	x				
Quality of Work	x				

2. **Professionalism:** ***(Please tick the appropriate box)***

	Excellent	Good	Acceptable	Poor	Very Poor
Interpersonal skills	x				
Communication skills	x				
Judgment	x				
Punctuality	x				
Attendance	x				

3. **Overall assessment of performance:** ***(Please tick the appropriate box)***

	Excellent	Good	Acceptable	Poor	Very Poor
Overall Performance	x				

4. **Remarks***: (We are particularly interested in any problems you encountered, and how we can better prepare students for their internships.)*

Tanya is a highly motivated and hard working student. She has shown resiliance to challenges faced her way throughout the course of this 6 week internship. Her work ethic is commendable and note worthy. As her mentor, I appreciate her efforts and would be happy to guide her again next year

Name: Viha Gupta *Signature:* 

Contact Number: +1 609 786 0246

You may return the completed form to : Department of Computer Science and Engineering
Maharaja Agrasen Institute of Technology, PSP Area, Rohini Sector 22, Delhi 110086
Email: cse@mait.ac.in

REFERENCES

[1] AH Infotech (www.ahinfotechusa.com)

[2] Sony Music Entertainment (www.sonymusic.com)

[2] GeeksForGeeks (www.geeksforgeeks.com)

[3] Kaggle (www.kaggle.com)

[4] StackOverFlow (www.stackoverflow.com)

[5] Wikipedia

[6] W3 Schools (www.w3schools.com)