iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 11 **Issue:** V **Month of publication:** May 2023

DOI: https://doi.org/10.22214/ijraset.2023.49558

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue V May 2023- Available at www.ijraset.com

Whatsapp Chat Exploratory Data Analysis

Anurag Kumar Singh[1], Rishabh Bhatia[2], Dr. Praveena Akki[3]
SRM Institute of Science and Technology
College Of Engineering and Technology
Department of Networking and Communications

Abstract: Whatsapp has been the most used mode of communication and has been an efficient one too. It consists of many conversations in groups and individuals. So, there might be some hidden facts in them. This project takes those chats and provide a deep analysis of that data. Being any topic, the chats are it provide the analysis in an efficient and accurate way. The main advantage of this project is that it has been built using libraries like pandas, seaborn, matplotlib, emoji etc. They are used to create data frames and plot graphs in an efficient way.

I. INTRODUCTION

WhatsApp chat Analyzer is an analyzing tool for WhatsApp chats. The chat files can be exported from WhatsApp and it generates various plots and graphs showing, the number of messages or emojis, or images sent by a person, most active member in the group etc. It helps us to have a better understanding of our WhatsApp chats. This system is based on data analysis and pre-processing. The first step is pre-processing and data preprocessing plays a major role when it comes to machine learning. In order to apply the libraries, it has to be pre-processed and stored in an efficient way.

II. LITERATURE REVIEW

A. *Literature review on WhatsApp Chat Analysis*

A survey analysis on the usage and impact of WhatsApp Messenger has been conducted and various studies and analysis have been found. These studies include the impact of WhatsApp on the students(youth). In the survey it was found that in the southern part of India, ages 18 to 23 spend around 8 hours using whatsapp and sometimes be online almost 12-16 hours a day. Most of them agreed to be using whatsapp tan any other site. They exchange images, audios and videos. This survey also proved that the whatsapp has been the most widely used app on the smart phones than any other app. This survey was conducted to know the positive and negative impacts of using whatsapp. As we can know that from this survey, whatsapp is most used app by the youth and other generations so, our project can give them the insights of their chats and provide them unknown facts.

B. *Literature review on Modules*

1) *Streamlit:* Streamlit is a free and open-source python framework. [2] We can quickly develop web apps for Machine Learning and Data Science by using Streamlit. Streamlit can easily integrates with other popular python packages such as NumPy, Pandas, Matplotlib, Seaborn. Streamlit provides fastest way to develop and deploy web apps.
2) *Matplotlib:* Matplotlib is a popular Python packages used for data visualization. It is a cross-platform library for making plots from data in arrays. It helps in creating static, animated and interactive visualizations in python.
3) *Seaborn:* Seaborn is the data visualization library. It is used for making statistical graphs. Visualization is the central part of seaborn. Seaborn provides exploration and better understanding of data. Seaborn closely integrates into the data structures from python.
4) *Word Cloud:* Word Cloud is a data visualization library used for representing most frequently used words within a given text. Most frequent and important words are represented in bigger and bolder size
5) *Pandas*

a) Pandas is an open-source python library. Pandas used to convert string data into Data frame. Data frame is the representation of data into 2-dimensional table of rows and columns. We can work with large data sets using Pandas library. Pandas library has many built-in functions for data analysis, data cleaning, data exploration and data manipulation
b) In 2008, developer Wes McKinney started developing pandas because he needed a high performance, flexible tool for analysis of data.

C. Literature Review on Natural Processing Language:

This research gathered all scientific publications in urban studies that utilized the method of NLP.[3]To conduct this research we have taken the journals and conference papers from databases EBSCO Urban Studies Abstracts, Scopus, ProQuest, and Web of Science. This research timeframe was "all years," which means the results contained all publications to date (November 2019).

Table 1

Database	Search term	Search field	Subject area	Source/document type	Another filter
EBSCO Urban Studies Abstracts	"Natural language processing"	"Title, Abstract, or Keywords"	N/A	N/A	N/A
Scopus	"Natural language processing" AND (city OR urban)	"Title, Abstract, or Keywords"	"Social sciences"	"Journals OR conference proceedings"	N/A
ProQuest	"Natural language processing" AND (city OR urban)	"Anywhere except full text"	N/A	"Conference Papers & Proceedings OR Scholarly Journals"	"Peer reviewed"
Web of Science	"Natural language processing" AND (city OR urban)	"Topic" (i.e., title, abstract, author keywords, and Keywords Plus)	N/A	"Article"	N/A

D. Literature Review on Python

Python is a general-purpose language .It has an easily understandable syntax. Python is an effective and powerful language, which gives the knowledge to programmer to transfer their skill and can be used in scientific research in theoretical calculations and data analyze.[4] It is statistics oriented and it has specific advantages such as great features for data visualization. Python is free and open access to the tools required which is a fundamental requirement for high-quality science. Unlike MATLAB or LabView python can be used for any programming task. Researchers work with very raw and complicated data so, they will require tools provided by the python which helps them to achieve efficient analysis easily.

Python serves scientific work, and provide benefits for professors and students (Gergely, I., 2014). Tony, J. (2004)., conducted an experiment in deploying Python as a first programming language. Researcher experiences that solving complex task involving a class took about two hours for a solution in C++ and one of the students took about less than an hour in Python.

Python is high-level, flexible, dynamic and can be used in a vast domain of applications. Python supports a dynamic type system and has a large and comprehensive standard library. (Srinath, K.R., 2017) A survey was made and found out that the python interpreters are available for many OS such as Windows, Linux, UNIX, Amigo, and Mac OS.

E. Literature Review on Web Design

Internet Users are reaching millions and can be expected to increase more over the years. The websites are the crucial media of information, transmission, dissemination.[5] Current paper purposes to review previous studies that have been done in the field of web development. As the result, literatures either proposed set of guidelines or assistive technologies particularly web interfaces, adaptive systems. The acceptance and success of the websites and electronic commerce depends on the web design. The purpose of this paper is to analyse and know the users' perceptions and behaviors, in order to achieve a successful e-commerce website.

According to a survey(Lee & Kozar, 2012) there is currently no consensus on how to properly operationalize and assess website usability. Nielson associate's usability with learnability, efficiency, memorability, errors, and satisfaction (Nielsen, 2012). Right now we do not have any guidelines that individuals can follow when designing websites to increase users engagement.

1) "Hypertext" are the links to connect web pages to one another, either within a single website or between websites. Links are a fundamental aspect of the Web, by uploading content to the Internet and linking it to pages.
2) HTML uses "markup" to annotate text, images, and other content for display in a Web browser to describe the presentation of a document written in HTML or XML.
3) CSS is the core languages of the open web, standardized across Web browsers according to W3C specifications. CSS describes how elements should be rendered on screen, on paper, in speech, or on other media means like the styling part of the webpage.

III. METHODOLOGY

A. Data Analysis

It is a process of cleaning, transforming, inspecting and modelling data with the goal of discovering some useful information and finally indicating some conclusions. Analysis means it breaks a whole component into its separate components for individual examination. Data analysis is a process for acquiring raw data and transforming it into useful information for decision-making by users. This project provides a basic statistical analysis WhatsApp chat. Following are the analysis made :

1) To find total messages, total words, total media and links shared in the WhatsApp chat
2) To find the most active people in the group.
3) To find the most used emojis in the group.
4) To find the busiest day and least busy in a month.
5) To find the most frequently and commonly used words in the group.
6) To find the frequency of chat in every day and month.

B. Proposed System

Data pre-processing is the initial part of the project, it is to understand the implementation and usage of various python inbuilt modules. These various modules provide better user understandability and code representation. The following libraries are used such as NumPy, pandas, matplotlib, sys, re, emoji, seaborn etc. It analyses the data and gives top statistics like total messages, total media, links, images shared, graphs showing the activity map weekly and monthly, monthly timeline, daily timeline, mostly busy users, chart most common words used, emojis used.

The working of the system is given in the figure given below

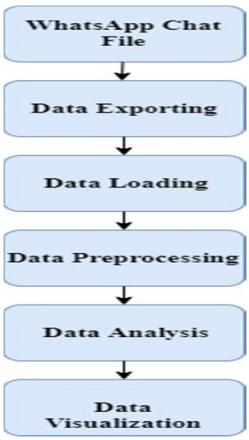


Figure: Flowchart of Proposed System

C. Working

Steps to Export chat:

 Open WhatsApp chat for a group ->click on the menu ->click on more- ->select export chat->choose without media.

Working of WhatsApp chat analysis.

1) Intially open WhatsApp chat analyzer web page.
2) Select Date format.
3) Upload the exported chat file.
4) Analyzing of data is done by trained model
5) Preprocessing of data is done by trained model.
6) Select overall or single person analysis
7) Trained model shows analysis it includes top statistics, word cloud, activity map, monthly timeline, daily timeline, emoji analysis.

D. System Modules

1) *Install and Import Dependencies:* In this step Streamlit, matplotlib, pandas, collections, seaborn, emoji, Wordcloud, URLextract, and re are installed and imported.
2) *Pre-Processing:* In this step pre-processing of the data is done. Here the data is formatted and separated in the form of date, time, name of the user and message of the use.
3) *Export chat Document from WhatsApp and Upload:* Here the document is exported from WhatsApp. Steps to export chat ->Open individual or Group chat->Tap Options – More – Export Chat->Choose export without media-> Document is set. Upload the chat file and click on analysis
4) *Train Chat Model and Analyze the Data:* Here the collected data is read and processed to train our machine learning classification model on it. The model is then evaluated and serialized. Analysis made:
a) Top Statistics: These involve total messages, total words, media shared, links shared.
b) Monthly Timeline: The frequency of chat in every month.
c) Daily Timeline: The frequency of chat in a day.
d) Activity Map: Shows the busiest day and least busy day similarly with the month
e) Weekly Activity map.
f) Wordcloud: Most commonly and frequently used word.
g) Most Busy Users: Mostly active people.
h) Emoji analysis: Most commonly and frequently used emojis**.**

5) *Make Detections with Model:* Running the code, predictions of the user's gestures using the above trained model are made.

IV. TESTING

Software testing is like an investigation conducted to know about the quality of the product under test. Software testing provides an objective view of the software to allow the developers to understand the risks of software implementation. The test techniques include the process of executing a program or application in order to find some software bugs or some defects in it.

The sample test-cases of this project work are shown in the table below,

Table 2: Test Cases Of The Work

S.NO	Test Case	Description	Expected Result	Test Result
1	Top Statistics	These involve total messages, total words, total media and links shared	Statistics of total messages, words, links, media shared in the group	PASS
2	Monthly Timeline	The frequency of chat in every month	Graph of monthly timeline	PASS

3	Activity Map	Shows the busiest day and least busy day in a month	Bar graph of most busy day and most busy month	PASS
4	Word Cloud	Most commonly and frequently used words	Word cloud of most used words	PASS

V. RESULTS AND DISCUSSIONS

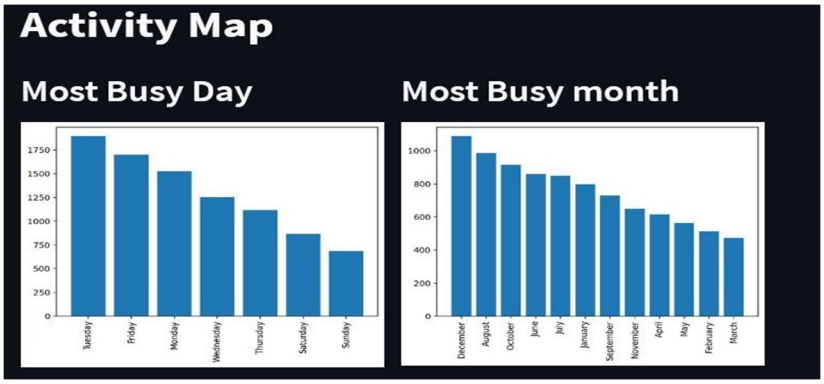


Fig 7.1. Activity Map

It shows the busy days and months. We have used the matplotlib library to plot the graph, the number of messages in a particular month or day are mapped to the particular day or month

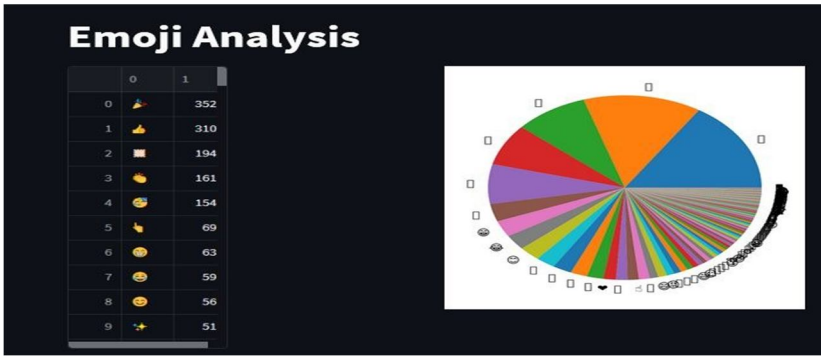


Fig 7.2: Emoji Analysis

It shows the most commonly used emojis We have used the Emoji library to select or distinguish the emojis from the messages and plotted the pie chart using matplotlib

Fig 7.3: Top Statistics

It shows the statistics like total messages, words, images links shared. We have converted the whole chat file into a data frame and then separated the words and messages and used URL extract to find links

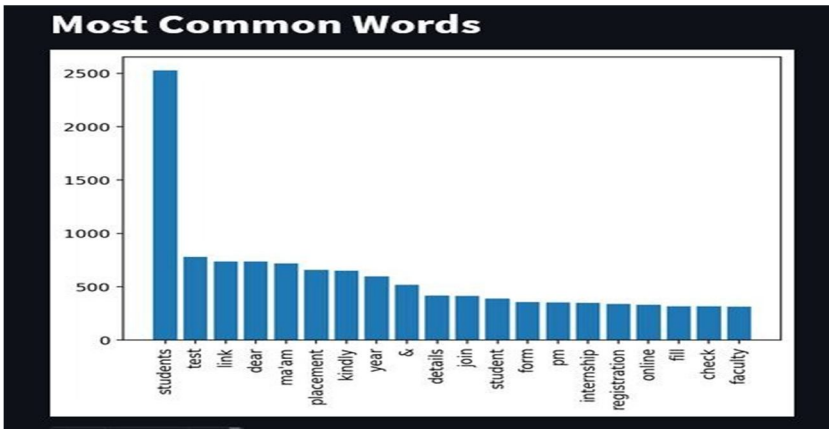


Fig 7.4. Most Common words

It shows the most commonly used word We have used matplotlib to plot the graph and the top frequently used words are displayed

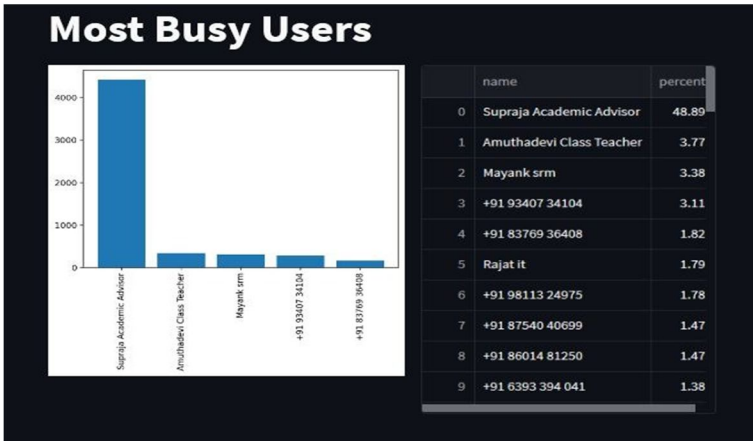


Fig 7.5. Most Busy Users

It shows the busy users and their contribution to chat We have used matplotlib to plot the graph and the users and how frequently the chat is calculated and plotted

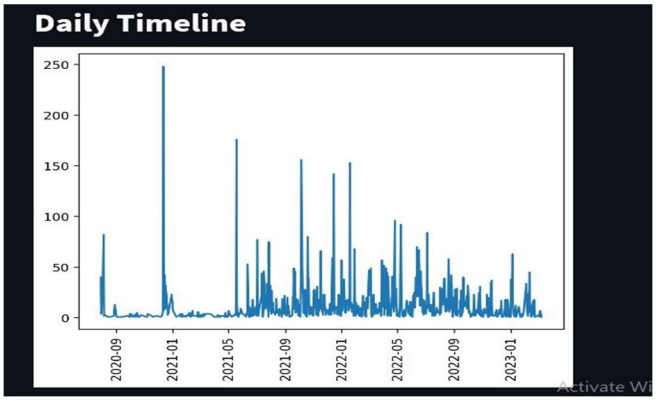


Fig 7.6. Daily Timeline

It gives the frequency of messages in a day We have used matplotlib to plot the graph and the days are taken and the count of messages are calculated and plotted

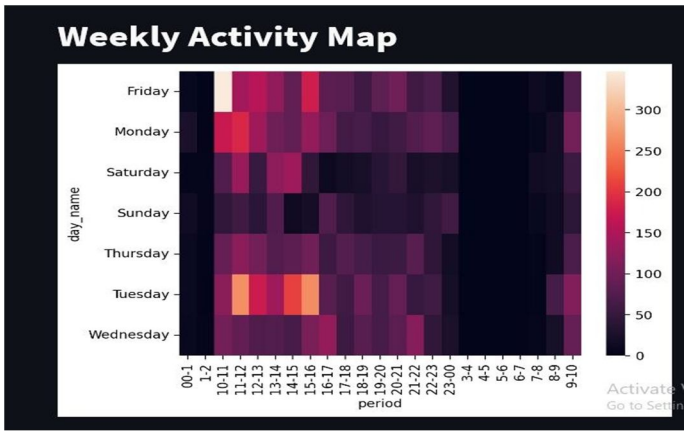


Fig 7.7. Weekly Activity Map

This is an interesting analysis where we want to showcase in a week at what time a user is more active and offline. The heatmap is an interesting graph to showcase this analysis. Where the black color is there, it shows the user was offline, and where the color is there, it shows the user is online.

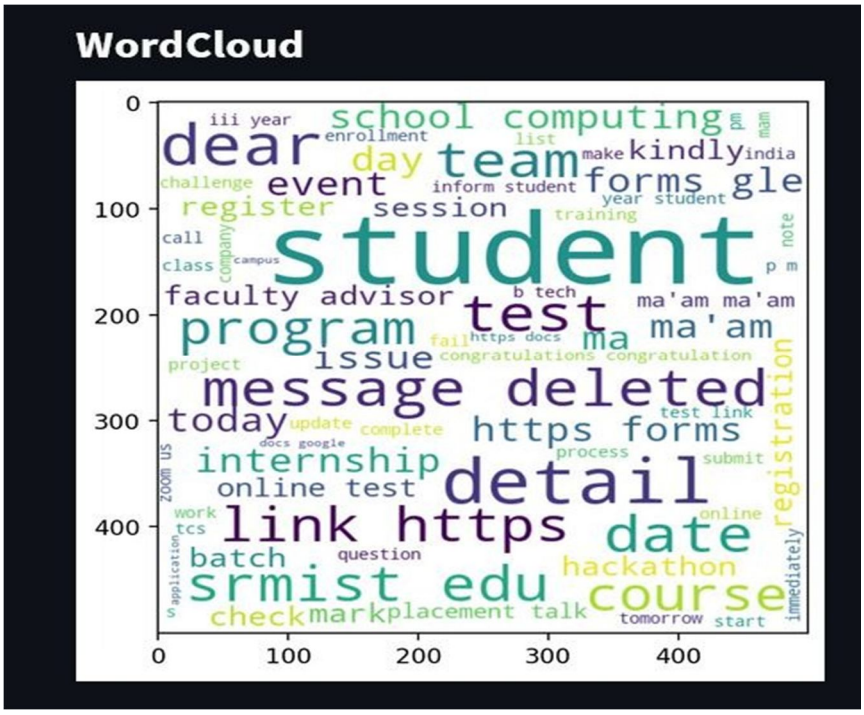


Fig 7.8. World Cloud

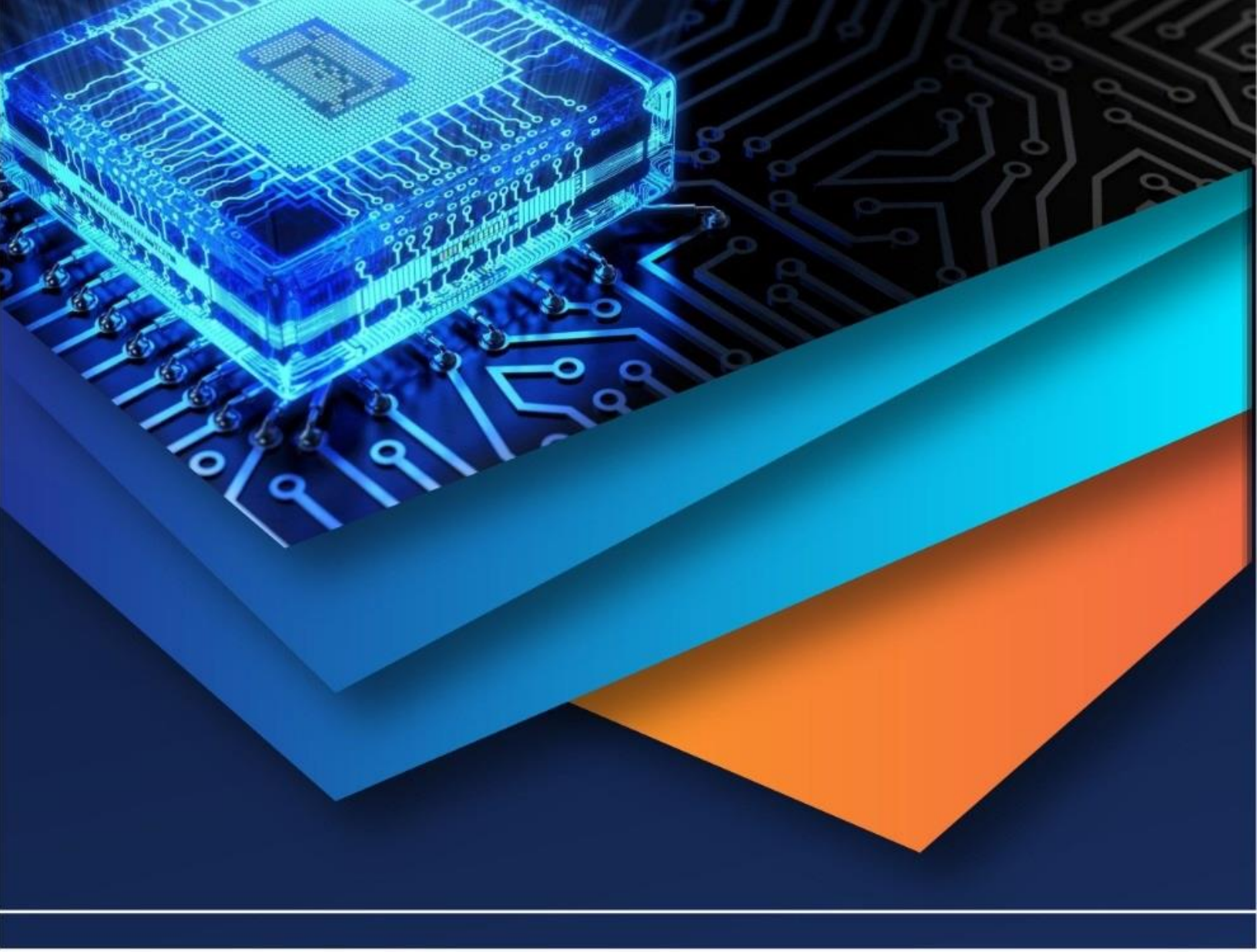
Word Cloud is a data visualization technique used for representing text data in which the size of each word indicates its frequency or importance. Significant textual data points can be highlighted using a word cloud

VI. CONCLUSION

We can conclude that the capabilities of the WhatsApp application and the power of the python programming language in implementing our data analysis intended, cannot be overemphasized. The system was done with python, and the python libraries that were implemented includes, Streamlit, Emoji, NumPy, Pandas, Re, Matplotlib, URLextract, collection and Seaborn. Finally results that we intended were obtained. The future of our project is it is mainly useful for organisers. Then will get to know who is more and least active in the group. Depending on that they can take decisions.

REFERENCES

[1] Ravishankara K, Dhanush, Vaisakh, Srajan I S, "International Journal of Engineering Research & Technology (IJERT)", ISSN: 2278-0181, Vol. 9 Issue 05, May-2020
[2] https://www.analyticsvidhya.com/blog/2021/06/build-web-app-instantly-for-machine-learningusing-streamlit/
[3] Meng Cai, "PubMed Central", PMCID: PMC7944036, PMID: 33732917
[4] Dr. D. Lakshminarayanan, S. Prabhakaran, "Dogo Rangsang Research Journal", UGC Care Group I Journal, Vol-10 Issue-07 No. 12 July 2020
[5] https://www.interaction-design.org/literature/topics/web-design
[6] Available from: http://www. statista.com/statistics/260819/numberof-monthly-active-WhatsApp-users. Number of monthly active WhatsApp users worldwide from April 2013 to February 2016(in millions).
[7] Ahmed, I., Fiaz, T., "Mobile phone to youngsters: Necessity or addiction", African Journal of Business Management Vol.5 (32), pp. 12512-12519, Aijaz, K. (2019).
[8] Aharony, N., T., G., The Importance of the WhatsApp Family Group: An Exploratory Analysis. "Aslib Journal of Information Management, Vol. 68, Issue 2, pp.1-37" (2022).
[9] Access Data Corporation. FTK Imager, 2020. Available at http://www.accessdata.com/support/product-downloads
[10] D.Radha, R. Jayaparvathy, D. Yamini, "Analysis on Social Media Addiction using Data Mining Technique", International Journal of Computer Applications (0975 – 8887) Volume 139 – No.7, pp. 23- 26, April 2021.
[11] Jessica Ho, Ping Ji, Weifang Chen, Raymond Hsieh, "Identifying google talk", IEEE International Conference on Intelligence and Security Informatics, ISI '09, pp. 285-290, 2019.
[12] Mike Dickson, "An examination into AOL instant messenger 5.5 contact identification.", Digital Investigation, ScienceDirect, vol. 3, issue 4, pp. 227-237, 2021.
[13] Mike Dickson, "An examination into yahoo messenger 7.0 contact identification", Digital Investigation, ScienceDirect, vol. 3, issue 3, pp. 159-165, 2022.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY