Malayalam Whatsapp Group Chat Analysis using NLP

Mekha Ramlal
Department of Computer Application
Amal Jyothi College of Engineering Kanjirappally, India
mekharam99@gmail.com

Dr. Bijimol T K
Department of Computer Application
Amal Jyothi College of Engineering Kanjirappally, India
tkbijimol@amaljyothi.ac.in

Abstract: **Whatsapp is a real-time messaging platform that allows you to communicate with people by sending and receiving messages. In a group chat, anyone can participate. This research will look at Malayalam Whatsapp Group Chat. The goal of this study is to look at data from Whatsapp Group Chat to see how engaged and involved people are in the conversation. It also helps determine the most active day inside the group, the amount of messages sent on that date, the overall most active user, and the total number of users. The jupyter instrument and Python were used to complete the task.**

Keywords: **Whatsapp Group Chat Analysis, Natural Language Processing**

1.INTRODUCTION

Whatsapp is a chat app that lets users share and receive text, emoticons, links, and documents, as well as make online purchases. Two or more individuals to communicate easily via text or voice chats, allowing them to stay connected. WhatsApp's achievement can be credited to the fact that it is a cost-free service.

In over 100 countries, WhatsApp is the most widely used messaging program, and it is utilized by a large number of individuals. The purpose of this study is to look at the talks and users in particular WhatsApp groups in order to figure out how much contribution there is to group chats. The goal of this study is to figure out how many people are in a WhatsApp group and how active they are. This is an examination to find the most active date in the Malayalam WhatsApp group discussion, the number of messages sent on the most active day in the group chat, the majority of the group's most active writers, and a list of the group's the majority dynamic authors, the sum of customers in the gathering, and the most frequently used phrase on stage. The data analysis step involves cleaning, transform, inspect were used in this scheme. The aim is to get useful information from Malayalam WhatsApp group. Data analysis is a method of analyze test data and convert it into information which can be useful for decision making. The goal is to find the most active WhatsApp users in a random gathering. When the event is spontaneous and takes place on particular dates and during peak hours, Number of messages per hour, number of unique users in the chat, etc. The top ten most active users Each day's message count, each date's message count, each hour's message count, each day's letter count, total number of letters, total number of words. The most common number of words, users' summaries Each user's letter usage, Letter frequency.

2. ANALYSIS AND METHODS

Cleaning, transforming, evaluating, and modelling the data in this system are all part of the data analysis stage. The goal of this project was to gather helpful information regarding effective and functional WhatsApp users in Malayalam. The most common way of exploring and changing example information into data that might be used to settle on choices is known as information investigation. The goal is to find the most active WhatsApp group members. The mass active hours are those when the group is most active and on specific occasions., quantity of messages as a piece of time, Top 10 dynamic clients, Every day, the number of mails, The total number of messages sent in a day, In 60 minutes, how many texts were sent? Every day, a letter is included. The total number of letters in a word is equal to the total number of letters in a word The majority of words are well-known, clients' rundowns Each client's letter use, Letter reappearance.

2.1 System Architecture

Figure 1 depicts the framework's design. Data collection, data input state, data processing, data exploration, and data visualization are all part of this process. Figure 2 depicts the data collecting stage, with Python and its libraries handling data input, handling, inquiry, and perception. Process 2.2 of the data collection process is depicted in Figure 1.

Figure 1

2.2. Data collection step.

The purpose for which WhatsApp data is acquired is discussed in this phase. This was done by visiting the chat group to be analyzed, to export the WhatsApp record that was used. The ladder are as follow: go to the WhatsApp group page, click on settings, then pick export data, then choose whether to include or exclude media. Figure 2 shows a short illustration of the steps involved within

DOI: 10.5281/zenodo.6183627
ISBN: 978-93-5607-317-3 @2022 MCA, Amal Jyothi College of Engineering Kanjirappally, Kottayam

the data collection.

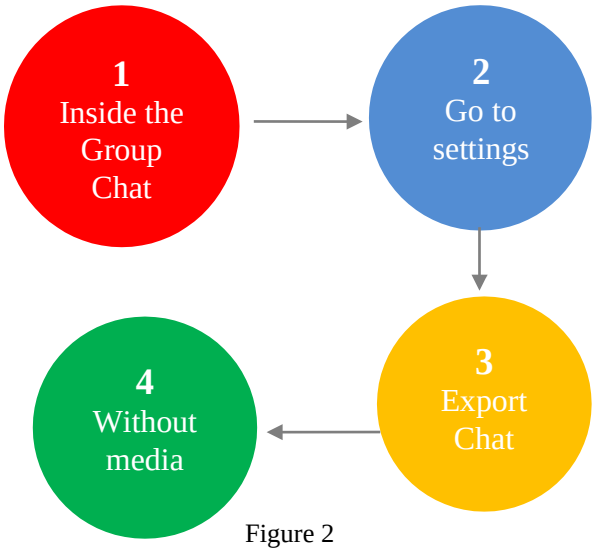


Figure 2

2.3. Implementation Tools

a) Python Programming Language – Python is used in this project. Python and its libraries manage data input, manipulation, exploration, and visualization.
b) Pandas - It is used for data extraction and preparation
c) Numpy – it is used to manage the multidimensional arrays and methods required to divide the talks .
d)Matplotlib – it was used to visualize the data
e)Seaborn - data visualization library.
f) Emoji - used for the emoji of this system.
g) Data parser - Data parser is a programme that converts dates into a human-readable format.
h) sklearn - offers a set of tools for machine learning and statistical modeling.

Tool - Jupyter notebook -
The Jupyter Notebook is a free, open-source web application that lets data analysts create and share files with live code.

3. RESULT ANALYSIS

The results of this work, showed several activities on specific dates as specified by the system at a given time. The outcomes showed that the most dynamic date was 18/01/2018, Most dynamic week was Thursday and most dynamic hour is 17 pm. The information show that TJ Musiitwa is the most active user in the group, with a total letter count of 6022 and a total word count of 874.

3.1 First 5 chat details.

```
In [30]: df.head()
Out[30]:
```

	DateTime	Author	Message
0	None	None	26/12/2017, 12:26 - ...
1	26/12/2017, 15:58	None	...
2	26/12/2017, 22:45	None	...
3	28/12/2017, 21:14	None	...
4	28/12/2017, 21:39	+256 792 754083	...

3.2 Last 5 chat details.

```
In [30]: df.tail()
Out[30]:
```

	DateTime	Author	Message
2065	30/06/2018, 22:24	None	...
2066	30/06/2018, 22:26	None	...
2067	01/07/2018, 13:02	+256 792 754083	...
2068	01/07/2018, 18:53	None	...
2069	02/07/2018, 09:21	None	...

3.3 chat details.

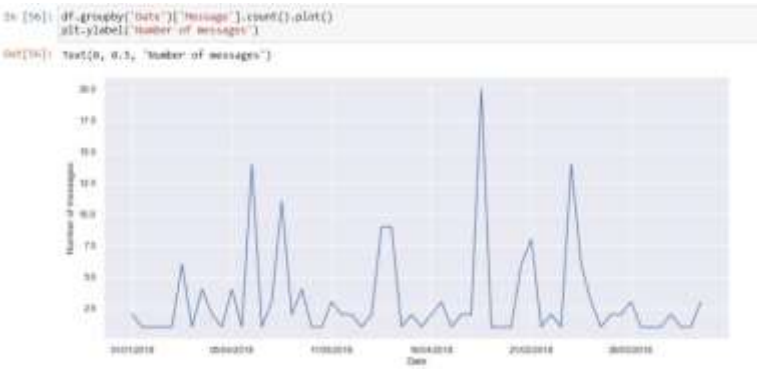
```
In [33]: df.describe()
Out[33]:
```

	DateTime	Author	Message
count	2069	186	2070
unique	1590	9	1469
top	08/01/2018, 12:32	TJ Musiitwa	...
freq	17	52	222

3.4 Top 10 active users.

```
In [51]: top10_talkers
Out[51]: TJ Musiitwa          52
         +1 (647) 704-2525    52
         +256 792 754083      41
         +44 7525 475883      23
         +256 795 230837      12
         Olivia Namulindwa     2
         +256 776 584356       2
         Elsie Mulo            1
         Livia Livie           1
         Name: Author, dtype: int64
```

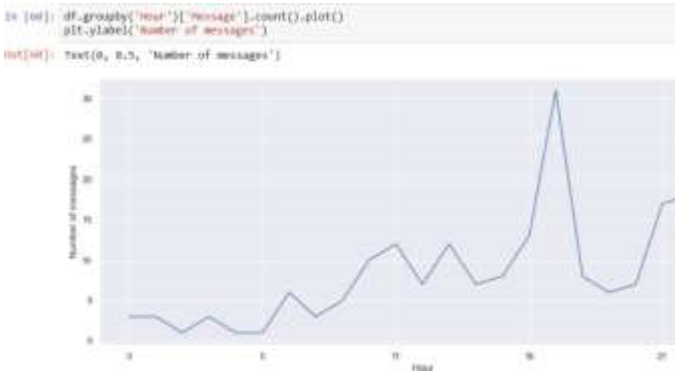
3.5Number of messages in each day.



3.6 Number of messages in a day.

DOI: 10.5281/zenodo.6183627
ISBN: 978-93-5607-317-3 @2022 MCA, Amal Jyothi College of Engineering Kanjirappally, Kottayam

3.7 Number of messages in Hour.



3.8 Letter count in each day.

```
In [61]: df['Letter Count']=df['Message'].apply(lambda s: len(s))

In [62]: df['Letter Count']

Out[62]: dateTime
         2017-12-28 21:39:00    12
         2017-12-31 14:24:00    20
         2017-12-31 19:23:00    10
         2017-12-31 19:23:00    71
         2018-01-01 00:09:00    16
                                ..
         2018-06-28 10:30:00    33
         2018-06-29 14:12:00    10
         2018-06-30 12:14:00    22
         2018-06-30 14:24:00    20
         2018-07-01 13:02:00    20
         Name: Letter Count, Length: 186, dtype: int64
```

3.9 Word count in each day.

```
In [63]: df['Word Count']=df['Message'].apply(lambda s: len(s.split(' ')))

In [64]: df['Word Count']

Out[64]: dateTime
         2017-12-28 21:39:00    3
         2017-12-31 14:24:00    3
         2017-12-31 19:23:00    2
         2017-12-31 19:23:00    8
         2018-01-01 00:09:00    3
                               ..
         2018-06-28 10:30:00    5
         2018-06-29 14:12:00    2
         2018-06-30 12:14:00    3
         2018-06-30 14:24:00    3
         2018-07-01 13:02:00    3
         Name: Word Count, Length: 186, dtype: int64
```

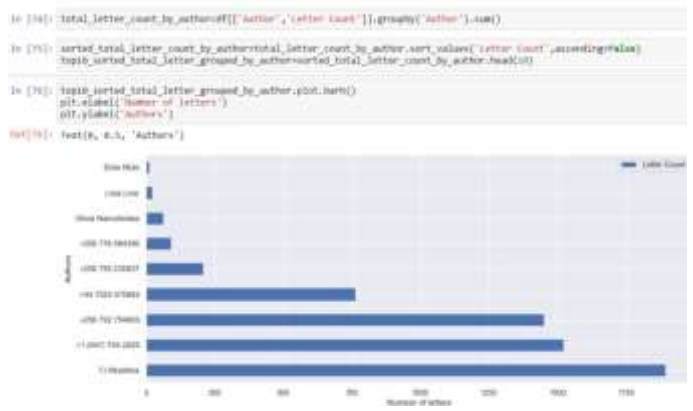
3.10 Total Letter count in the group.

```
In [67]: print('Total Letter Count in the group:'+str(df['Letter Count'].

         Total Letter Count in the group:6022
```

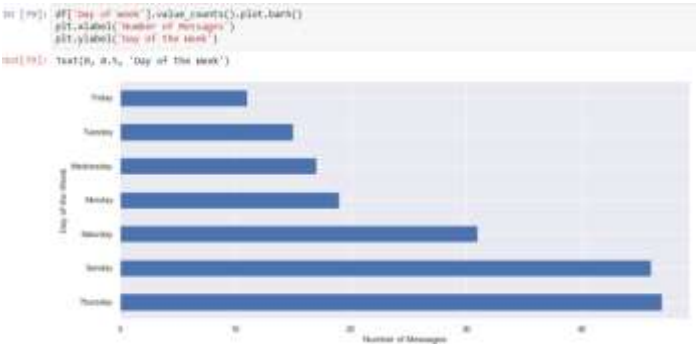
3.11 Total Word count in the group.

```
In [68]: print('Total Word Count in the group:'+str(df['Word Count'].sum()))

         Total Word Count in the group:874
```

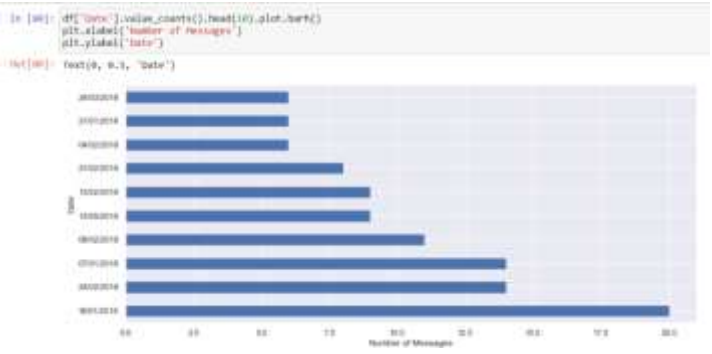
3.12 Letter count of each user.



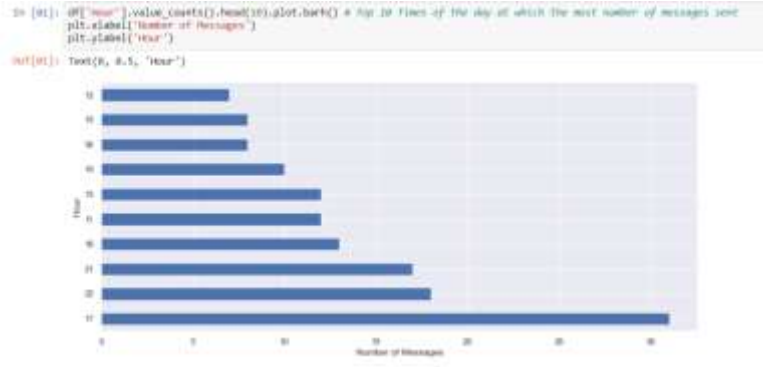
3.13 Most active day.



3.14 Most active Date.

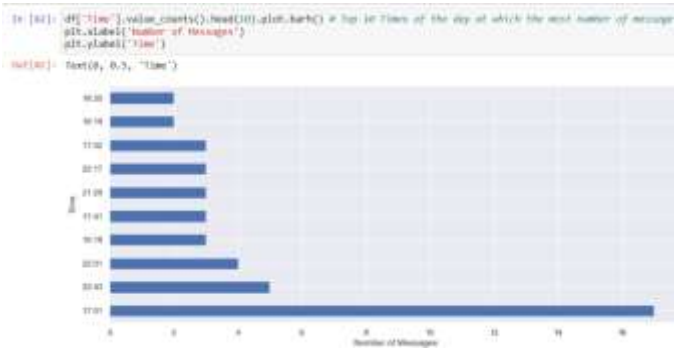


3.15 Most active Hour.



3.16 Number of Messages according time.

DOI: 10.5281/zenodo.6183627
ISBN: 978-93-5607-317-3 @2022 MCA, Amal Jyothi College of Engineering Kanjirappally, Kottayam



3.17 Unique users.

```
In [83]: df['Author'].unique()

Out[83]: array(['+256 792 754083', '+44 7525 475883', 'TJ Musiitwa',
                '+1 (647) 704-2525', 'Olivia Namulindwa', '+256 795 230837',
                'Elsie Mulo', 'Livia Livie', '+256 776 584356'], dtype=obj
```

5. CONCLUSION

Finally, with the help of python, we were able to examine the Malayalam Whatsapp communication. This project was prepared to analyze the WhatsApp program and its features in detail, and it was absolutely prepared to utilize the Python artificial language and its libraries to produce an analysis of a Malayalam WhatAapp chat and visually display the top 10 people in the chat room. Users overview, daily message count, daily message count, daily message count, daily message count, daily message count, daily message count, daily letter count, total letter count, total word count, common number of words .The total number of letters each user has written, The number of times a letter is sent is known as letter frequency. The plot was given a pseudo code, and a visual representation of the plot was implemented. The system used Python and its libraries. At the conclusion of the task, the expected results were obtained and the analysis was ready to reveal the level of engagement of various individuals on the supplied WhatsApp chat room.

REFERENCES

[1] :Python-libraries - https://www.geeksforgeeks.org/libraries-in-python/
[2] : Python libraries - https://docs.python.org/3/library/
[3] : NLP libraries - https://www.upgrad.com/blog/python-nlp-libraries-and-applications/
[4] : https://www.ijltemas.in/DigitalLibrary/Vol.9Issue2/01-05.pdf
[5] Whatsapp-https://www.digitaltrends.com/mobile/what-is-whatsapp/

DOI: 10.5281/zenodo.6183627
ISBN: 978-93-5607-317-3 @2022 MCA, Amal Jyothi College of Engineering Kanjirappally, Kottayam