PROJECT DOCUMENTATION

(FEBRUARY 2024-JULY 2024)

Social Media Analytics and Central Media Dashboard(Telegram)

Submitted by

Mohamed Aslam M(211721104083) Keerthika R (211721104070)





Under the mentorship of

Mr.Solai Murugan V-Joint Director
Center for Development of Advanced Computing
Chennai

Department of Computer Science and Engineering Rajalakshmi Institute of Technology Chennai

July 2024

ABSTRACT

In the age of digital communication, Telegram has emerged as a prominent platform for private and public messaging, creating a need for sophisticated tools to analyse and monitor social media activities. This documentation presents a comprehensive social media analytics and monitoring dashboard specifically designed for Telegram, integrating advanced data analytics with real-time monitoring capabilities to deliver actionable insights into user engagement, content performance, and community demographics. The dashboard features scraping of messages and storing it in MongoDB database, detailed analysis on telegram users and their messages and significant insights on telegram channels.Leveraging MongoDB for robust data storage and Flask for a responsive web interface, the dashboard provides a scalable solution for managing and analysing Telegram channel data. Asynchronous tasks enhance data scraping and processing efficiency, making the tool a valuable asset for administrators and content creators aiming to make data-driven decisions, optimise engagement strategies, and improve overall channel performance and behaviour.

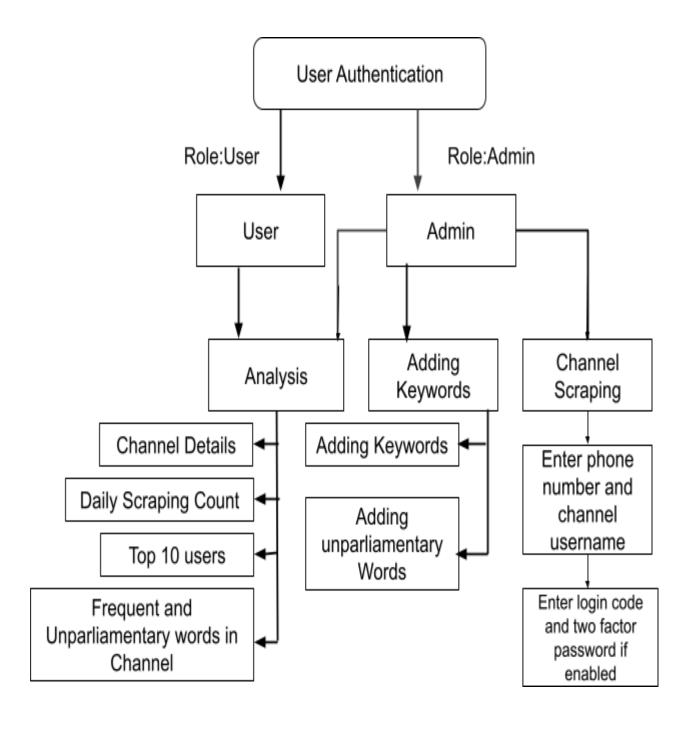
TECHNOLOGIES USED

- 1)HTML
- 2)CSS
- 3)Javascript
- 4)Python
- 4.1) The libraries used: telethon, asyncio, pytz, pymongo, datetime, nltk, os, corpus, flask
- 5)Telegram API ID and API Hash
- 6)MongoDB

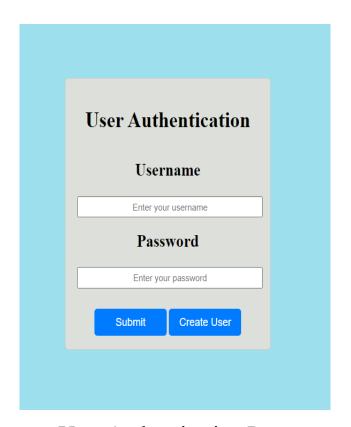
For obtaining telegram API, refer:

https://core.telegram.org/api/obtaining_api_id

WORKFLOW



1) USER AUTHENTICATION

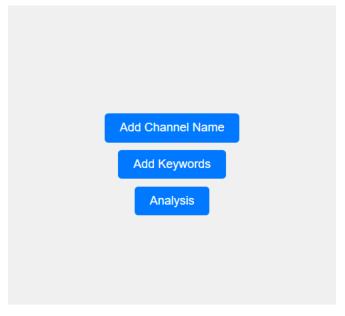


User Authentication Page

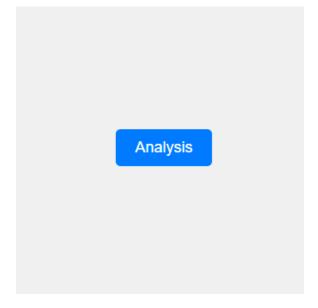
In the first phase, users enter their username, password, and role, which are stored in a MongoDB collection. Roles can be either Admin or User.

If the credentials are valid and the user is an Admin, they are redirected to a page with options to add keywords, add a channel name, or perform analysis. If the user is a User, they are redirected to the same page but only have the analysis option available.

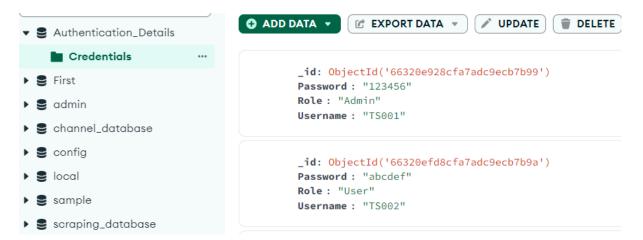
If the credentials are invalid, the user is redirected to re-enter their details. If a user has no credentials, administrators can create them by clicking on the "Create User" button. This action directs the administrator to a page where they can enter and store new user details in the MongoDB collection holding user credentials.



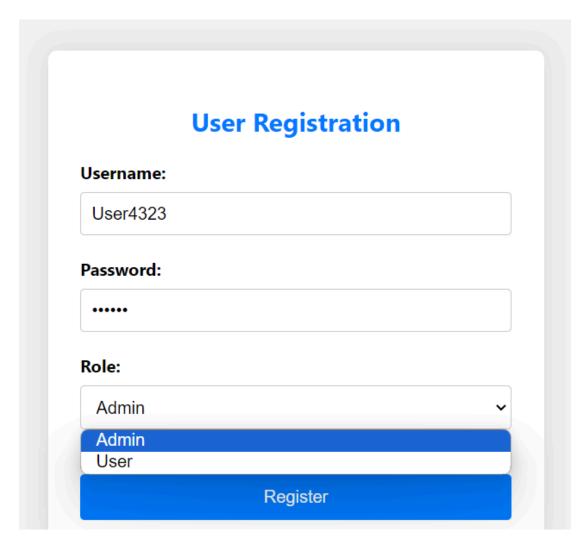
Admin Role



User Role



MongoDB Collection with user credentials



User Creation/Registration

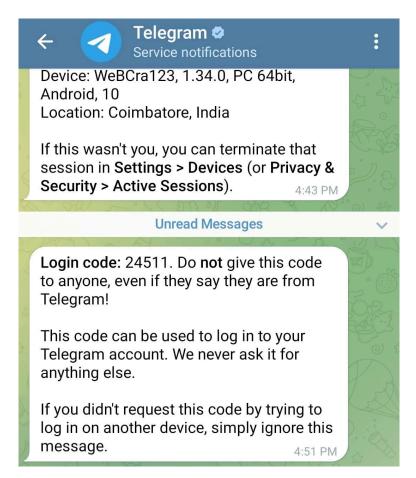
2)SCRAPING OF MESSAGES

Once an admin gets logged in,he can access features to scrape messages,to add keywords and to view detailed analytics. In order to scrape messages, the admin has to click "Add channel name" button where upon clicking it, the admin is redirect to a form where they first enter their phone number or any phone number with international code and channel username. In addition they can also select duration for scraping (for every 5 mins, for every hour etc) and submit the form by clicking send login code button.

Telegram Scraper Enter phone number with international code*: +911234567098 Enter the Channel Name*: killingshop Select Time: Minutes V Select Sub-Time: 5 minutes V Send Login Code Reset

After submitting their form the user gets navigated to a new page where they enter the login code sent to the phone number linked to a telegram account. This is done with the help of Telegram API ID and Telegram API Hash. The admins enter login code and two factor password if enabled and submits form . If the user is validated based on login code and 2FA then scraping starts based

on time duration specified previously till all messages are scraped from the channel.



Login code sent to user

Authentication Details

Submit

Login code will be sent to your Telegram account. Please wait!

Enter two-factor password if enabled.

Enter the login code:	24511
Two Factor Authentication Password:	Enter the 2FA passcode

Once the scraping is completed, the channel information is displayed to the user in a table.

Authentication Details

Login code will be sent to your Telegram account. Please wait!

Enter two-factor password if enabled.

Enter the login code:	Enter	login	COC
-----------------------	-------	-------	-----

Two Factor Authentication Password:

Enter the 2FA passcode

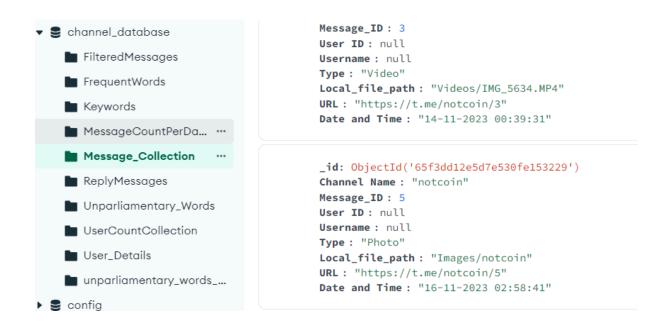
Submit

Details	Data
Channel Name	Sickipedia
Channel Description	The sickest, rudest, most offensive, inappropriate & politically incorrect jokes — we've got them all!
Channel Access Hash	2140466380762987982
Usercounts	13964
Admin Count	None
Kicked Count	None
Banned Count	None
Linked Chat Type	Unknown Type
Linked Chat Title	None
Megagroup	False
Gigagroup	False
Verified	False
Restriction Reason	None

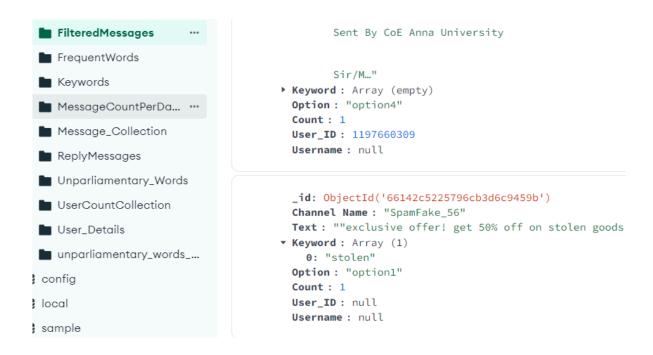
2.1) Message Analysis

During scraping of messages, for each message we check whether the message contains a specified set of keywords where each of them belongs to a category of keywords and unparliamentary words where words of these types are stored in separate collection ,where admin can update them .The message containing these offensive words or unparliamentary words are stored in separate collections.

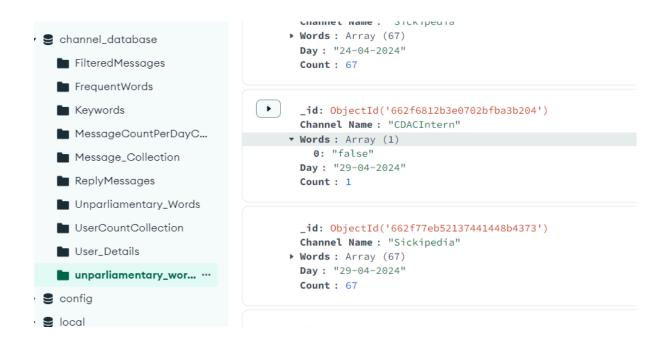
The text messages from each channel are stored in a collection where other types of messages such as video, images and audio in telegram channels are downloaded and stored in separate folders where their local paths and url are stored in the same collection used for text messages.



All the messages from all channels are stored in "Message_Collection". Each message is stored with its message id which helps us to keep track of most recent message scraped from a channel.



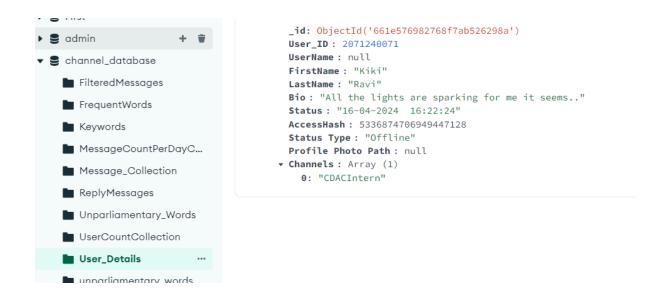
Messages which consist of keywords which are offensive and belong to a category of keywords are stored in "Filtered Messages" collection.



Messages which consist of unparliamentary keywords are stored in a collection known as "unparliamentary words in messages".

2.2) User Analysis

From a message we can track their ID and username if they have made it public. This could be useful to identify more information about the user such as their bio, recent status (when they were online or offline), profile photo, channels which they are a member of etc. For channels which a user is a member of, we track them based on messages they have sent to multiple channels and store each channel in an array. Certain user information requires admin privileges.



The user demographics are stored in "User_Details" collection

Based on the number of channels which a user is a member of,we store the top ten users who are participants of the most number of channels.



Top ten users are stored in "UserCountCollection" collection

If an admin wants to identify whether someone has replied to their messages, or whether action is taken by replying to their messages, this is done by checking whether they have replied to the specific message using that message's id. If yes, the corresponding message and response can be stored in a collection.



Reply to messages are stored in the ReplyMessages collection.

2.3) Channel Analysis

Based on channel name, we can get more information about the channel such as description, user count, linked chats etc. These information could be stored in a collection which can be useful for reference. However, certain information about the channel requires admin privileges.

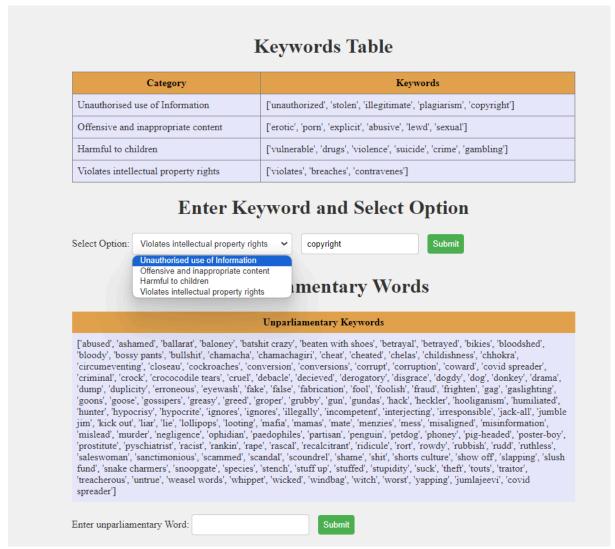
```
_id: ObjectId('6620f2ab0b055a10a737ab03')
Channel Name: "Sickipedia"
Channel Description: "The sickest, rudest, most offensive, inappropriate & politically incor..."
Channel_Access_Hash: 2140466380762987982
Channel_Usename: "Sickipedia"
Start_date: "18-04-2024"
Recent_Date: "28-07-2024"
Number of Users: 13964
Admin_Count: null
Kicked_Count : null
Banned Count: null
Linked Chat Title: null
Linked_Chat_Type: "Unknown Type"
Megagroup: false
Gigagroup: false
Verified: false
Restriction_reasons : null
                                           false
Status: true
Count: 8866
Channel_Username: "Sickipedia"
```

Channel demographics are stored in the "scrapping_status" collection.

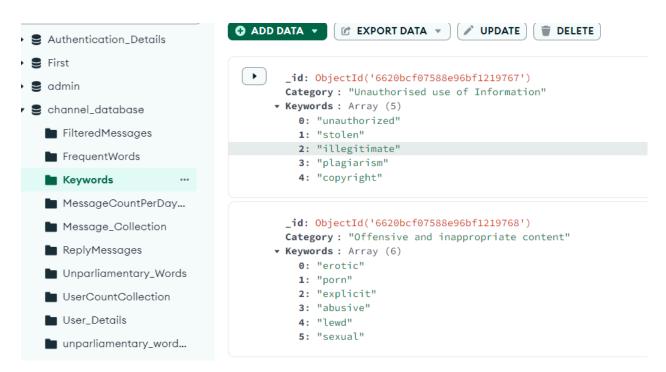
3)ADDING KEYWORDS

If the user logged in as admin,by clicking on "Add Keywords" button,he is navigated to a webpage where the existing keywords of different categories which are offensive and unparliamentary are separately shown in a table where these words are used to

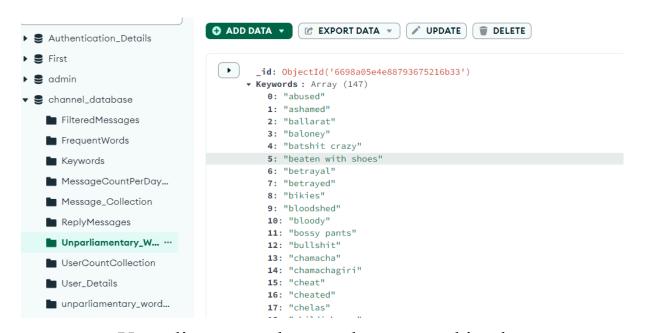
filter messages. In addition, admin can add new words to these categories which are automatically updated to the collections holding them.



Keywords page



Offensive keywords of different categories are stored in the Keywords collection.



Unparliamentary keywords are stored in the "Unparliamentary_words" collection.

4) Analysis Dashboard

Whether the user logged in is an admin or normal user,he/she can access the analysis dashboard by clicking the "Analysis "button. They are various tables which are displayed in the dashboard.

The first table is "Scraping of Channels" where it displays the channel demographics which are stored in "scrapping_status" collection. We can also search for specific channel by entering their name in the search bar and also click on specific field to display the field entries in ascending or descending order.

	Scraping of Channels						
Search for channel							
	Channel Name	Channel Description	Start Date	Recent Date	Number of Users	Status	Messages Scraping Count
	iplbh	आज कोण जितेगा ? आज कौन सी टीम जिथेगी ? और कोई नहीं बतायेगा ? सिर्फ हम बतायेंगे ? . #IPL भविष्यवाणी 🏑	01- 04- 2024	23-04- 2024	29	True	23
	KANYEsol	Excuse me! Did you say something? https://kanyecoin.site	01- 04- 2024	01-04- 2024	585	True	2
	Government_Jobs_Sarkari_Naukari	India's Best Telegram Channel for Free Government Job Alert. Government Jobs Sarkari Naukri [Latest Vacancy Alert for Freshers HR LAW Police Bharti Free Job Alert Employment News]	01- 04- 2024	01-04- 2024	2	True	3
	warmonitors	Breaking News Geopolitics Buy ads: @orthodoxplant Discussion: @WarMonitors_Discussion Appeal a ban: @WarMonitorsAppeal	01- 07- 2024	18-07- 2024	104264	True	32

The next table displays the number of messages which are scraped in a daily basis with channel name and count of messages which are scraped. The search and ordering functionalities of previous are also incorporated here.

Dail Search for channel Search	Daily Scraping Count				
Channel Name ▼	Day	Scraping Count			
x_movies_adult_videos_18_english	03-04-2024	0			
x_movies_adult_videos_18_english	16-04-2024	0			
x_movies_adult_videos_18_english	24-04-2024	0			
taylor	08-04-2024	2			
real_madrid	31-03-2024	18			
real_madrid	01-04-2024	5			
killingshop	07-04-2024	58			
kanyesolxyz	04-07-2024	7			
iplbh	01-04-2024	0			
iplbh	08-04-2024	0			
ipIbh	16-04-2024	0			
iplbh	18-04-2024	0			
iplbh	23-04-2024	0			

The next table displays the top 10 users based on number of channels they are a member of

Top 10 users							
UserID	Channels	ChannelsCount					
5559914153	None	10 Rules to upgrade your life: 1. Lift weights 2. Eat real foods 3. Face your fears 4. Positive self-talk 5. Never stop learning 6. Prioritize your sleep 7. Lower your screentime 8. Wake up with gratitude 9. Enjoy the present moment 10. Choose your circle carefully	['SpamFake_56', 'CDACIntern']	2			
1484373805	None	No Photo Available	['anna_university_cse']	1			
1749283142	None	No Photo Available	['Vikram_Singh_Ranbir_Roy']	1			
1013583367	None	No Photo Available	['anna_university_cse']	1			

The following table displays the filtered messages which are filtered based on keywords in the "Keywords" collection. We can also view them in separate pages with the help of pagination.

	Submit	O Violates int	ellectual p	property rights	O Harmful to chil	dren Offens	sive and inappropriate cont	ent Unauthorised use of
Channel Name	Message	Keyword	Count	User_ID	UserName		The keywords are:	
SpamFake_56	: "get the hottest deals on erotic content! watch free porn videos! explicit offers for adults.	['erotic', 'porn', 'explicit']	1	None	Not Available		erotic pom explicit	
CDACIntern	: "get the hottest deals on erotic content! watch free porn videos! explicit offers for adults.	['erotic', 'porn', 'explicit']	1	5559914153	Not Available		abusive lewd	
Sickipedia	i couldn't believe it when my mum confided in me about how her father sexually abused her as a child. then my sisters told me about being sexually abused by our dad. i got home shocked and was telling my wife about this, and she too told me about being sexually abused as a child by her father. fuck it, i can't wait for my daughter to get home from school, i'm not going to be the one to break a family tradition. #sexandshit	[ˈsexˈ]	1	None	Not Available		sexual	
Sickipedia	the pope calls for action on sex abuse. sex abuse. directed by his holiness is now available to stream on netflix. #sexandshit	['sex']	1	None	Not Available			
Sickipedia	i was once abused by a prolific sex offender, but he wasn't very good at it. #mehtoo #sexandshit	[ˈsexˈ]	1	None	Not Available			

By hovering the cursor on the options, we can view the keywords for each category. By clicking the number below we can view different messages which belong to same category of offensive messages.

The following two tables displays the most frequent words and unparliamentary words in a channel, where in case of frequent words we display each frequent word along with their count and in case of unparliamentary words we display the unparliamentary words, the channel name in which we observed those words and the count of them.

Frequent Words in Channels

Channel Name	Word
CDACIntern	[{'hi': 4}, {'children': 4}, {'sexual': 2}, {'harm': 2}, {'violates': 2}]
notcoin	[{'notcoin': 17}, {'new': 8}, {'fun': 8}, {'beta': 8}, {'get': 7}]
techbossindian	[{'buy': 2}, {'12hrs': 1}, {'uv': 1}, {'speaker': 1}, {'hi': 1}]
Sickipedia	[{'said': 2276}, {'says': 2251}, {'man': 1989}, {'nthe': 1596}, {'one': 1418}, {'said': 2279}, {'says': 2259}, {'man': 1993}, {'one': 1424}]

Unparliamentary Words in Channels

Channel Name	Day	Words	Count
Sickipedia	24- 04- 2024	['shame', 'rubbish', 'corruption', 'slapping', 'fraud', 'bloodshed', 'scandal', 'grubby', 'incompetent', 'murder', 'goose', 'stupidity', 'frighten', 'betrayed', 'betrayal', 'bloody', 'corrupt', 'coward', 'liar', 'humiliated', 'yapping', 'mess', 'stench', 'fool', 'cheated', 'false', 'suck', 'dog', 'gaslighting', 'sanctimonious', 'mate', 'prostitute', 'abused', 'fake', 'shit', 'ashamed', 'penguin', 'species', 'ignores', 'wicked', 'gun', 'cruel', 'donkey', 'racist', 'bullshit', 'paedophiles', 'phoney', 'cockroaches', 'hunter', 'stuffed', 'lie', 'foolish', 'illegally', 'witch', 'mafia', 'mamas', 'traitor', 'cheat', 'rape', 'hack', 'dump', 'disgrace', 'gag', 'theft', 'criminal', 'worst', 'drama']	67
CDACIntern	29- 04- 2024	['false']	1
Sickipedia	29- 04- 2024	['donkey', 'gun', 'cruel', 'paedophiles', 'phoney', 'shit', 'hunter', 'bloodshed', 'stuffed', 'cheated', 'mamas', 'penguin', 'witch', 'bloody', 'corrupt', 'worst', 'ignores', 'shame', 'fraud', 'frighten', 'abused', 'gag', 'lie', 'false', 'rape', 'stench', 'yapping', 'scandal', 'fool', 'rubbish', 'betrayal', 'tratior', 'hack', 'goose', 'ashamed', 'mafia', 'stupidity', 'racist', 'betrayed', 'liar', 'disgrace', 'foolish', 'grubby', 'ashatimonique', 'bullshit', 'acuyard', 'maes', 'dag', 'corruption'	67

The last table in the analysis dashboard displays messages in "ReplyMessages" collection.

Reply to messages Reply Message Message ID Origin Time User ID Message_Type Message Username Action Reply Reply_User_ID Reply_Username Reply_Type Reply_Time 2024-29-04-2024 CDACIntern Ok nice 95 5559914153 Aslam_200356 04-25 Ok nice 103 2071240071 None Text Text 11:50:59 12:06:11 29-04-2024 49 5559914153 Aslam_200356 2071240071 CDACIntern 03-26 Ok aslam None Text Text 11:51:20 19:53:43 2024-Why 29-04-2024 Aslam_200356 5559914153 2071240071 Text CDACIntern 03-14 None Text aslam 11:51:42 aslam 13:42:54 sad 2024-29-04-2024 Aslam_200356 CDACIntern 5559914153 03-14 Mass uh 2071240071 None Text aslam 11:52:21 13:31:37 2024-29-04-2024 Aslam_200356 CDACIntern 5559914153 04-29 2071240071 important 14:52:38 aslam 14:52:12 questions 2024-29-04-2024 112 5559914153 Aslam_200356 2071240071 Text CDACIntern Verified 04-29 Verified None Text 14:53:38

14:53:27