Department of Electronics and Telecommunication Engineering

WhatsApp Chat Analysis

Vaishali Jabade, Bhavit Khandelwal, Om Kakde, Kartikey Dwivedi, Vedant Khandekar, Mandar Kulkarni

***Abstract*-- Social media apps like WhatsApp are indeed a boon to society, but there are some exceptions to it, some groups use them to promote violence and hatred in society. This paper aims at reducing these incidences by monitoring text, images, videos, pdfs, and audio message sent on What's app groups. Various libraries, algorithms and pretrained model are used to achieve the goal. If the questionable messages in a group crosses a threshold, then warning calls are made to the sender. The main aim of this paper is to try to eliminate and break the network of all violence promoting groups.**

Keywords — Data mining, security

I. INTRODUCTION

Data mining has become very important aspect ever since social media has grown in size. It is crucial to monitor all the activities happening on social media sites like Twitter, Instagram, Facebook, WhatsApp, etc. to avoid misinformation and wrong influential messages to float on it. Our project aims at scrutinizing the activities taking place on WhatsApp groups. Nowadays there are lots of riots, terrorist activities, etc. happening due to misuse of WhatsApp. Our project aims at extracting the WhatsApp messages through various algorithms and know about the activities in a particular group. We propose this model to be implemented by cyber security in order to respect the privacy as well.

Project monitors all the aspects of communication media send through WhatsApp that are, text messages, videos, images, documents and also audio messages. If the level of such harmful messages crosses the margin, then we will report such WhatsApp groups to the Cyber security and other Security agencies. Word cloud, scikit learn, regex, speech recognition, etc. are some libraries used for analysis. Extraction of words from images and text from the pdf are some other techniques used. Also we have used Twilio python library which will call up the admin or the authority, if the WhatsApp chat content is found to be more than 70% violent or having malicious content.

II-LITERATURE REVIEW

Kabita Sahoo et al. [1] have discussed that EDA analyses the data in statistical manner which includes measures of central tendency, measures of spread, the shape of the distribution and the existence of outliers. 4 steps of data analysis, 1)Data Exploration, 2)Data Cleaning, 3)Model Building and 4)Present Results are discussed. Graphical Exploratory Data Analysis (GEDA) has also been discussed in the paper. Amazon review dataset which contains reviews of electronic data items has been taken as example. EDA visualize data distributions; bar charts, histograms, box plots. Calculate and visualize correlations (relationships) between variables; heat map. [1]

YupingJin [2] has discussed about word cloud which can be briefly defined as weighted list to visualize language or text data, which gains increasing attention and more application opportunities as the big data time approaches. There are 3 major word cloud maps applied in social networks distinguished by their algorithm instead of appearance, i.e. Frequency, Categorization and Mixed.[2]

Swathikiran Sudhakaran et al. [3] have discussed about the network consists of a series of convolutional layers followed by max pooling operations for extracting discriminant features and convolutional long short memory (convLSTM) for encoding the frame level changes, that characterizes violent scenes, existing in the video. The convolutional neural network along with the convolutional long short term memory is capable of capturing localized spatiotemporal features which enables the analysis of

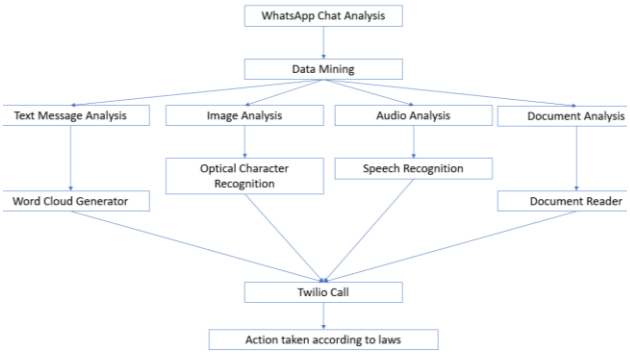
local motion taking place in the video. We also propose to use adjacent frame differences as the input to the model thereby forcing it to encode the changes occurring in the video.

Amal Rekik et al.[4] have discussed about the method violent vocabulary extraction based on a set of annotated dangerous and safe profiles. The active user's identification and taking action against these dangerous profiles. The danger degrees computation of the identified users to extract radical communities.

A. Jović et al. [5] put forward the idea of Data mining (DM) which deals with preparation of data obtained from various information sources (e.g. databases, text files, streams) as well as data modeling using a variety of techniques, depending on the goal that one wants to achieve (e.g. classification, clustering, regression, association rule mining, etc.). Out here we want to target keywords which aren't civilized.[5]

III-METHODOLOGY/EXPERIMENTAL

Flowchart



Text Message Analysis

1. Generating Wordcloud

The first step is to remove filler words like a, an, the, that etc. It is done using the Stopwords function . Wordcloud is generated using the wordcloud library . Author specific word cloud is also generated by enlisting the members in a group and then developing the Wordcloud for all members individually , it is done using pandas library. Links, emoji count message word count are also calculated and stored in a data frame to process the data further.

2. Analyzing links

A lots of video links are shared daily on what's app. Some videos may contain some offensive and violence triggering actions which may cause disturbance in the society. Pretrained model video_mamonreader is used to scan through the links. The links are first downloaded through a Python code and the downloaded file is then automatically fed to the model where the violence index is calculated. Score from 0 to 1 is given and any score above 0.7 is considered high.

Image Analysis

OCR (Optical Character Recognition) is the manner of electronical conversion of Digital photographs into system-encoded textual content. Where the digital photo is typically a photo that carries areas that resemble characters of a language.

For permitting our python application to have Character recognition capabilities, we'd be using pytesseract OCR library.

First we install PIL library and then import it to read the image file then pytesseract function from pytesseract library.

Then we described the path_to_tesseract variable which includes the route to the executable binary (tesseract.exe).

After this, we assigned the pytesseract.tesseract_cmd variable the route saved in path_to_tesseract variable this will be utilized by the library to discover the executable and use it for extraction) of text from the images.

Audio Analysis

Audio messages or recordings is another way to communicate on WhatsApp. As project aims to cover each communication media, audio messages are also monitored. Speech Recognition is the library in Python that has been used to extract words from audio files. Along with it Pydub is the Python library which works with wav files. It also converts mp3 files into wav files as speech recognition is done on wav files.

Document Analysis

Document Analysis plays a very crucial role in this project as documents are the 2nd most shared data type on WhatsApp. Documents hold valuable information like addresses, bills, catalogues, magazines, and scanned images.

To mine data from a document we have used a python library called PyPDF. This library reads all the text present in a document and creates a temporary text file. At the end of the process compiles all the information on it and displays the contents.

Twilio Calling Feature

With the increase in inciting violence through WhatsApp group it becomes inevitable for the authority to keep an eye on such groups. Since, our project works extracting words from the WhatsApp group and then processing it, there is a possibility that it can detect some casual chats as harmful, so to compensate that we have set the limit of 70%.

If the WhatsApp chat content is found to be more than 70% violent or having malicious content, then the system will automatically make a call to the concerned authority informing about the phone numbers present in that group. The calling feature has been added using Twilio library. Adding to it, one can also add SMS feature using the same library.

V – RESULTS

Text Message Analysis

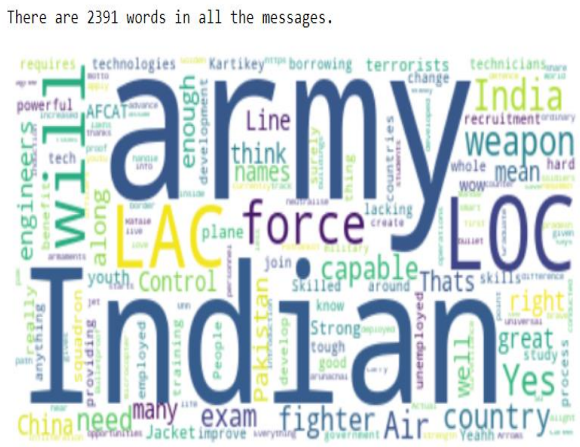


Figure 1 WhatsApp Chat Wordcloud

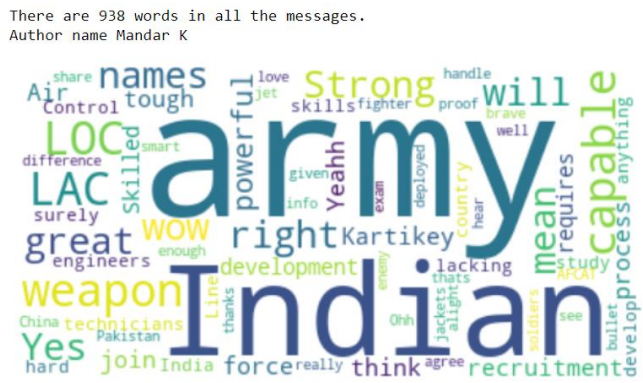


Figure 2 Author wise Wordcloud.

```
res = main_fight('C:/Users/user/EDI_sem2/Toy Guns and Equipments - M4 Rifle - Dese
```

```
(30, 160, 160, 3)
reading video
1622037729849
1622037736913
```

```
res
```

```
{'fight': True, 'percentegeoffight': '0.9728403', 'processing_time': '7064'}
```

Figure 3 Analysis of videos / links sent and percentage of harm in it based on violence content in the video.

Audio Analysis

```
In [2]: from os import path
        import speech_recognition as sr
        from pydub import AudioSegment
        recognizer = sr.Recognizer()

        with sr.AudioFile("audio_file.wav") as source:
            recorded_audio = recognizer.listen(source)
            print("Done recording")
        try:
            print("Recognizing the text")
            text = recognizer.recognize_google(
                    recorded_audio,
                    language="en-US"
                )
            print("Decoded Text : {}".format(text))

        except Exception as ex:
            print(ex)

Done recording
Recognizing the text
Decoded Text : I am a failure sorry I'm sorry for long t
```

Figure 3 Audio to text conversion to analyze audio messages.

Document Analysis

```
In [38]: import PyPDF2
         pdfFileObj = open('ak47.pdf', 'rb')
         pdfReader = PyPDF2.PdfFileReader(pdfFileObj)
         print(pdfReader.numPages)
         pageObj = pdfReader.getPage(0)
         print(pageObj.extractText())
         pdfFileObj.close()

1
Weapon Identi˜cation Sheet
Compiled by the Small Arms Survey with the technical assistance
dom. For further
information or if you have identi˜cation queries, contact weapo
˜cation sheet may be
downloaded in printable PDF format at www.smallarmssurvey.org/w
WARNING:
 Make a full and informed appraisal of the local security situa
light weapons.
Kalashnikov AK-47
 (& close derivatives)
TYPE:
 ASSAULT RIFLE
ALSO REFERRED TO AS:
 KALASHNIKOV/AK
Model illustrated: AK-47, Bulgaria
Technical Information
Calibre
7.62 x 39 mm M1943Length
```

Figure 5 PDF to text conversion to analyze text.

Twilio Calling Feature

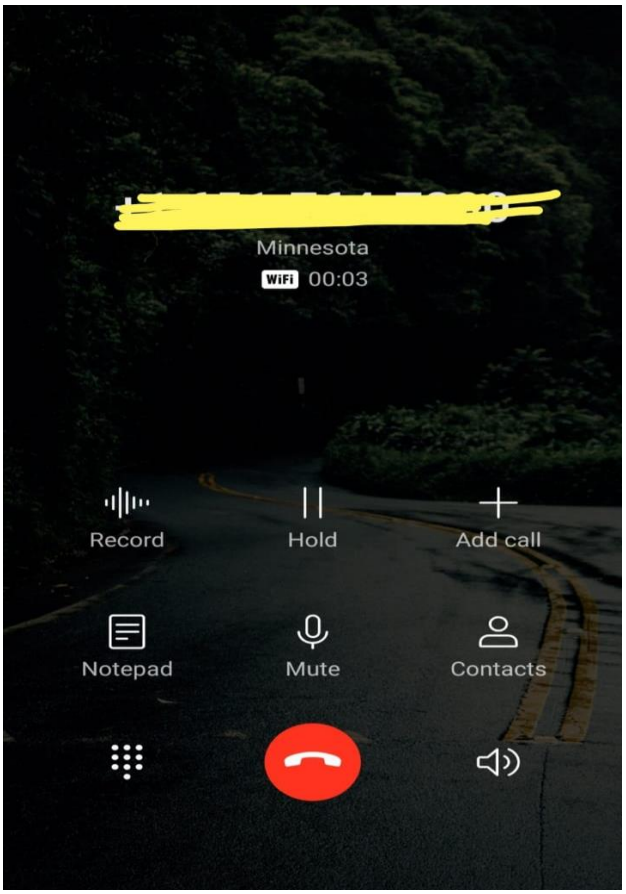


Figure 4 Warning call to the sender if offensive content is found.

VI – ADVANTAGES

1. It creates a user specific wordcloud which tells us the frequency and the unfriendly words used by a specific user

2. It analyzes images which helps in detecting violence if any.

3. It can read contents shared in forms of links and analyzes the content present in the link.

4. Document analysis helps in analyzing the contents of the PDFs to detect violent content.

5. Implementation of a warning feature which gives calls the user if he/she has shared a significant amount of violence-related content.

VII - LIMITATIONS

1. Our system faces issues whenever it is trying to read deleted messages. If users delete their messages, it becomes very hard to track.

2. This project requires the approval of WhatsApp and government, which will help us enable this and read group chats when reported.

VIII-FUTURE SCOPE

1. This system can be implemented on other social media sites like Facebook, Twitter, and Instagram to name a few.

2. After approval we can use this system for protecting our country from riots, terrorist attacks and all types of criminal activities.

IX-CONCLUSION

The spread of false and riot inciting news through WhatsApp is always a concern for the government. We have provided the solution for this problem through our project, obviously the privacy is still a matter of concern which we have to look upon. The methods for checking the malicious activities in our dummy WhatsApp group has given us satisfactory results. The project can be further improved if we work upon the limitations provided.

X-ACKNOWLEDGMENT

Our sincere appreciation goes to Prof Vaishali Jabade our guide and staff members of the Department of Electronics and Telecommunication, Vishwakarma Institute of Technology, Pune for the successful completion of this project.

XI- REFERENCES

[1] Kabita Sahoo, Abhaya Kumar Samal, Jitendra Pramanik, Subhendu Kumar Pani , “Exploratory Data Analysis using Python”, International Journal of Innovative Technology and Exploring Engineering (IJITEE)ISSN: 2278-3075,Volume-8, Issue-12,October 2019

[2] Jin, Y. (2017). Development of Word Cloud Generator Software Based on Python. Procedia Engineering, 174, 788–792. doi:10.1016/j.proeng.2017.01.223

[3] Sudhakaran, S., & Lanz, O. (2017). Learning to detect violent videos using convolutional long short-term memory. 2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS). doi:10.1109/avss.2017.8078468

[4] Amal Rekik, Salma Jamoussia, Abdelmajid Ben Hamadoua, A recursive methodology for radical communities’ detection on social networks, 24th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems

[5] Stancin, I., & Jovic, A. (2019). An overview and comparison of free Python libraries for data mining and big data analysis. 2019 42nd International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO). doi:10.23919/mipro.2019.8757088.