CHAT SYSTEM WITH CONTENT FILTERING

Mohammadi Shaikh, Suwaiba Sayyed, Gauri Kolhe, Sumit Mandage

Student of Final Year Information Technology
MET’s Institute of Engineering, Nashik, India,

Abstract- Social media refers to the means of interactions among people in which they create, share, and/or exchange information and ideas in virtual communities and networks. Social media is about conversations, community, connecting with the audience and building relationships. In today’s world social media has been an integral part of modern life. Starting from children to adults everybody is using various social media for several purposes such as Marketing, Entertainment and Content Creation etc. Everything has its pros and cons so does social media, on one hand, social media has been a really helpful as well as important part of our lives which keeps us updated but on the other hand social media has some serious issues, and one of them is spreading harmful, vulgar or offensive content which may create violence in society. Now, these problems have been considered and we have come up with a solution and developed an android-based application named "CHAT SYSTEM WITH CONTENT FILTERING" in which if a sender sends some message to the receiver and if any inappropriate word is detected in then our algorithms will control such messages by sending them back to the sender with an alert box 'This message cannot be forwarded further'. If the user tries to send the same message more than three times then the user will get blocked for 24 hours. Because of that security of our system will be increased. The best security features are assured for all activities and this application will make sure that your data is safe from all types of threats.

Keywords: Vulgar, Harmful, Offensive, Social Media, Android application.

I. INTRODUCTION
Social media refers to the means of interactions among people in which they create, share, and/or exchange information and ideas in virtual communities and networks. Social media is about conversations, community, connecting with the audience and building relationships. It is not just a broadcast channel or a sales and marketing tool. Social media not only allows you to hear what people say about you, but enables you to respond. Listen first, speak second.

In today’s world social media has been an integral part of modern life. Starting from children to adults everybody is using various social medias for several purposes such as Marketing, Entertainment, Content Creation etc. Everything has its pros and cons so does social media, on one hand social media has been really helpful as well as important part of our lives which keeps us updated but on other hand social medias has some serious issues, and one of them is SPAM.

SPAM could be Hate Speech, Symbols, Bullying, Harassments, Violence, Dangerous Organization, Fraud, Nudity. Now, these problems have been considered and we have come up with a solution and developed an android based application named ‘CHAT SYSTEM WITH CONTENT FILTERING’.

In our ‘CHAT SYSTEM WITH CONTENT FILTERING’ application, we aim to provide splendid modules which are beneficial for user and their privacy. Our application mainly focuses on filtering the spam in 3 main categories i.e.: “TEXT”, “IMAGE” and “URL”.

For TEXT: if a sender tries to send some message to the receiver and if any spam (i.e.: Hate Speech, Symbols, Bullying, Harassments, Violence, Dangerous Organization, Fraud, Nudity) is present in that message then our algorithms will detect such spam and will control them by not sending that particular message further to the receiver and it will display an alert box with the message “Invalid Content”.

For IMAGE: if a sender tries to send an image which contains spam (i.e.: Hate Speech, Symbols, Bullying, Harassments, Violence, Dangerous Organization, Fraud, Nudity) to the receiver, in that case our algorithms will detect the spam and will control them by not sending that particular image further to the receiver and it will display an alert box with the message “Invalid Content”.

For URL: if a sender tries to send some URL to the receiver and if that particular URL contains any type of spam (i.e.: Hate Speech, Symbols, Bullying, Harassments, Violence, Dangerous Organization, Fraud, Nudity), in that case our algorithms will detect such spam and will control them by not sending that URL further to the receiver and it will display an alert box with the message “Invalid Content”.

As our application is all about automatic actions hence it provides another favorable aspect that is, the communication between two or more users is completely secured with help of encryption. If any person opens our application, the last message sent to any contact from chats will be displayed in encrypted format and also the communication between the users are displayed in encrypted format even at the backend. For reading the entire communication user will have to open the specific chat of any user so that he/she will be able to view the messages. Another exciting aspect is that, there is no room for screenshots in our application which means user cannot take screenshots which is an incredible feature. The best security features are assured for all the activities and this application will make sure that your data is safe from all type of threats.

II. PURPOSE
The online social networking sites are becoming the significant tools and are providing a common medium for number of users to communicate with each other. Everything has its pros and cons so does social media, on one hand social media has been really helpful as well as important part of our lives but on other hand social medias has some serious issues that are sharing of Objectionable, inappropriate, or illegal content which can distract a person. The motive is often to stop access to content that the
user might determine is objectionable. The purpose of our project is to build healthy environment on social media, to reduce inappropriate data, it can often be a strong step towards providing security to any and all users, by detecting unwanted contents we can prevent spam messages from creeping into the user’s inbox.

III. MOTIVATION
The reason to do this is simple: by detecting unsolicited and unwanted content, we can prevent spam (i.e.: Hate Speech, Symbols, Bullying, Harassments, Violence, Dangerous Organization, Fraud, Nudity) content from creeping into the user’s inbox, thereby improving user experience. Motivation behind making this project is that in our application, user can indisputably and effortlessly share their information and data which will be protected from all kind of offensive and vulgar words.

IV. OBJECTIVE
1. How to use machine learning techniques for text/keywords detection.
2. How to use Image Classification Algorithms in a Project.
3. To modify machine learning algorithm in computer system settings.
4. To leverage modified machine learning algorithm in knowledge analysis software.
5. To test the machine learning algorithm real data from machine learning data repository.
6. The aim of the project is to detect the offensive and vulgar words from messages which create negative impact on society.

V. LITRATURE SURVEY
1. Srinivasan et al. [1] present the effect of word embedding in deep learning for email spam detection, the proposed method performed better compared to other classical representation methods.
2. Egozi et al. [2] tried to approve the effectiveness of applying NLP techniques to detect phishing messages by processing the samples content and extract features focused on word counts, stop word counts, punctuation counts, and uniqueness factors. The 26 extracted features were used to train an ensemble learning model based on linear kernel SVM and it was able to successfully identify over 80 percent.
3. Seth et al. [3] propose a hybrid CNN model analyzing both the textual and image content of the message to classify it into spam or ham. Their model achieves a high accuracy of 98.87 percent. 4. Bibi et al. [4] propose a comparative study for previous spam filtering systems in terms of accuracy and dataset used.
4. The author [S.K Tuteja] (2016) [5] has worked with different machine learning algorithms for email classification such as Neural Network (NN), Support Vector Machine (SVM), J48 Decision Tree based classifier, NaA˘Zve Bayes. The dataset used by the author was Spam Base dataset. In this paper work, the author didn’t mention advantages and disadvantages of any algorithm.
5. Since tokenization is one of the first steps in any Information Retrieval or Natural Language Processing system, the importance of using a tokenization algorithm is highlighted in early studies [6].
6. The prevalent tokenization algorithms in the literature, Byte Pair Encoding (BPE) [7] and WordPiece [7] are of recent interest in language model pre training research. Many noteworthy studies in the literature focus on enhancing these sub word tokenization methods.
7. Explore the impact of the number of BPE merges on the machine translation performance. [8] propose a drop-out method for each merge step of BPE in order to break the deterministic nature of BPE, which provides a performance improvement in machine translation.
8. The aim of this task is to determine whether a given text sequence includes hate speech towards other individuals or communities with different backgrounds. Hate Speech Detection is a challenging problem with a limited number of resources in the literature, since there is no decisive consensus on the definitions of the hate or offensive speech, and hate language can have various forms in natural language. In this study, we use a recent hate speech dataset in Turkish, curated by [9].
9. They compared Support Vector Machine (SVM) and Decision Tree for email filtering. [A.S Yuksell] [S.F. Cankaya] [I.S. Uncu] (2017) [10] The given dataset was divided into training set and testing set. Each of the model gets trained separately and based on its training, its accuracy is measured. The author made use of supervised learning for both the algorithms and obtained an accuracy of 92 percent on decision.

VI. SYSTEM ARCHITECTURE
In proposed system first user will register or login through mobile number then list of the recent user or as per our contact list dashboard is displayed. After that user will select other user to chat or communicate. User can send message in the form of Text, Image and URL. If user try to send the text message which contain any vulgar and offensive word then text message cannot be send. If the user try to send the Image which contains dirty content then it cannot be send same or URL message as well. We can also provide the encryption security because of which unknown user cannot see the message from outside the chat window. All the vulgar word and message and all other details can store on firebase cloud. By using pattern matching algorithm or technique we can detect the inappropriate word.
VII. SYSTEM IMPLEMENTATION

An Overview of Project Modules

The proposed system consists of four modules containing:

1. Application Module:
An application module represents the data model that your client uses. An application module provides a container for your app’s source code, resource files, and app level settings. It encapsulates the data model associated with a task, plus the custom code to implement the task. To create the data model, the application module contains named instances of view objects and view links. Each component in your business logic tier can have Java code associated with it, to perform a particular role.

2. Stop word removal:
Stop words are the frequently used word in sentence like preposition the articles (eg. as, are, the, of, an, a, this, that, if, he). Such words are removed for next step of processing text message as there is nothing inappropriate in this words. And to lower the processing time of next processing step,

3. Pattern matching:
Pattern matching is basically string searching algorithm. In this we will have to search the inappropriate words in users message. And if found restricting that message and also showing user using toast message that message you are trying to send is inappropriate.

B Tools and Technologies Used

- **Android Studio**
  Android Studio is the official Integrated Development Environment (IDE) for Android app development. Android Studio provides a unified environment where you can build apps for Android phones, tablets, Android Wear, Android TV, and Android Auto. Structured code modules allow you to divide your project into units of functionality that you can independently build, test, and debug. A unified environment where you can develop for all Android devices. The official language for Android development is Java. Large parts of Android are written in Java and its APIs are designed to be called primarily from Java. Android Studio provides the fastest tools for building apps on every type of Android device.

- **Firebase**
  The Firebase Realtime Database is a cloud-hosted database in which data is stored as JSON. The data is synchronized in real-time to every connected client. All of our clients share one Realtime Database instances and automatically receive updates with the newest data, when we build cross-platform applications with our iOS, and JavaScript SDKs. The Firebase Realtime Database is a NoSQL database from which we can store and sync the data between our users in real-time. It is a big JSON object which the developers can manage in real-time. By using a single API, the Firebase database provides the application with the current value of the data and updates to that data. Real-time syncing makes it easy for our users to access their data from any device, be it web or mobile.

- **JavaScript**
  JavaScript is a text-based programming language used both on the client-side and server-side that allows you to make web pages interactive. Where HTML and CSS are languages that give structure and style to web pages, JavaScript gives web pages interactive elements that engage a user. It is an interpreted programming language with object-oriented capabilities.
Imagga API

This API makes use of machine learning technology and using it, it simply assign the tags to the images based on the content inside the images. So this tags will be checked with the words in the database of the words we have stored. Basically this words are indication of inappropriate content. Example, if the image gives the tags like gun and gun is the word we have stored in database. So Image will not be sent to another end. But is image tags are not matched with the words in the database then it’s the indication that image is appropriate and safe to send to another end.

XML

Extensible Markup Language (XML) is a markup language that defines a set of rules for encoding documents in a format that is both human-readable and machine-readable. The design goals of XML emphasize simplicity, generality, and usability across the Internet. It is a textual data format with strong support via Unicode for different human languages. Although the design of XML focuses on documents, the language is widely used for the representation of arbitrary data structures such as those used in web services.

C Algorithm Details

The algorithms used in this project are as follows:

Stop word removal

1. Step 1: The target document text is tokenized and individual words are stored in array.
2. Step 2: A single stop word is read from stopword list.
3. Step 3: The stop word is compared to target text in form of array using sequential search technique.
4. Step 4: If it matches, the word in array is removed, and the comparison is continued till length of array
5. Step 5: After removal of stopword completely, another stopword is read from stopword list and again algorithm follows step 2. The algorithm runs continuously until all the stopwords are compared.

Pattern Matching

Pattern matching is basically string searching algorithm. In this we will have to search the inappropriate words in user’s message. And if found restricting that message and also showing user using toast message that message you are trying to send is inappropriate.

1. Step 1: Take the word from the list of inappropriate words and check whether it is contained in the message string.
2. Step 2: If it contained that word restricting the message from sending and letting user know about it by showing toast message.
3. Step 3: If not message will be encrypted and sent.

Advanced Encryption Standard

1. Step 1: Derive the set of round keys from the cipher key.
2. Step 2: Initialize the state array with the block data (plaintext).
3. Step 3: Add the initial round key to the starting state array.
5. Step 5: Perform the tenth and final round of state manipulation.
6. Step 6: Copy the final state array out as the encrypted data.

The encryption process uses a set of specially derived keys called round keys. These are applied, along with other operations, on an array of data that holds exactly one block of data; the data to be encrypted. This array we call the state array.

VII. RESULT ANALYSIS

The output of the proposed system consist of a report. The report format is designed as per the requirement of the organisation. The report will contain the basic information of the patient followed by the findings for left breast and right breast. After findings report will display the output of neural network classifier. This report will also contain some suggestions given by the radiologist. The following report will contain anomaly highlighted images for all the views [Front view, right lateral, left lateral, right oblique, left oblique]. Followed by edge detected images. In the end the report will contain the history meter and diagnostic meter readings.
XI. CONCLUSION

Spam is a serious issue that is not just annoying to the end-users, but also financially damaging and a security risk. Message containing vulgar words or images will not be sent from the messenger apps. Without spam filtering and security people can lack interest and loose trust on such applications, our application is secured. Also, this pandemic resulted to lots of cyber crimes through social media which would be avoided.

X. FUTURE WORK

Apart from the promising results presented, there are different aspects for future research. As a future work, group chatting feature will be added. Even in group chatting spam would be detected to restrict the spam messages getting viral. Or in group chat one should be targeted of trolls.

REFERENCES: