Professional Chat Application based on Natural Language Processing

Karthick S[1], R John Victor[2], Manikandan S[3], Bhargavi Goswami[4]
Department of Computer Science, Christ University,
Bengaluru, Karnataka, India.
Email: karthick.sowndarajan@gmail.com[1], johnvictor2406@gmail.com[2],
manikandanshan42@gmail.com[3], bhargavigoswami@gmail.com[4]

***Abstract*— There has been an emerging trend of a vast number of chat applications which are present in the recent years to help people to connect with each other across different mediums, like Hike, WhatsApp, Telegram, etc. The proposed network-based android chat application used for chatting purpose with remote clients or users connected to the internet, and it will not let the user send inappropriate messages. This paper proposes the mechanism of creating professional chat application that will not permit the user to send inappropriate or improper messages to the participants by incorporating base level implementation of natural language processing (NLP). Before sending the messages to the user, the typed message evaluated to find any inappropriate terms in the message that may include vulgar words, etc., using natural language processing. The user can build an own dictionary which contains vulgar or irrelevant terms. After pre-processing steps of removal of punctuations, numbers, conversion of text to lower case and NLP concepts of removing stop words, stemming, tokenization, named entity recognition and parts of speech tagging, it gives keywords from the user typed message. These derived keywords compared with the terms in the dictionary to analyze the sentiment of the message. If the context of the message is negative, then the user not permitted to send the message.**

Keywords*— Android Application, Chatting, Dictionary, Named Entity Recognition, Natural Language Processing, Networking, Parts of Speech tagging, Sentimental Analysis, Stemming, Tokenization.

I. INTRODUCTION

Online chatting refers to the process of sending and receiving messages using the internet. There are various chatting applications available in the market. At the first quarter of 2017, the total number of users using chat applications are more than 5.03 Billion [1], and widely used apps are WhatsApp, Facebook Messenger, We Chat, QQ Mobile, etc., All these applications provide various features to ensure security, integrity, and consistency. All these apps let the user send any messages, and the messages can be lewd or inappropriate. There are many cases filed for sending lewd or inappropriate messages in various online mediums [2,3,4,5,6]. It may also be possible for the user to send inappropriate messages by mistake. According to Section 66A of the Information Technology (Amendment) Act, 2008 says that transmitting of obscene information using a transmission equipment which may result in three years of incarceration including a fine [7,8]. In order to solve these concerns, the proposed mechanism implemented. The proposed network-based android chat application used for chatting purpose with remote clients or users connected to the internet, and it will not let the user send inappropriate messages. The application developed for Android, because Android is one of the most widely used mobile operating systems having major market share when compared to other mobile operating systems like iOS, Windows and Blackberry [9,10].

The rest of the paper organized as Section II approaches related work. Section III outlines the proposed methodology. Section IV explains implementation and conclusion of the paper.

II. RELATED WORK

A. Enhanced Education Chat Application Based Chat Application Based on Interested Keyword with Username and Password Authentication Security

This paper states that since there is a lot of changes which had occurred in the recent past and the customers are changing their needs from time to time, it has become a necessity to build a chat application that matches the present crowd. This system was mainly developed to provide a chat room for all kind of educationalist by offering them a platform to chat and solve study related queries [11].

B. Design of Chatting Application Based on Android Bluetooth

This paper proposes a method of sending and receiving text messages using the Bluetooth connection. Since the Bluetooth technology consumes low cost and power, they incorporated a system that enables the user to chat with each other over a Bluetooth connection. But this method might also face constraint as the range of the Bluetooth connection is restricted [12].

C. Android forensics analysis: Private chat on social messenger

This paper discusses regarding the relic data from secret, hidden and private chat. They also provide reports of formed

messages along with how it associates to one another. From the inquiry outcomes of Android forensics and interpretation, an inquisitor or investigator will be able to understand, reproduce, and confer the chronology of the information which has been generated by the user [13].

D. Extracting Intrauterine Device Usage from Clinical Texts using Natural Language Processing

This paper discusses the devices are much useful regarding preventing pregnancy which was not intended to happen. The Clinical texts help the patient to overcome this risk factor. But after a series of research, the results showed that these data are structured and cannot be efficiently used for processing and analyzing the risk factor but would be used in the traditional method of maintaining health record. To overcome this problem, a clinical factor extraction tool used called EasyCIE which was used to identify the patients who are contraception for counseling [14].

E. Dialogue-Oriented Interface for Linguistic Human-Computer Interaction: A Chat-based Application

This paper discusses a model allowing human-computer interaction by utilizing natural language utterance. Development of interactive agents like chatbots, in the recommended environment, the human-machine discussion is dedicated to a question or answer monotonic approach proposed at reducing semantic uncertainty within information. The prevalence of the chat as asynchronous transmission device permits design be fit for an infinite diversity of forms [15].

F. Lexical and Discourse Analysis of Online Chat Dialog

This paper discusses one of the final goals of natural language processing (NLP) methods is getting the definition of what is transpiring sent, irrespective of the means (e.g., printed versus lectured) or the pattern (e.g., static reports versus productive arguments). Although significant work has transpired performed in conventional language domains such as speech and static written text, understanding has yet been done in the latest information areas approved by the Internet, e.g., online chat and immediate messaging. That is in part due to the point that there are no explained chat corpora accessible to the broader research area. This dissection aims to develop a chat corpus, marked with lexical (part-of-speech token labels), syntactic, and speech data. Such a corpus can be practiced to improve more complicated, analytical-based NLP forms that do tasks such as artist profiling, item description, and social network reports [16].

The related work states that there are no methods or provisions for identification of lewd or vulgar context in the typed message of a user and stop the user from sending the message. The proposed methodology solves these issues by developing an NLP based android tool for identification of lewd or vulgar context in the message and prevents the user from sending those messages with a warning notification.

III. PROPOSED METHOD

Failure to identify the inappropriate context in the text message is the main reason for various problems. The proposed system solves these issues by developing an NLP based android tool which identifies and warns the user if the sentiment of the message contains lewd or vulgar context. The proposed system illustrated in Fig.1 consists of three phases.

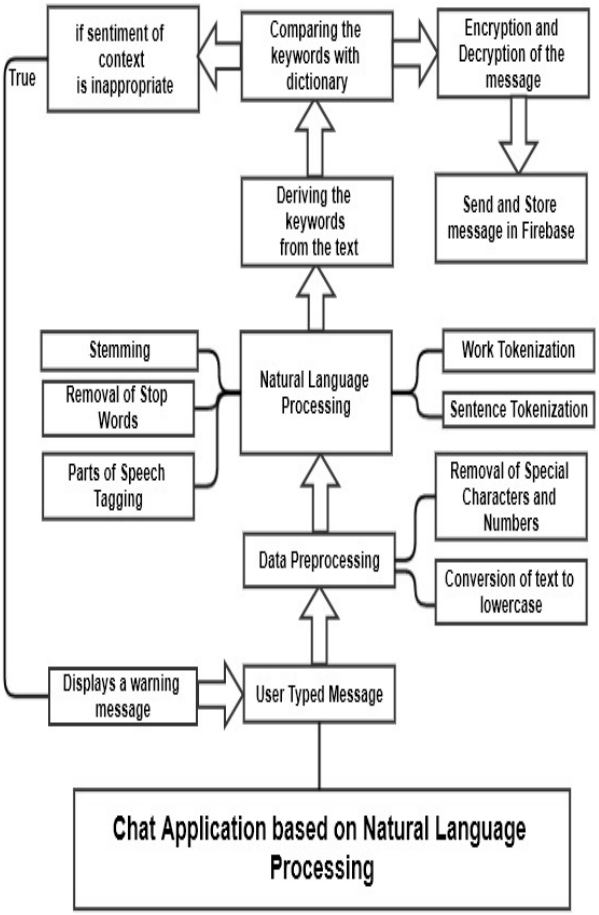

Fig. 1. Block Diagram of the Proposed System

The first phase deals with data pre-processing. The result of the first step given as an input to the second phase. The second phase deals with implementation of NLP concepts like the removal of stop words, stemming, entity recognition, tokenization and parts of speech tagging which derive keywords from the user typed the message. These keywords compared with user dictionary to identify irrelevant terms. The third phase deals with sending and receiving the messages using the internet and saving the messages in encrypted form in the real-time database "FireBase".

IV. IMPLEMENTATION

The NLP based tool implemented using Android, Java, and Firebase. The tool accepts the user typed message as an input. Data pre-processing applied to the user typed message.

Fig. 2. Splash Screen of Professional Chat Application

Data pre-processing is a procedure of data mining, executed on real-time data, which is vociferous, inadequate or erratic. The subsequent data scrubbing measures are employed:

(i) Eliminating all the distinctive letters from the line.

(ii) Trimming undesired tabs, spaces, newlines and additional nonprintable letters in the line.

(iii) Transforming the entire text to lower case.

All these steps performed to make computation easy.

Natural Language processing concepts applied to the pre-processed data. Natural Language Processing is a method for machines to evaluate, determine, and obtain the semantics of human language wisely. Java Programming language is employed to accomplish several tasks of Natural Language Processing on cleaned data. Word Tokenization, Sentence Tokenization, Removal of Stop words, Stemming, Entity Recognition, Parts of Speech Tagging, etc. are practiced to reconstruct the data to a pattern suitable for interpretation. Word Tokenization is the method of changing the text into tokens and saving it in the list. Sentence tokenization is the method of transforming the text into different phrases. Stop words deemed inapplicable or pointless because they have limited importance in capturing the semantics of the text and additionally stop words increase the searching time which leads to wastage of many computational resources. Stop words are omitted to preserve both time complexity and space complexity. Stemming is a crude method that cuts off the ends or beginning of words. The primary aspiration of stemming is to change a derivative word into its standard form and keeps the root word. Parts of Speech Tagging examines the text and assigns parts of speech to each token as a verb, adjective, noun, etc., Entity Identification helps to classify the named entities like persons, organizations, etc., from the text. After implementing NLP techniques on the pre-processed cleaned data, definitive keywords derived. These keywords compared with the keywords available in the user dictionary. The vocabulary contains all the keywords. From the keywords procured from the text, the sentiment of the context determined. If the meaning of the message is inappropriate, then user not allowed to send the message.

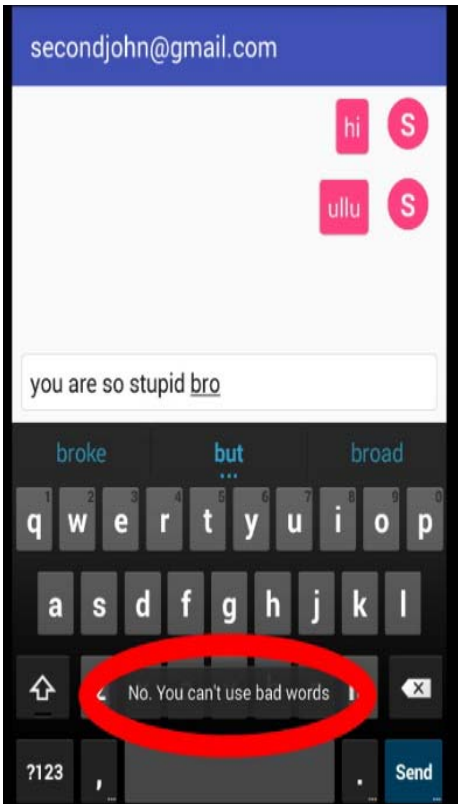

Fig. 3. Displaying warning message if the user tries to send inappropriate message

The information stored in a secured real-time database Firebase [17] [18]. The data stored in a JSON file which is easy to process. The MD5 algorithm used to convert the user typed message into fixed size hash of 32 digits' hexadecimal value [19].

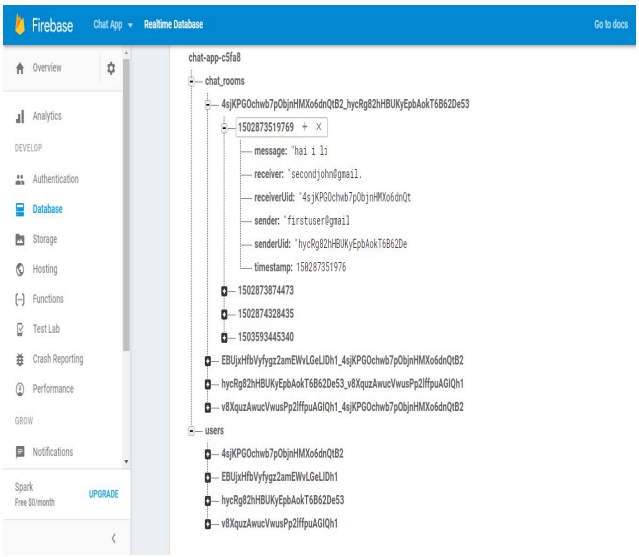

Fig. 4. Encrypted data stored in Firebase Server

V. CONCLUSION

It is the need of the hour to create a professional application that will not let the user send inappropriate messages. The proposed application developed in one of the most widely used operating systems in the world that is an Android OS and cryptographic techniques incorporated.

ACKNOWLEDGMENTS

This research is conducted at **Christ University,** Bengaluru, Karnataka, India. We would like to thank **Dr. Surendra Umesh Kulkarni,** Dean of Sciences for his support for the research work. We would also like to thank **Prof. Joy Paulose,** HOD, Department of Computer Science for his consistent guidance, support and encouragement throughout the research work. We would also like to thank **Dr. Rohini V,** Coordinator of MCA for her support for the research work.

REFERENCES

[1] Most popular mobile messaging apps worldwide as of January 2017, b. (2017). Most popular messaging apps 2017 | Statista. [online] Statista. Available at: https://www.statista.com/statistics/258749/most-popular-global-mobile-messenger-apps/.

[2] Internetdemocracy.in. (2017). Section 66A: Do not send offensive messages – The Internet Democracy Project. [online] Available at: https://internetdemocracy.in/laws/the-information-technology-amendment-act-2008/section-66a/.

[3] Avasarala, S. (2015). AN ANALYSIS OF VALIDITY OF SECTION 66A OF IT ACT, 2000 IN SHREYA SINGHAL V. UNION OF INDIA. [online] http://ijlljs.in. Available at: http://ijlljs.in/wp-content/uploads/2015/06/CASEANALYSISFORIJLLJS.pdf.

[4] Citizen Matters, Bengaluru. (2014). How not to get arrested under 66A for your online chatter |. [online] Available at: http://bengaluru.citizenmatters.in/how-not-to-get-arrested-under-66a-for-your-online-chattering-6589.

[5] Mhaiske, S. (2017). Abusive Chat. [online] Cybercrimcomplaints. Available at: https://www.cybercrimecomplaints.com/category/complaint-types/abusive-chat.

[6] The Hindu Business Line. (2015). Man arrested for sending lewd SMS to lady prof denied bail. [online] Available at: http://www.thehindubusinessline.com/news/national/man-arrested-for-sending-lewd-sms-to-lady-prof-denied-bail/article4512487.ece.

[7] The Times of India. (2014). Sending vulgar messages, photographs amount to outraging modesty: HC - Times of India. [online] Available at: https://timesofindia.indiatimes.com/india/Sending-vulgar-messages-photographs-amount-to-outraging-modesty-HC/articleshow/28286434.cms.

[8] DNA. (2014). Engineer arrested for sending obscene picture to woman | Latest News & Updates at Daily News & Analysis. [online] Available at: http://www.dnaindia.com/mumbai/report-engineer-arrested-for-sending-obscene-picture-to-woman-1983464.

[9] S. Karthick and S. Binu, 'Android Security Issues and Solutions', International Conference on Innovative Mechanisms for Industry Applications (ICIMIA-IEEE), pp. 686-689, 2017.

[10] S. Karthick and S. Binu, 'Static Analysis Tool for Identification of Permission Misuse by Android Applications', International Journal of Applied Engineering Research, 12(24), pp.15169-178, 2017

[11] Er. Kavindra Singh, Enhanced Education Chat Application Based on Interested Keyword with Username and Password Authentication Security. (2017). International Journal of Advanced Research in Computer Science and Software Engineering, 6(6).

[12] Nikita Mahajan, Design of Chatting Application Based on Android Bluetooth. (2017). International Journal of Computer Science and Mobile Computing, 3(3).

[13] Satrya, G., Daely, P. and Shin, S. (2016). Android forensics analysis: Private chat on social messenger. 2016 Eighth International Conference on Ubiquitous and Future Networks (ICUFN).

[14] Shi, J., Mowery, D., Zhang, M., Sanders, J., Chapman, W. and Gawron, L. (2017). Extracting Intrauterine Device Usage from Clinical Texts Using Natural Language Processing. 2017 IEEE International Conference on Healthcare Informatics (ICHI).

[15] Di Lecce, V., Calabrese, M., Soldo, D. and Quarto, A. (2010). Dialogue-oriented interface for linguistic human-computer interaction: A chat-based application. 2010 IEEE International Conference on Virtual Environments, Human-Computer Interfaces and Measurement Systems.

[16] Forsythand, E. and Martell, C. (2007). Lexical and Discourse Analysis of Online Chat Dialog. International Conference on Semantic Computing (ICSC 2007).

[17] Firebase. (2017). User Based Security | Firebase Realtime Database | Firebase. [online] Available at: https://firebase.google.com/docs/database/security/user-security.

[18] Firebase. (2017). Learn to Secure Files | Firebase. [online] Available at: https://firebase.google.com/docs/storage/security/secure-files.

[19] SearchSecurity. (2017). What is MD5? - Definition from WhatIs.com. [online] Available at: http://searchsecurity.techtarget.com/definition/MD5.