IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 15, May 2023

A Machine Learning Approach for Social Media Content Filtering

Gayatri Jawharkar, Prof. Rahul M. Raut, Sarvesh Sonawane, Shubham Kandekar, Parag Thorat
Department of Information Technology
Sandip Institute of Technology and Research Centre, Nashik, India

Abstract: *Communication has become stronger due to exponential increase in the usage of social media in the last few years. People use them for communicating with friends, finding new friends, updating any important activities of their life, etc. Due to their growing popularity and deep reach, these mediums are infiltrated with huge Volume of spam messages. Spam message randomly sent to multiple addressees by all sorts of groups, but mostly lazy advertisers and criminals who wish to lead you to phishing sites. In this we are using various machine learning techniques for detecting spam in the short text messages and also Google vision API for detecting spam images.*

Keywords*:* Spam Filtration ,Google Vision API, OWASP, Naive bayes classifier, Dictionary Based Algorithm

I. INTRODUCTION

Spam may or may not be harmful to the intended person. It might range from just a funny text message to a deadly virus that may corrupt the entire machine or a code written to steal all the information on your machine. Initially, the spam started spreading with email, but with the increase in the use of the Internet and the advent of social media , they started to spread like an epidemic. Popular Social media apps :-WhatsApp, Facebook , Instagram , Twitter ,etc. These apps includes personal information and extra information about everyone .We should think about increasing cyber crimes through social media, spam messages i. e fraud messages. These apps are just for entertainment there are no extra features preventing cyber crimes, vulgarity, obscences word, etc.

Spam message randomly sent to multiple addressees by all sorts of groups, but mostly lazy advertisers and criminals who wish to lead you to phishing sites. It has a higher response rate as compared to email spam. Apart from emails, and SMS , social networking like Twitter , Facebook, instant messenger like WhatsApp etc. are also contributing to a major chunk of spam over the network. The Project App refers to irrelevant or unsolicited messages sent over the messengers for abusing or may harm someone's personal life. The spam may or may not be harmful to the intended people. Message Protection is a tedious task and in the absence of automatic measure for filtering of message, the task of spam filtering is taken up with the person at the receiving end. Communication has become stronger due to exponential increase in the usage of social media in the last few years. People use them for communicating with friends, finding new friends, updating any important activities of their life, etc.

Among different types of social A Machine Learning Approach For Social Media Content Filtering media, most important are social networking sites and mobile networks. Due to their growing popularity and deep reach, these mediums are infiltrated with huge Volume of spam messages

II. PROBLEM STATEMENT

To develop a chatting application which blocks spam messages , URL's and images. Spam comments refer to the unwanted comments with rude words, advertisement, political or religious views. Massive spam comments seriously decrease users' reading experience and hinder the healthy development of social media. Thus, it is essential to detect and Filter spam comments.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 15, May 2023

III. PROPOSED METHODOLOGY

Stop word Removal

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

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

Advanced Encryption Standard

- Step 1 : Derive the set of round keys from the cipher key.
- Step 2 : Initialize the state array with the block data (plaintext).
- Step 3 : Add the initial round key to the starting state array.
- Step 4 : Perform nine rounds of state manipulation.
- Step 5 : Perform the tenth and final round of state manipulation.
- 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. Theseare applied, along with other operations, on an array of data that holds exactly oneblock of data/ the data to be encrypted. This array we call the state array.

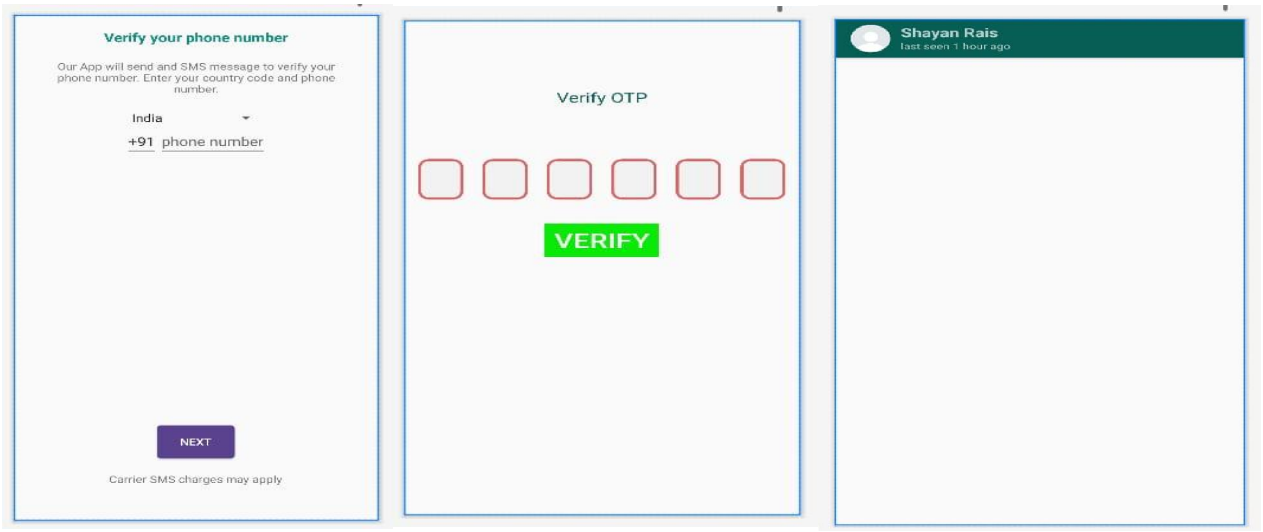
For Image :

- Input : Input Image
- Upload image on imagga API
- Pass Token
- Get Image BOW(Bag Of Word Lists)
- Check BOW present in Image word list
- If word present Content not uploaded
- Else Upload Image
- Output : Detect Image has Vulgar Content

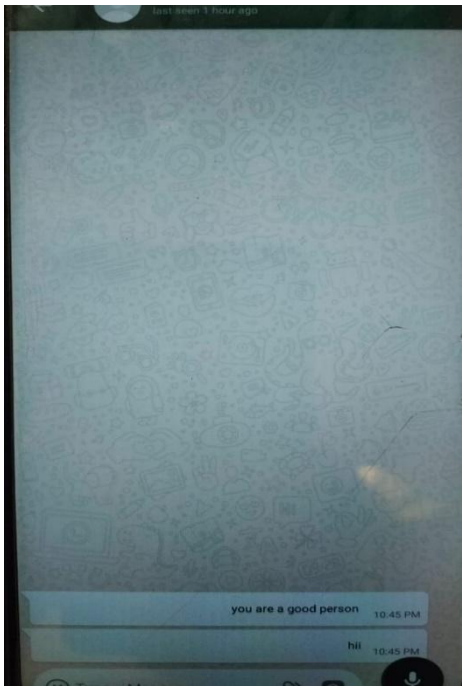
IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 15, May 2023

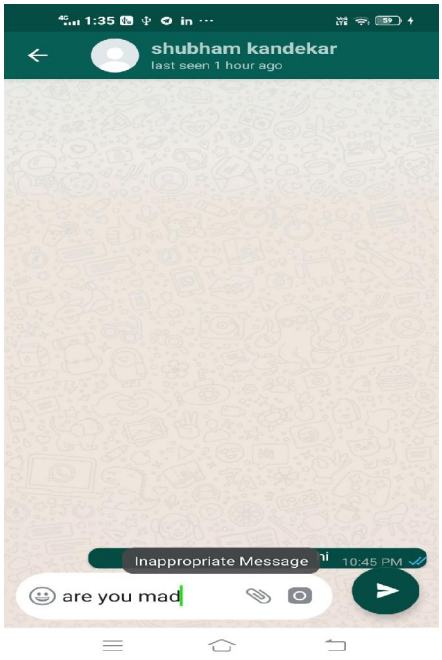
IV. RESULTS & SCREENSHOTS

Step 1 : Login / Registration Then Verify



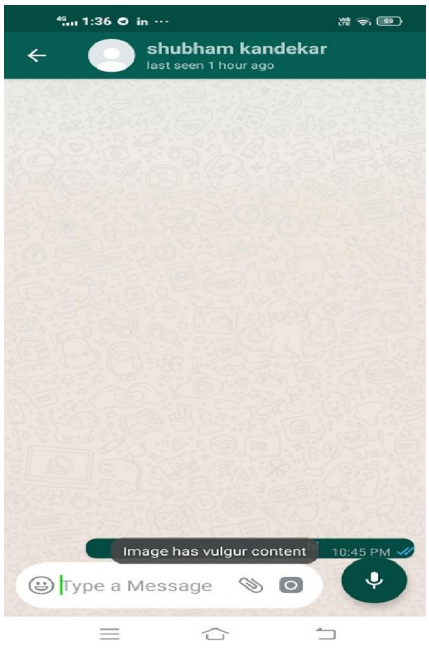
Step 2 : Try Sending normal Message and Then Any Vulgar word(ex- mad)





IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 15, May 2023

Similarly , for Images Classify them Keyword (BOW)



V. 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.

VI . FUTURE WORK

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 trolls.

REFERENCES

[1] Gianluca et al. Stringhini. Detecting spammers on social networks. In Proceedings of the 26th Annual Computer Security Applications Conference, pages 1–9. ACM, 2010.
[2] Chengfeng Lin et al. Analysis and identification of spamming behaviors in sina weibo microblog. In Proceedings of the 7th Workshop on Social Network Mining and Analysis. ACM, 2013.
[3] Jong Myoung Kim, Zae Myung Kim, and Kwangjo Kim. An approach to spam comment detection through domain-independent features. In Big Data and Smart Computing (BigComp), 2016 International Conference on, pages 273–276. IEEE, 2016.
[4] Chenwei Liu, Jiawei Wang, and Kai Lei. Detecting spam comments posted in micro-blogs using the self-extensible spam dictionary. In 2016 IEEE International Conference on Communications (ICC
[5] Archana Bhattarai, Vasile Rus, and Dipankar Dasgupta. Characterizing comment spam in the blogosphere through content analysis. In Computational Intelligence in Cyber Security, 2009., pages 37–44. IEEE, 2009.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 15, May 2023

[6] Michael Crawford, Taghi M. Khoshgoftaar, Joseph D. Prusa, Aaron N. Richter, and Hamzah Al Najada. Survey of review spam detection using machine learning techniques. In Journal of Big Data, volume 23, 2015.

[7] Fangzhao Wu, Jinyun Shu, Yongfeng Huang, and Zhigang Yuan. Co-detecting social spammers and spam messages in microblogging via exploiting social contexts. In Neurocomputing, volume 201, pages 51–65, 2016.

[8] T. M. Mahmoud and A. M. Mahfouz, "Sms spam filtering technique based on artificial immune system," IJCSI International Journal of Computer Science Issues, vol. 9, no. 1, pp. 589–597, 2012.

[9] X. Huang and M. Xu, "An Inter and Intra Transformer for Hate Speech Detection," 2021 3rd International Academic Exchange Conference on Science and Technology Innovation (IAECST), 2021, pp. 346-349, doi: 10.1109/IAECST54258.2021.9695652

[10] D. Sahnan, S. Dahiya, V. Goel, A. Bandhakavi and T. Chakraborty, "Better Prevent than React: Deep Stratified Learning to Predict Hate Intensity of Twitter Reply Chains," 2021 IEEE International Conference on Data Mining (ICDM), 2021, pp. 549-558, doi: 10.1109/ICDM51629.2021.00066.

[11] H. Watanabe, M. Bouazizi and T. Ohtsuki, "Hate Speech on Twitter: A Pragmatic Approach to Collect Hateful and Offensive Expressions and Perform Hate Speech Detection," in IEEE Access, vol. 6, pp. 13825-13835, 2018, doi: 10.1109/ACCESS.2018.2806394.

[12] S. Alsafari and S. Sadaoui, "Semi-Supervised Self-Learning for Arabic Hate Speech Detection," 2021 IEEE International Conference on Systems, Man, and Cybernetics (SMC), 2021, pp. 863-868, doi: 10.1109/SMC52423.2021.9659134.

[13] K. -Y. Lin, R. K. -W. Lee, W. Gao and W. -C. Peng, "Early Prediction of Hate Speech Propagation," 2021 International Conference on Data Mining Workshops (ICDMW), 2021, pp. 967-974, doi: 10.1109/ICDMW53433.2021.00126.

[14] R. A. Ilma, S. Hadi and A. Helen, "Twitter's Hate Speech Multi-label Classification Using Bidirectional Long Short-term Memory (BiLSTM) Method," 2021 International Conference on Artificial Intelligence and Big Data Analytics, 2021, pp. 93-99, doi: 10.1109/ICAIBDA53487.2021.