Survey Paper on Content Filtering

Suwaiba Sayyed[1], Mohammadi Shaikh[2], Gauri Kolhe[3], Sumit Mandage[4] Student of

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

1. 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.

2. 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, and 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 hate messages which contain offensive and vulgar words. Offensive and vulgar words could be bad words and hate speeches. 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 words is present in that message then our algorithms will detect them and will control by sending them back to the sender with an alert box" This message contains inappropriate words/ Cannot be forwarded further". As our application is all about automatic actions hence it provide another favorable aspect that is, If any message sent by sender for maximum three times than sender will get blocked for 24 hours.

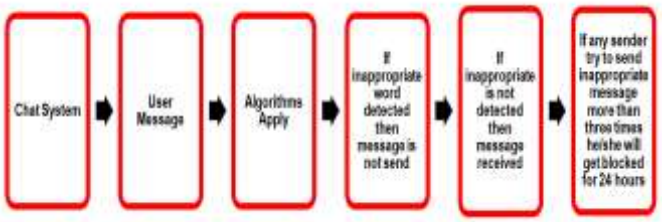

Fig 1. Introduction

3. 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.

4. OBJECTIVE

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

5. 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.

5. 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.

6. 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].

7. 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. 11

8. 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

9. 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]

10. They compared Support Vector Machine (SVM) and Decision Tree for email filtering. [A.S Yuksel] [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.

6. PROBLEM STATEMENT

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, and 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 a messages which contain offensive and vulgar words.

Now, these problems have been considered to design or provide 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 words are present in that message then our algorithms will detect such spam will control them by sending them back to the sender with an alert box 'This message cannot be forwarded further'.
As the application is all about automatic actions hence it provides another favorable aspect that is, if any message sent by sender for maximum three times than sender will get blocked for 24 hours.

7. MOTIVATION

The reason to do this is simple: by detecting unsolicited and unwanted content, we can prevent spam messages from creeping into the user's inbox, thereby improving user experience.
Purpose behind making this project is that in our application user can easily and freely share their information and data which will be protected from all kind of offensive and vulgar words.

8. 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 if user try to send the message to the other user and that message will contain any offensive and vulgar word then that message is not send. We can store user data and our dataset on the cloud while sending the message our algorithm tries to analyses that message.

If the message does not contain any offensive and vulgar word, then message is successfully send to the other user.
The proposed system is all about automatic actions hence it provides another favorable aspect that is, if any message sent by sender for maximum three times than sender will get blocked for 24 hours and much more.

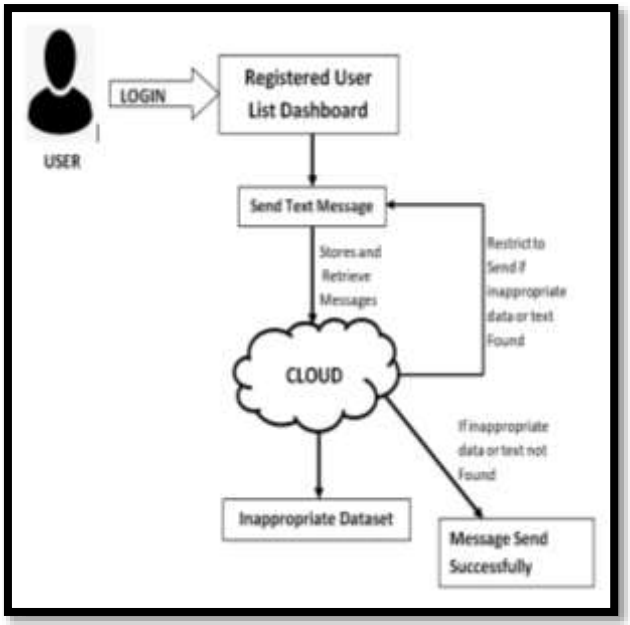

Fig 2. System Architecture

9. SYSTEM IMPLEMENTATION

9.1 Overview of Project Module

The aim of the project is to detect offensive and vulgar words from messages which create a negative impact on society. The motive is often to stop access to content that the user might determine is objectionable. The purpose of our project is to build a healthy environment on social media, to reduce inappropriate data, it can often be a strong step towards providing security to all users, by detecting unwanted content we can prevent messages from creeping into the user's inbox. Our goal is that users can not able to send inappropriate data or messages. If any sender will try to send an inappropriate message more than one time, then the sender will get blocked for 24 hours. In this way, we are trying to improve the security of the chat system.

9.2 The Module of our system:

- User Login
- Dataset Creation
- Detect offensive and vulgar words
- Trained Dataset or Keywords

• Display the Result

10. ALGORITHMS

Natural Language Processing

Natural language processing is one of the fields in programming where the natural language is processed by the software. This has many applications like sentiment analysis, language translation, fake news detection, grammatical error detection etc. The input in natural language processing is text. The data collection for this text happens from a lot of sources. This requires a lot of cleaning and processing before the data can be used for analysis.

10.1 Tokenization

Tokenization is breaking the raw text into small chunks. Tokenization breaks the raw text into words, sentences called tokens. These tokens help in understanding the context or developing the model for the NLP. Tokenization can be done to either separate words or sentences. If the text is split into words using some separation technique it is called word tokenization and same separation done for sentences is called sentence tokenization.

ALGORITHM STEPS:
Input: - Message or Text
Process: -

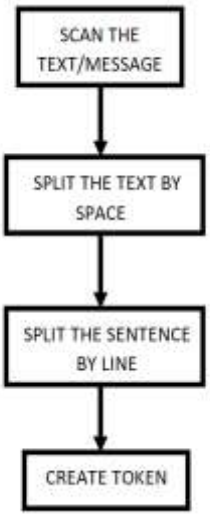

Fig 2.Tokenization Step

Output: - Token List

10.2 Stop Words

A stop word is a commonly used word (such as "the", "a", "an", "in") that a search engine has been programmed to ignore, both when indexing entries for searching and when retrieving them as the result of a search query. We would not want these words to take up space in our database, or taking up valuable processing time. For this, we can remove them easily, by storing a list of words that you consider to stop words. NLTK (Natural Language Toolkit) in python has a list of stop words stored in 16 different languages. Stop words are the words in any language which does not add much meaning to a sentence. They can safely be ignored without sacrificing the meaning of the sentence. For some search engines, these are some of the most common, short function words, such as the, is, at, which, and on. In this case, stop words can cause problems when searching for phrases that include them, particularly in names such as "The Who" or "Take That".

ALGORITHM STEPS:
Input: - Message or Text
Process: -

Fig 3.Stop words Steps

Output: - Clean the Data

11. ADVANTAGES

1) User will be unable to send inappropriate data.
2) User can connect anyone at any time.
3) Chat system Application is Easy to use.
4) Cost effective.

12. APPLICATIONS

- Google Gmail
- Google drive
- Microsoft outlook
- Facebook
- Twitter

13. CONCLUSION

Hence, due to the increase in the usage of social media networks crime has increased. Such as abusive words, and offensive words been increased. Due to this unhealthy weather is been created on social media and one can easily get depressed. So the purposed system will detect offensive and vulgar words from the messages and if any inappropriate words detect then the message cannot be forwarded further. And if the message does not contain any inappropriate words then the message can send successfully. Because of this system can only be forwarded or send valuable and knowledgeable content to the users. Also, it provides high security because if any user can try to send an inappropriate message more than three times then the user will get blocked for 24 hours. Because of this security will increase.

14. FUTURE SCOPE

1) In the future scope feature of image filtering and video filtering can be added.
2) Groups can be created.
3) Groups will improve engagement among users.
4) Video call feature can be added.
5) Language translation feature can be implemented on the application.

15. REFERENCES

[1] S. Srinivasan, V. Ravi, M. Alazab, S. Ketha, A.-Z. Ala'M, and S. K. Padannayil, "Spam emails detection based on distributed word embedding with deep learning," in Machine Intelligence and Big Data Analytics for Cybersecurity Applications, Springer, 2021, pp. 161– 189.

[2] G. Egozi and R. Verma, "Phishing email detection using robust nlp techniques," in 2018 IEEE International Conference on Data Mining Workshops (ICDMW), IEEE, 2018, pp. 7–12.

[3] S. Seth and S. Biswas, "Multimodal spam classification using deep learning techniques," in 2017 13th International Conference on SignalImage Technology & Internet-Based Systems (SITIS), IEEE, 2017, pp. 346–349.

[4] A. Bibi, R. Latif, S. Khalid, W. Ahmed, R. A. Shabir, and T. Shahryar, "Spam mail scanning using machine learning algorithm.," JCP, vol. 15, no. 2, pp. 73–84, 2020. [17] W. Awad and S. ELseuofi, "Machine learning methods for spam e-mail classification," International Journal of Computer Science & Information Technology (IJCSIT), vol. 3, no. 1, pp. 173–184, 2011.

[5] S. K. Tuteja, "Classification Algorithms for Email Spam Filtering", 2016.

[6] Alejandro Metke Jimenez, Kerry Raymond, and Ian MacColl. 2011. Information Extraction from Web Services: A Comparison of Tokenisation Algorithms. In Proceedings of the 2nd International Workshop on Software Knowledge 2011, in conjunction with 3rd International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management. Scitepress, 12–23.

[7] Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural Machine Translation of Rare Words with Subword Units. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, Berlin, Germany, 1715–1725

[8] Ivan Provilkov, Dmitrii Emelianenko, and Elena Voita. 2020. BPE-Dropout: Simple and Effective Subword Regularization. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, Online, 1882–1892.

[9] Cagri Toraman, Furkan Ş¸ahinuç¸, and Eyup Halit Yilmaz. 2022. Large-Scale Hate Speech Detection with Cross-Domain Transfer. arXiv preprint arXiv:2203.01111 (2022).

[10] Yüksel, A. S., Cankaya, S. F., & Üncü, İ. S. "Design of a Machine Learning Based Predictive Analytics System for Spam Problem", 2017.