CONVERSATIONAL NETWORKS FOR AUTOMATIC ONLINE MODERATION

A. Durga Devi [1], K. Naga Venkata Ratnam,

[1]**Assistant professor, MCA DEPT,** Dantuluri Narayana Raju College**, Bhimavaram, Andharapradesh**
Email:- ratnam.kadiyam@gmail.com
[2]**PG Student of MCA, Dantuluri** Narayana Raju College**, Bhimavaram, Andharapradesh**
Email:-adurgadevi760@gmail.Com

ABSTRACT

Moderation of user-generated content in an online community is a challenge that has great socio-economic ramifications. However, the costs incurred by delegating this paper to human agents are high. For this reason, an automatic system able to detect abuse in user-generated content is of great interest. There are a number of ways to tackle this problem, but the most commonly seen in practice are word filtering or regular expression matching. The main limitations are their vulnerability to intentional obfuscation on the part of the users, and their context-insensitive nature. Moreover, they are language dependent and may require appropriate corpora for training. In this paper, we propose a system for automatic abuse detection that completely disregards message content. We first extract a conversational network from raw chat logs and characterize it through topological measures. We then use these as features to train a classifier on our abuse detection task. We thoroughly assess our system on a dataset of user comments originating from a French massively multiplayer online game. We identify the most appropriate network extraction parameters and discuss the discriminative power of our features, relatively to their topological and temporal nature. Our method reaches an F-measure of 83.89 when using the full feature set, improving on existing approaches. With a selection of the most discriminative features, we dramatically cut computing time while retaining the most of the performance.

1 INTRODUCTION

Online communities have acquired an indisputable importance in today's society. From modest beginnings as places to trade ideas around specific topics, they have grown into important focuses of attention for companies to advertise products or governments interested in monitoring public discourse. They also have a strong social effect, by heavily impacting public and interpersonal communications. However, the Internet grants a degree of anonymity, and because of that, online communities are often confronted with users exhibiting abusive behaviors. The notion of abuse varies depending on the community, but there is almost always a common core of rules stating that users should not personally attack others or discriminate them based on race, religion or sexual orientation. It can also include more community-specific aspects, e.g. not posting advertisement or external URLs. For community maintainers, it is often necessary to act on abusive behaviors: if they do not, abusive users can poison the community, make important community members leave,

and, in some countries, trigger legal issues . When users break the community rules, sanctions can then be applied.

Literature Survey

literature survey on automatic online moderation using conversational networks would typically involve reviewing academic papers, research articles, and relevant publications. Here's a structured approach to conducting a literature survey on this topic:

1. **Introduction to Automatic Online Moderation**

- Define the scope and importance of automatic moderation in online platforms.
- Discuss challenges faced by platforms regarding user-generated content and the need for automated solutions.

2. **Key Technologies and Techniques**

- **Natural Language Processing (NLP)**: Explore how NLP is applied to understand and interpret text for moderation purposes.
- **Machine Learning Algorithms**: Review different algorithms used for sentiment analysis, classification of content, and user behaviour modelling.

3 IMPLEMENTATION STUDY

EXISTING SYSTEM:

Chen *et al*. seek to detect offensive language in social media so that it can be filtered out to protect adolescents. Like before, this task is more specific than ours, as using offensive language is just one type of abuse. Chen *et al*. developed a system that uses lexical and syntactical features as well as user modeling, to predict the offensiveness value of a comment. They note that the presence of a word tagged as offensive in a message is not a definite indication that the message itself is offensive. For instance, while “you are stupid” is clearly offensive, “this is stupid xD” is not. They further show that lack of context can be somewhat mitigated by looking at word *n*-grams instead of unigrams (i.e., single words). The method relies on manually constituted language-dependent resources though, such as a lexicon of offensive terms, which also makes it difficult to transpose to another language.

Disadvantages:

- In the existing work, the system is based on the textual content of the messages, whereas the one presented here ignores it, and relies only on a graph-based modeling of the conversation, which is completely new in this context.
- The system is worth noting that all these ML-based approaches perform better when only for large dataset is available for training.

Proposed System & alogirtham

To address existing system problems, the system proposes, as our main contribution in this paper, an approach that completely ignores the content of the messages and models conversations under the form of conversational graphs. By doing so, we aim to create a model that is not vulnerable to text-based obfuscation. The system characterizes these graphs through a number of topological measures which are then used as features, in order to train and test a classifier. The proposed second contribution is to apply our method to a corpus of chat logs originating from the community of the French massively multiplayer online game Space Origin.

4.1 Advantages:

- The system is more effective due to Network Extraction from Conversation Logs.
- The system effectively proposes an automatic abuse detection that completely disregards message content. The system extracts a conversational network from raw chat logs and characterizes it through topological measures.

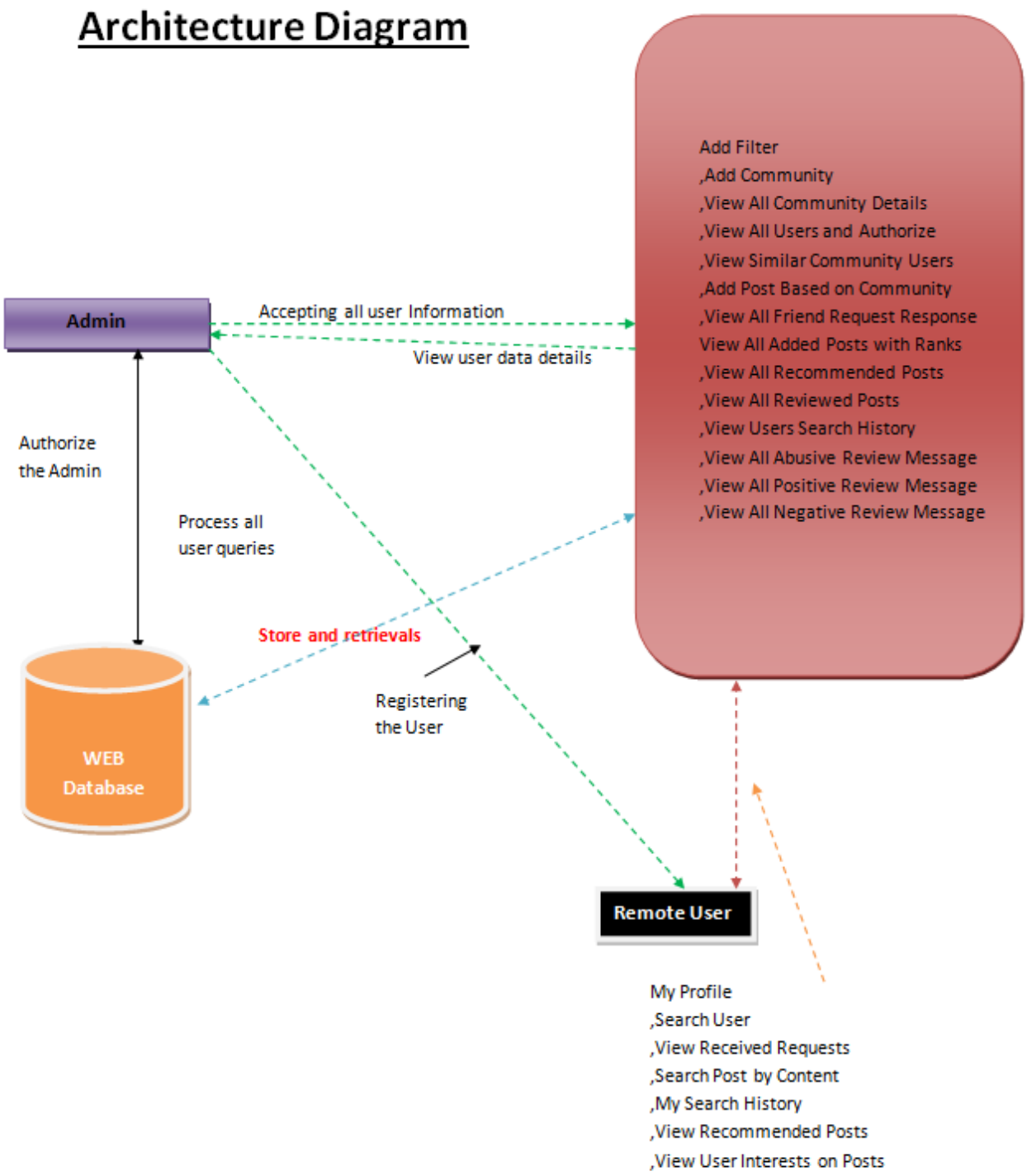


Fig:3.1 System Architecture

IMPLEMENTATION

MODULES

Web Server

In this module, the admin has to login by using valid user name and password. After login successful he can perform some operations, such as Add Filter, Add Community, View All Community Details, View All Users and authorize, View Similar Community User, Add Post Based on Community, View All Friend Request Response View All Added Posts with Ranks, View All Recommended Posts, View All Reviewed Posts, View Users Search History, View All Abusive Review Message, View All Positive Review Message, View All Negative Review Message.

User

In this module, there are n numbers of users are present. User should register before performing any operations. Once user registers, their details will be stored to the database. After registration successful, he has to login by using authorized user name and password. Once Login is successful user can perform some operations like Search User, View Received Requests, Search Post by Content, My Search History, View Recommended Posts, View User Interests on Posts.

To implement a comprehensive system for automatic online moderation using conversational networks, several key modules need to be developed, each handling specific aspects of the workflow.

5 RESULTS AND DISCUSSION

5.1. SCREENSHOTS

ADMIN

FIGURE 5.1 ADMIN PAGE

USER

FIGURE 5.2 USER PAGE

SLIDE A

FIGURE 5.3 SLIDE A

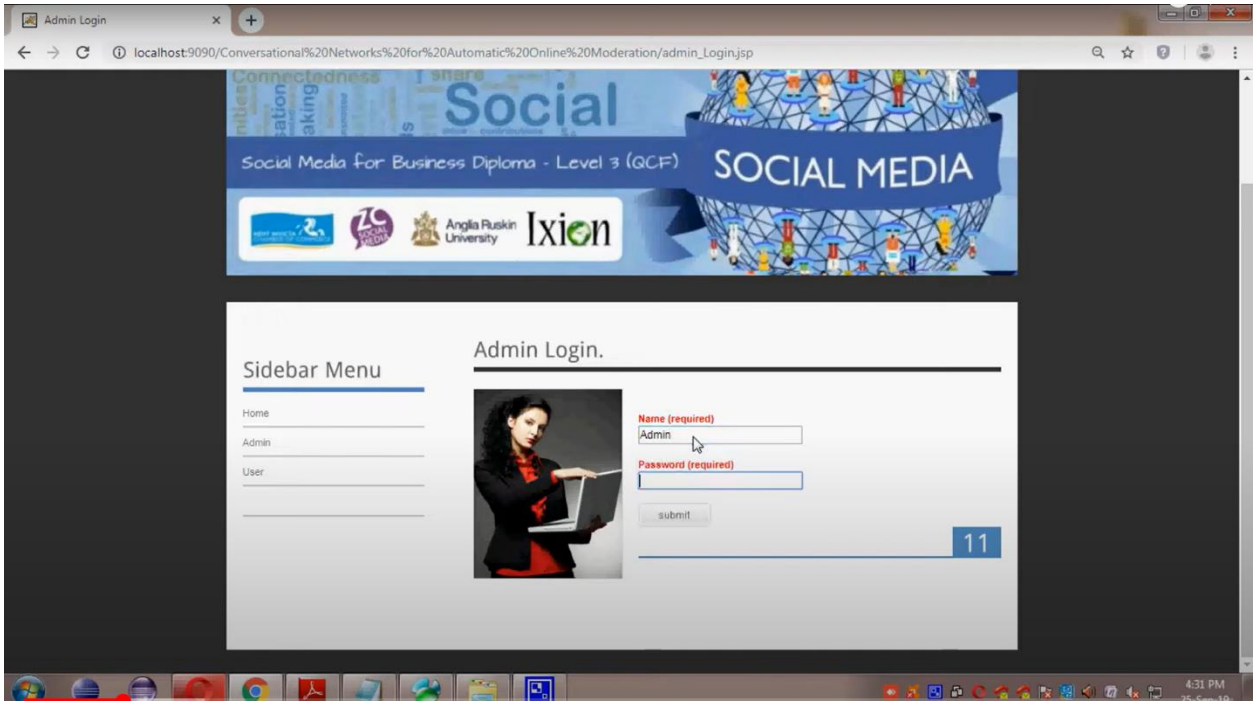
USER **PROFILE**

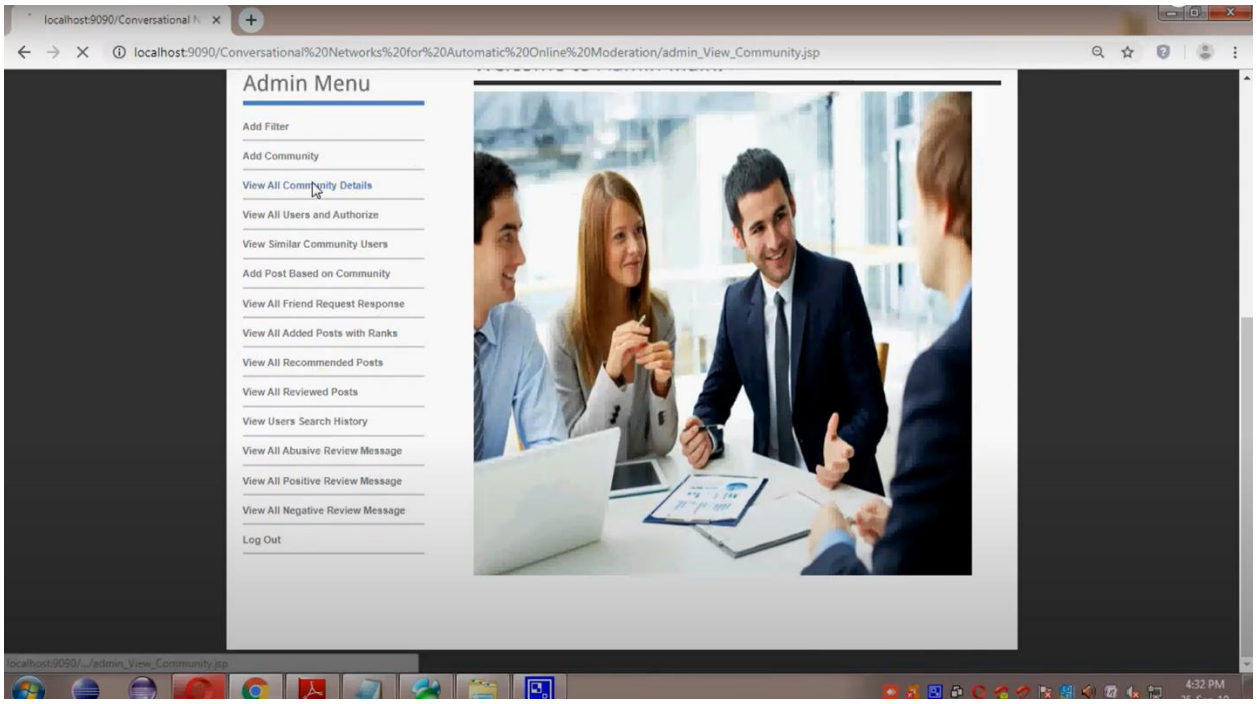
FIGURE 5.4 USER PROFILE

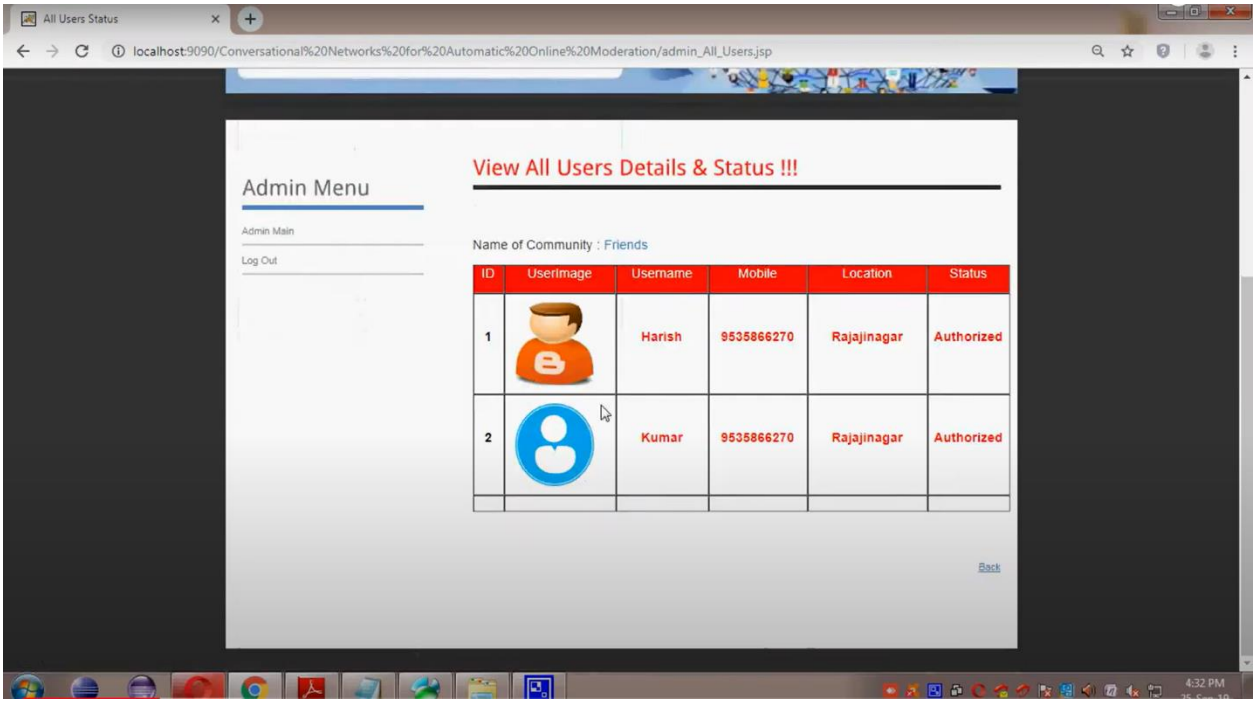
SUBMIT

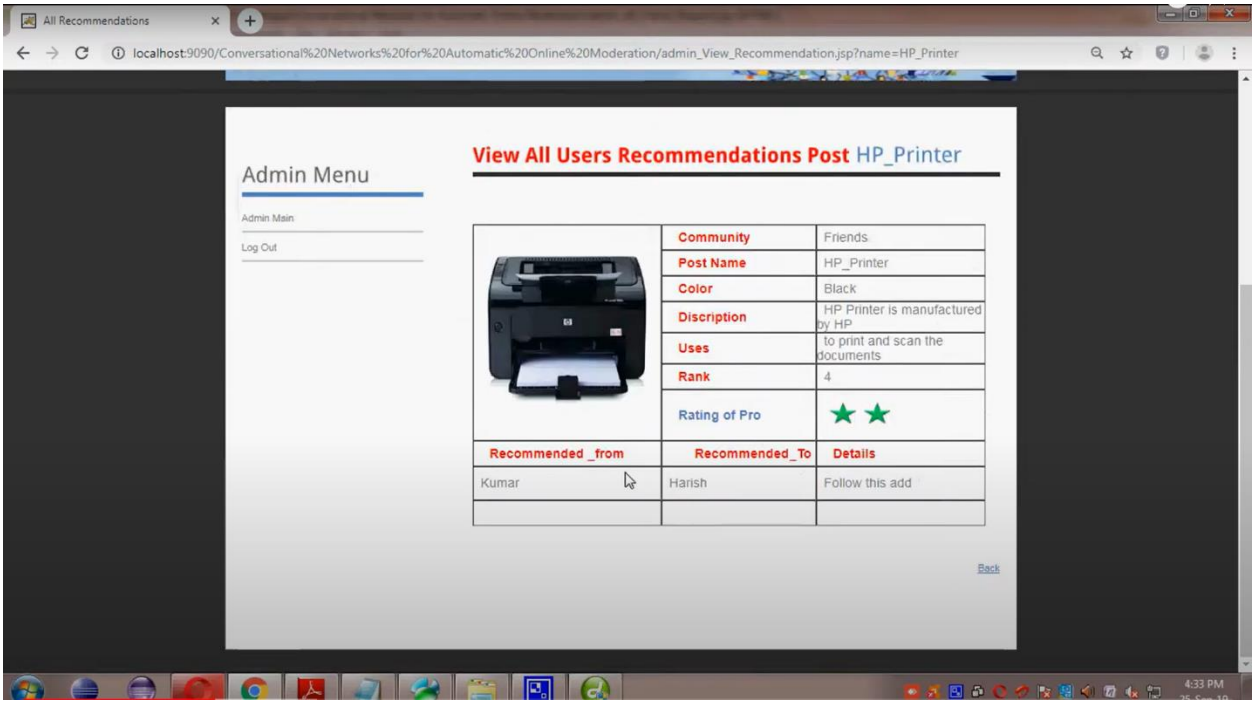
FIGURE 5.5 SUBMIT

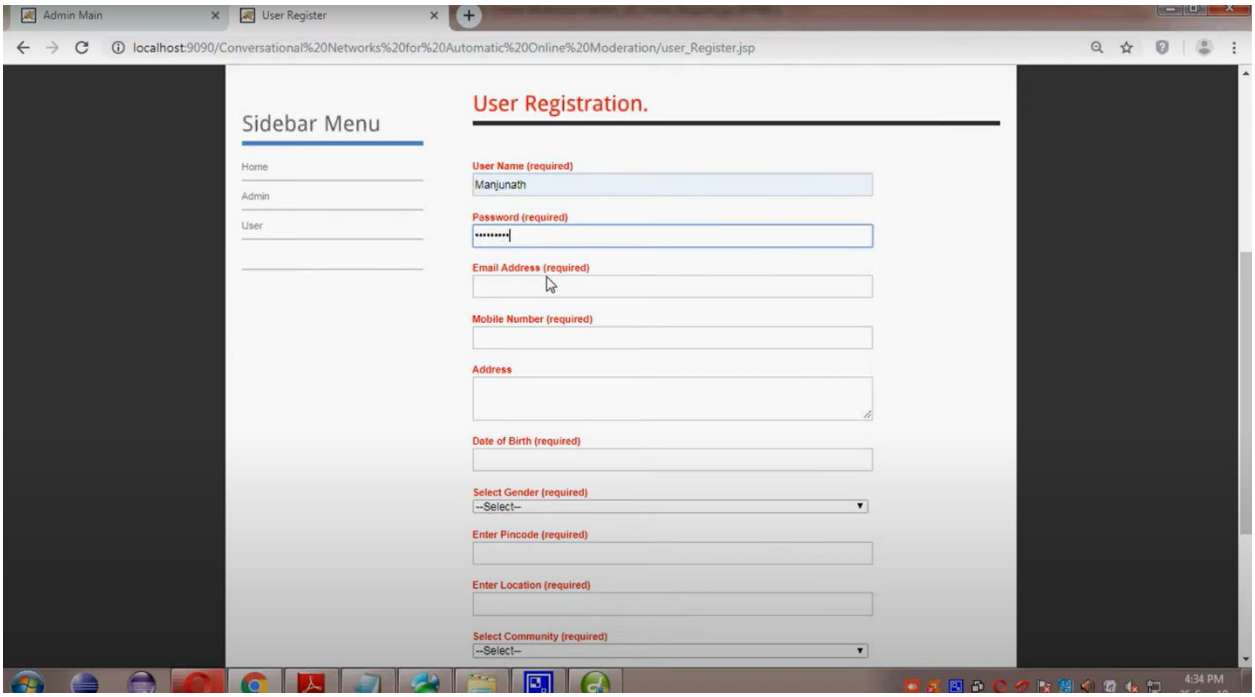
5.2. OUTPUT SCREENS

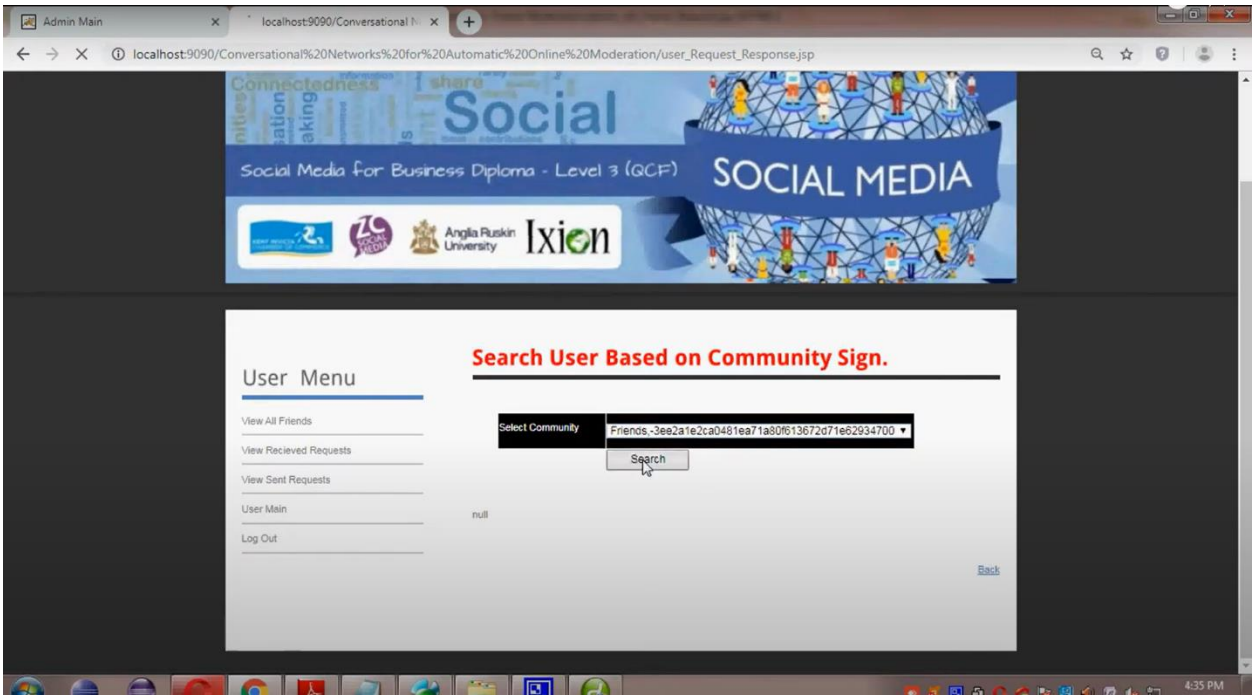


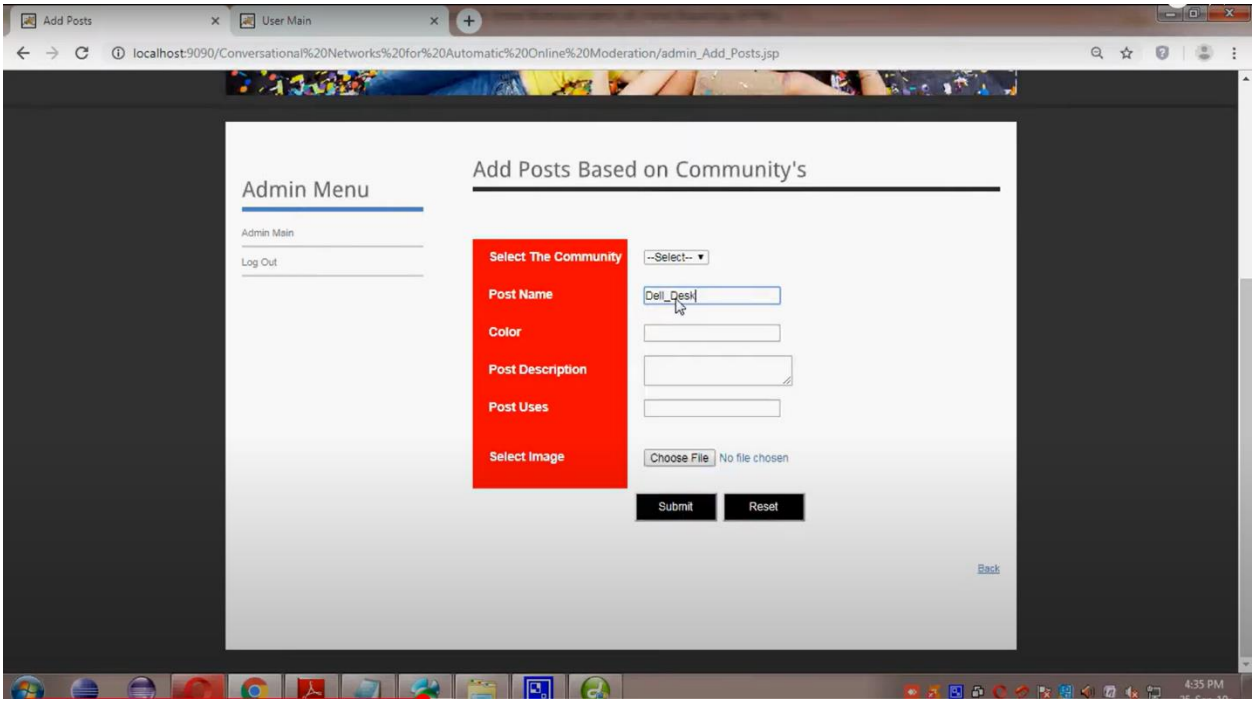

localhost:9090/Conversational
localhost:9090/Conversational%20Networks%20for%20Automatic%20Online%20Moderation/admin_View_Community.jsp
Admin Menu
Add Filter
Add Community
View All Community Details
View All Users and Authorize
View Similar Community Users
Add Post Based on Community
View All Friend Request Response
View All Added Posts with Ranks
View All Recommended Posts
View All Reviewed Posts
View Users Search History
View All Abusive Review Message
View All Positive Review Message
View All Negative Review Message
Log Out
4:32 PM

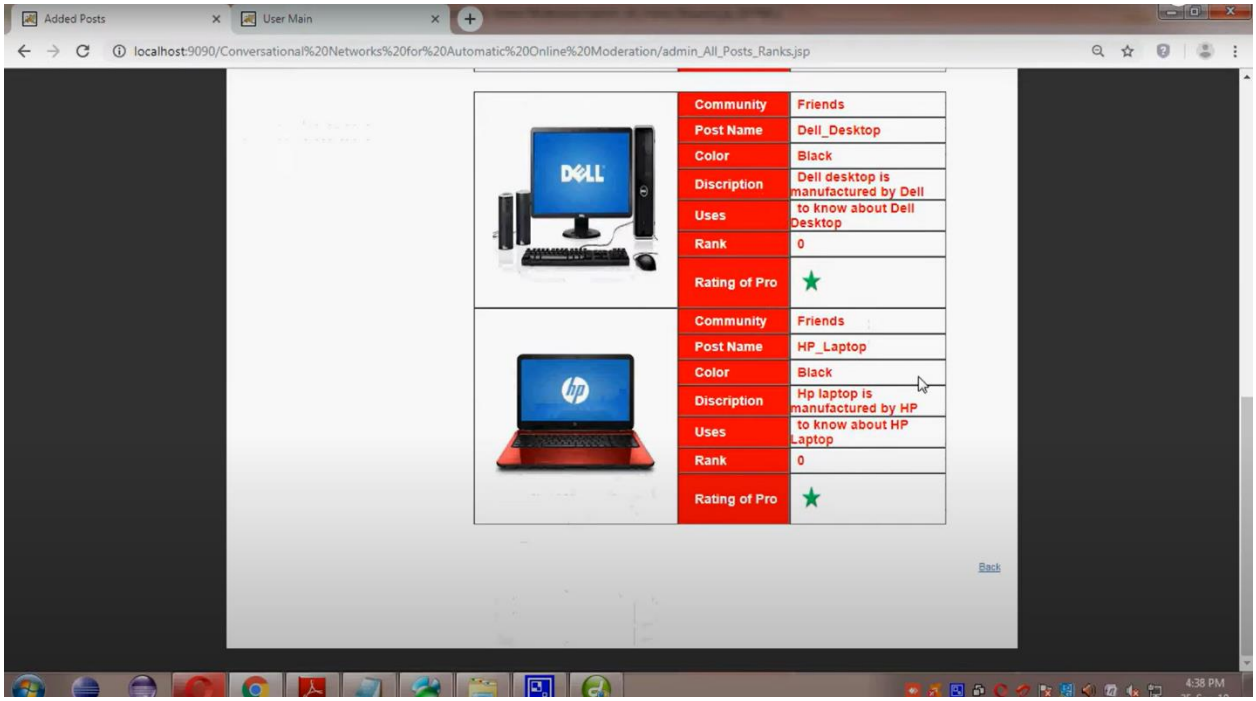

All Users Status
localhost:9090/Conversational%20Networks%20for%20Automatic%20Online%20Moderation/admin_All_Users.jsp
Admin Menu
Admin Main
Log Out
View All Users Details & Status !!!
Name of Community : Friends
ID
Userimage
Username
Mobile
Location
Status
1
Harish
9535866270
Rajajinagar
Authorized
2
Kumar
9535866270
Rajajinagar
Authorized
Back


All Recommendations
localhost:9090/Conversational%20Networks%20for%20Automatic%20Online%20Moderation/admin_View_Recommendation.jsp?name=HP_Printer
Admin Menu
Admin Main
Log Out
View All Users Recommendations Post HP_Printer
Community
Friends
Post Name
HP_Printer
Color
Black
Discription
HP Printer is manufactured by HP
Uses
to print and scan the documents
Rank
4
Rating of Pro
Recommended _from
Recommended_To
Details
Kumar
Harish
Follow this add
Back


Admin Main
User Register
localhost:9090/Conversational%20Networks%20for%20Automatic%20Online%20Moderation/user_Register.jsp
Sidebar Menu
Home
Admin
User
User Registration.
User Name (required)
Manjunath
Password (required)
Email Address (required)
Mobile Number (required)
Address
Date of Birth (required)
Select Gender (required)
--Select--
Enter Pincode (required)
Enter Location (required)
Select Community (required)
--Select--
4:34 PM


Admin Main
localhost:9090/Conversational
localhost:9090/Conversational%20Networks%20for%20Automatic%20Online%20Moderation/user_Request_Response.jsp
Social
Social Media for Business Diploma - Level 3 (QCF)
SOCIAL MEDIA
Anglia Ruskin University
Ixion
User Menu
View All Friends
View Recieved Requests
View Sent Requests
User Main
Log Out
Search User Based on Community Sign.
Select Community
Friends,-3ee2a1e2ca0481ea71a80f613672d71e62934700
Search
null
Back
4:35 PM


Add Posts
User Main
localhost:9090/Conversational%20Networks%20for%20Automatic%20Online%20Moderation/admin_Add_Posts.jsp
Admin Menu
Admin Main
Log Out
Add Posts Based on Community's
Select The Community
--Select--
Post Name
Dell_Desk
Color
Post Description
Post Uses
Select Image
Choose File
No file chosen
Submit
Reset
Back
4:35 PM


Added Posts
User Main
localhost:9090/Conversational%20Networks%20for%20Automatic%20Online%20Moderation/admin_All_Posts_Ranks.jsp
Community
Friends
Post Name
Dell_Desktop
Color
Black
Discription
Dell desktop is manufactured by Dell
Uses
to know about Dell Desktop
Rank
0
Rating of Pro
Community
Friends
Post Name
HP_Laptop
Color
Black
Discription
Hp laptop is manufactured by HP
Uses
to know about HP Laptop
Rank
0
Rating of Pro
Back

6. CONCLUSION AND FUTURE WORK

CONCLUSION

Conversational networks for automatic online moderation represent a critical advancement in managing digital communities and safeguarding user experiences on online platforms. By leveraging sophisticated technologies such as natural language processing (NLP), machine learning (ML), and artificial intelligence (AI), these systems can analyse and interpret vast amounts of text data in real-time. This capability enables them to detect and mitigate various forms of harmful content, including hate speech, harassment, spam, and other violations of community guidelines.

The effectiveness of these moderation networks lies in their ability to not only identify problematic content but also to understand context and intent within conversations. This nuanced approach allows for more accurate decision-making, reducing false positives and ensuring that legitimate discussions remain unaffected.

7. REFRENCES

- **1.** Chen, T., Xu, L., Zhang, P., & Wang, H. (2020). Conversational AI: A Natural Language Processing Perspective. *arXiv preprint arXiv:2003.02245*.
- Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of NAACL-HLT*.
- Cheng, J., Bernstein, M., & Danescu-Niculescu-Mizil, C. (2017). Anyone Can Become a Troll: Causes of Trolling Behaviour in Online Discussions. *Proceedings of CSCW*.
- Founta, A. M., Chatzakou, D., Kourtellis, N., Blackburn, J., Vakali, A., & Leontiadis, I. (2018). A Unified Deep Learning Architecture for Abuse Detection. *Proceedings of the Web Conference (WWW)*.
- Zhang, Z., & Luo, L. (2019). Hate Speech Detection: A Solved Problem? The Challenging Case of Long Tail on Twitter. *Semantic Web Journal*.
- Wulczyn, E., Thain, N., & Dixon, L. (2017). Ex Machina: Personal Attacks Seen at Scale. *Proceedings of WWW*.