ONLINE PUBLIC SHAMING ON TWITTER(X): DETECTION, ANALYSIS AND MITIGATION

B.S.Murthy [1], G. Venkata Lakshmi

[1]**Assistant professor(HOD) , MCA DEPT,** Dantuluri Narayana Raju College**, Bhimavaram, Andhrapradesh**
Email:- suryanarayanamurthy.b@gmail.com
[2]**PG Student of MCA, Dantuluri** Narayana Raju College**, Bhimavaram, Andhrapradesh**
Email:- kavya534201@gmail.com

ABSTRACT

Public shaming in online social networks and related online public forums like Twitter(x) has been increasing in recent years. These events are known to have devastating impact on the victim's social, political and financial life. Notwithstanding its known ill effects, little has been done in popular online social media to remedy this, often by the excuse of large volume and diversity of such comments and therefore unfeasible number of human moderators required to achieve the task. In this project, we automate the task of public shaming detection in Twitter(x) from the perspective of victims and explore primarily two aspects, namely, events and shamers. Shaming tweets are categorized into six types- abusive, comparison, passing judgment, religious/ethnic, sarcasm/joke and whataboutery and each tweet is classified into one of these types or as non-shaming. It is observed that out of all the participating users who post comments on a particular shaming event, the majority of them are likely to shame the victim. Interestingly, it is also the shamers whose follower counts increase faster than that of the non-shamers in Twitter(x). Finally, based on categorization and classification of shaming tweets, and web application called "Block Shame" has been designed and deployed for on-the-fly muting/blocking of shamers attacking a victim on Twitter(x).

1 INTRODUCTION

The internet plays a crucial role in various aspects of human life. The Internet is a collection of computers connected through telecommunication links such as phone lines, fiber optic lines, and wireless and satellite connections. It is a global computer network. The internet is used to obtain information stored on computers, which are known as hosts and servers. For communication purposes, they used a protocol called Internet protocol/transmission control protocol (IP-TCP). The government is not recognized as an owner of the Internet; many organizations, research agencies, and universities participate in managing the Internet. This has led to many convenient experiences in our lives regarding entertainment, education, banking, industry, online freelancing, social media, medicine, and many other fields in daily life. The internet provides many advantages in different fields of life. In the field of information search, the Internet has become a perfect opportunity to search for data for educational and research purposes. Email is a messaging source in fast way on the

Internet through which we can send files, videos, pictures, and any applications, or write a letter to another person around the world.

Literature Survey

Efforts to moderate user-generated content on the Internet started very early. Smokey is one of the earliest computational works in this direction which builds a decision tree classifier for insulting posts trained on labeled comments from two web forums. Although academic research in this area started that early, it used different nomenclatures including abusive, flame, personal attack, bullying, hate speech, etc., often grouping more than a single category under a single name [6]. Based on the content (and not the specific term used), we divide the related work into five categories: profanity, hate speech, cyberbullying, trolling, and personal attacks**.**

3 IMPLEMENTATION STUDY

EXISTING SYSTEM:

The existing systems for online public shaming detection, analysis, and mitigation on Twitter(x) primarily focus on identifying and categorizing shaming tweets, understanding the dynamics of shaming events and the behaviour of shamers, and developing tools to protect victims from such online harassment.

Disadvantages:

Most of the previous works mentioned above do not make a distinction between acceptability- ty and non-acceptability of a comment based on the presence or absence of a predefined victim.

Proposed System & alogirtham

In the proposed system, the system proposes a methodology for the detection and mitigation of the ill effects of online public shaming. We make three main contributions in this work-

(a) Categorization and automatic classification of shaming tweets

(b) Provide insights into shaming events and shamers

4.1 Advantages:

The System is very effective due to AUTOMATED CLASSIFICATION OF SHAMING TWEETS

IMPLEMENTATION

4.1 MODULES

Admin:

In this module, the Admin has to login by using a valid user name and password. After login successful he can do some operations such as view all user and their details and authorize them, Add and View All Filters, View All Created Tweets, View All Recommended Tweets, View All Shared Tweets, View All Transactions, View Tweets Using Tripartite Graph, View Positive Retweets, View Negative or Shameful Retweets, Find Rank For All Tweets ,Find Vote For All Tweets, Find Rating For All Tweets

- **View and Authorize Users**

In this module, the admin can view the list of users who all registered. In this, the admin can view the user's details such as, username, email, address and admin authorize the users.

User:

In this module, there are 'n 'numbers of users present. Users should register before doing some operations. After registration successful he must wait for admin to authorize him and after admin authorized him. He can login by using an authorized username and password. Login successful he will do some operations like View My Profile, Search Friend and Find Friend Request, View All My Friends, Create Tweets, View All Tweets, Search Tweets By Keyword, View All My Friends Tweets and Recommend, View All My Friends Shared Tweets Details, View All Recommended Tweets and Recommend.

5 RESULTS AND DISCUSSION

SCREEN SHOTS

HOME PAGE

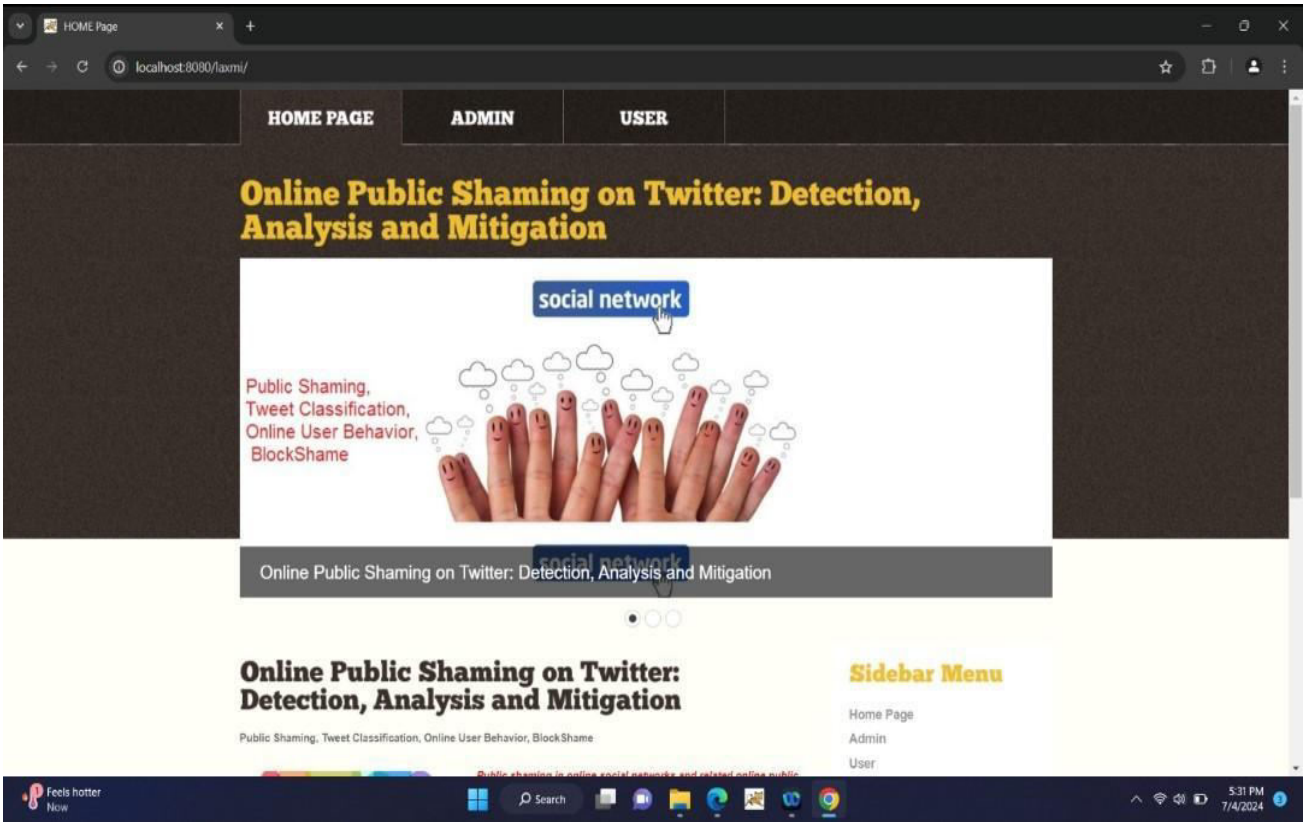

Fig5.1: HOME USER LOGIN PAGE

USER LOGIN

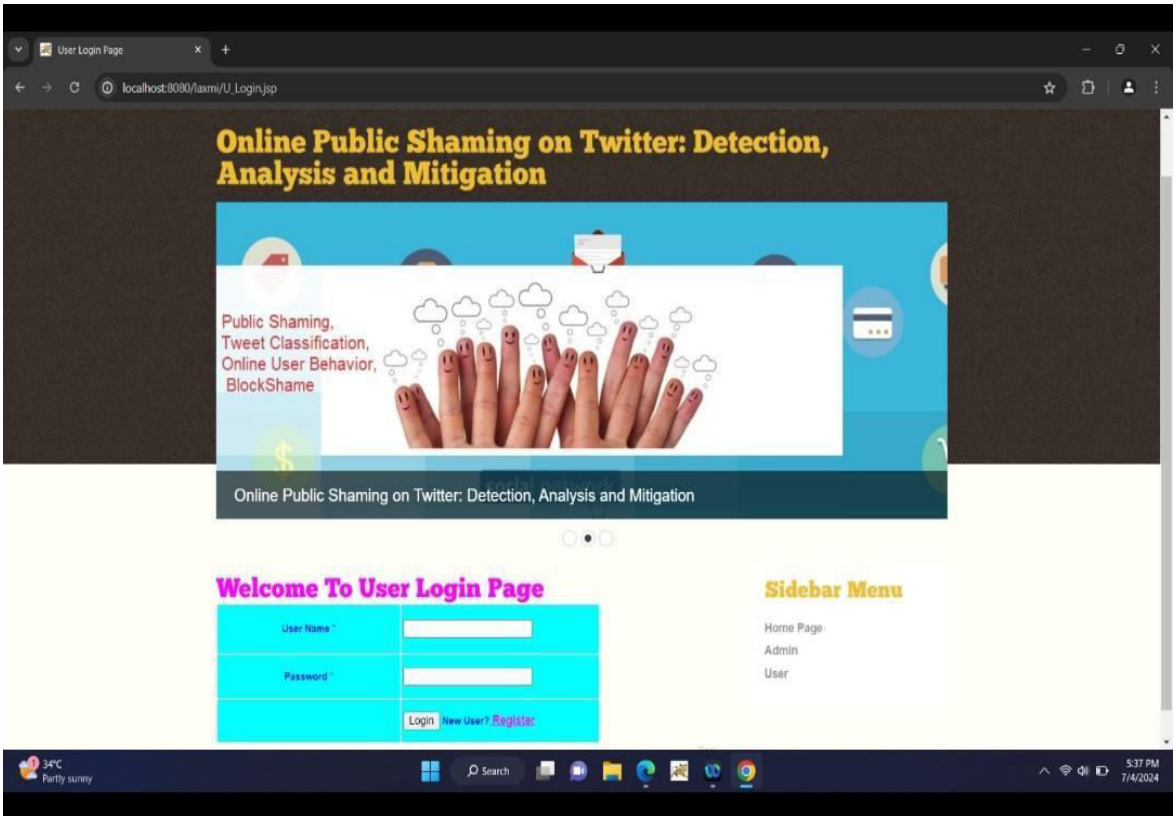

Fig5.2:User Login

USER REGISTRATION

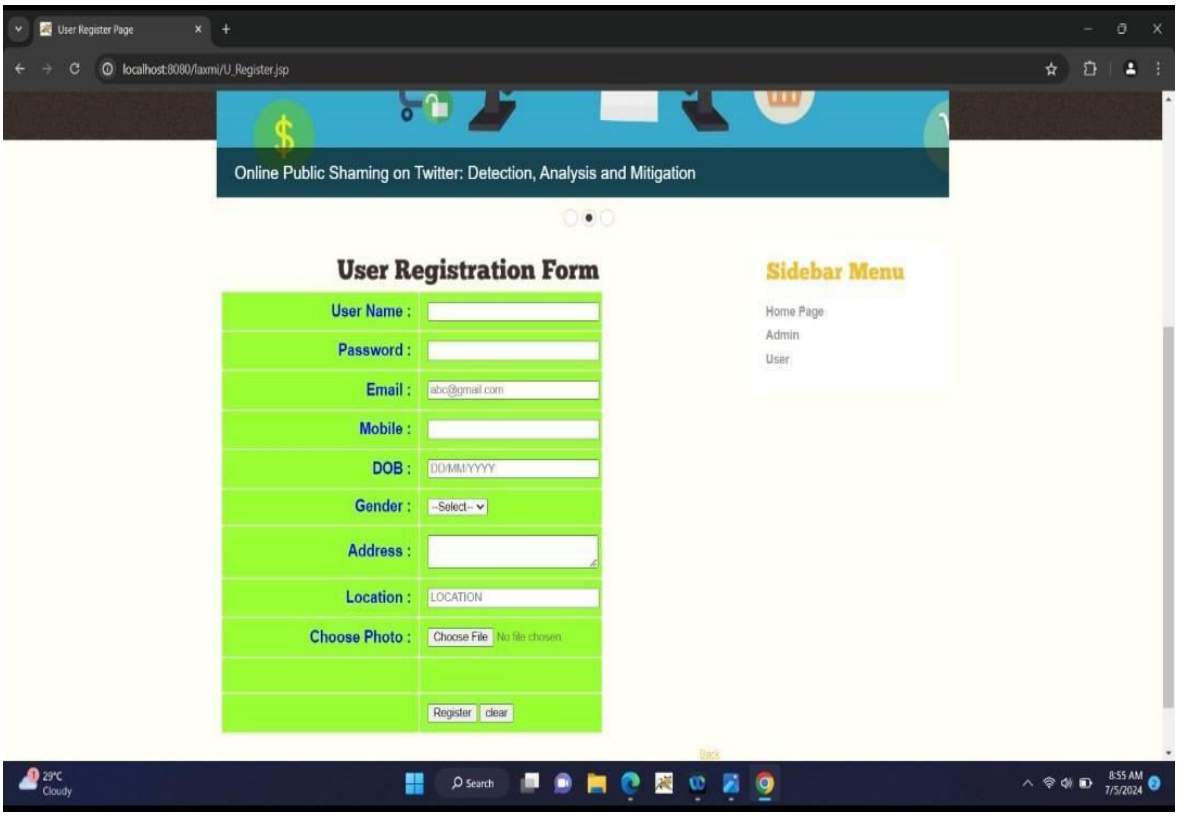

Fig5.3: USER REGISTRATION

USER PROFILE DETAILS

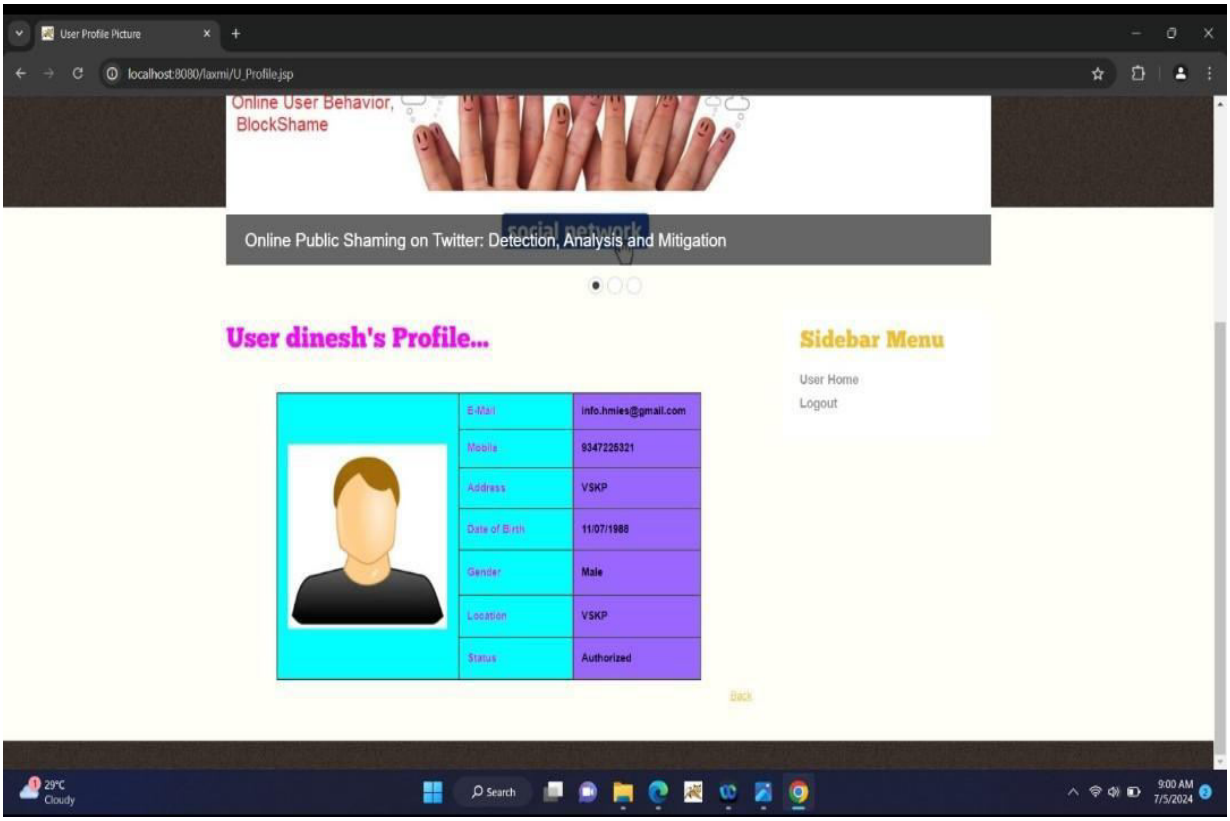

Fig5.4: USER PROFILE DETAILS

SEARCH FRIENDS

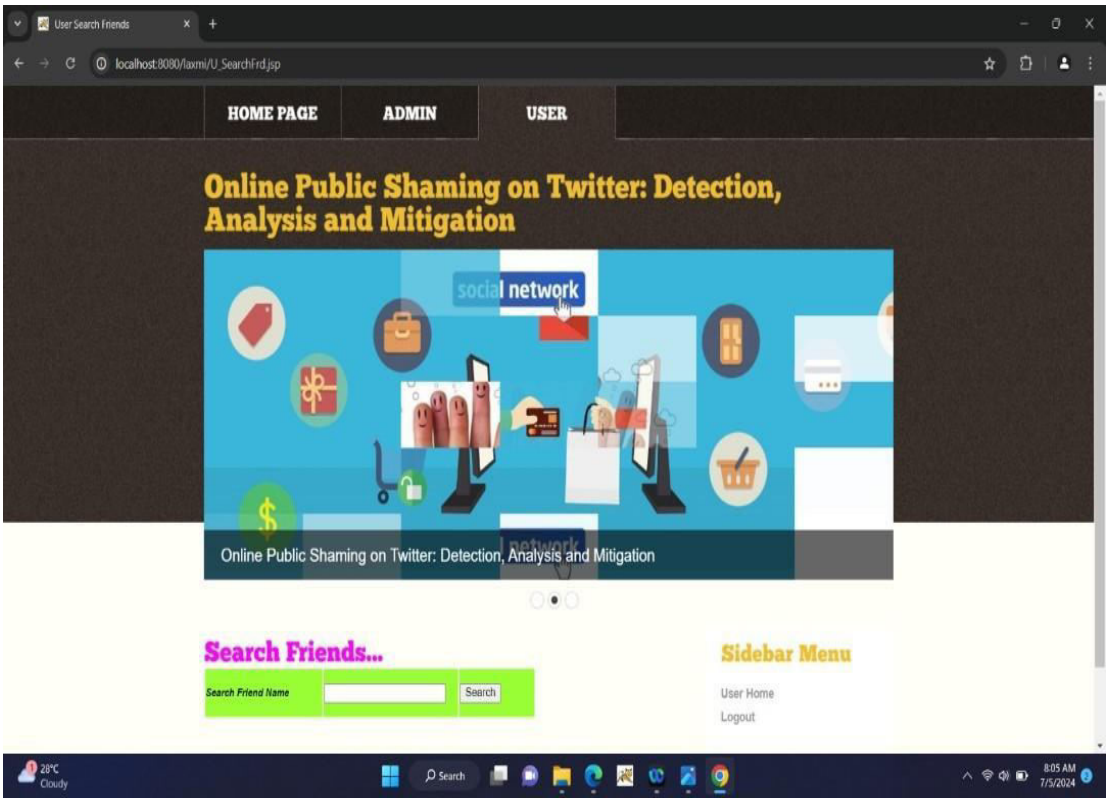

Fig5.5: SEARCH FRIENDS

VIEW FRIEND REQUESTS

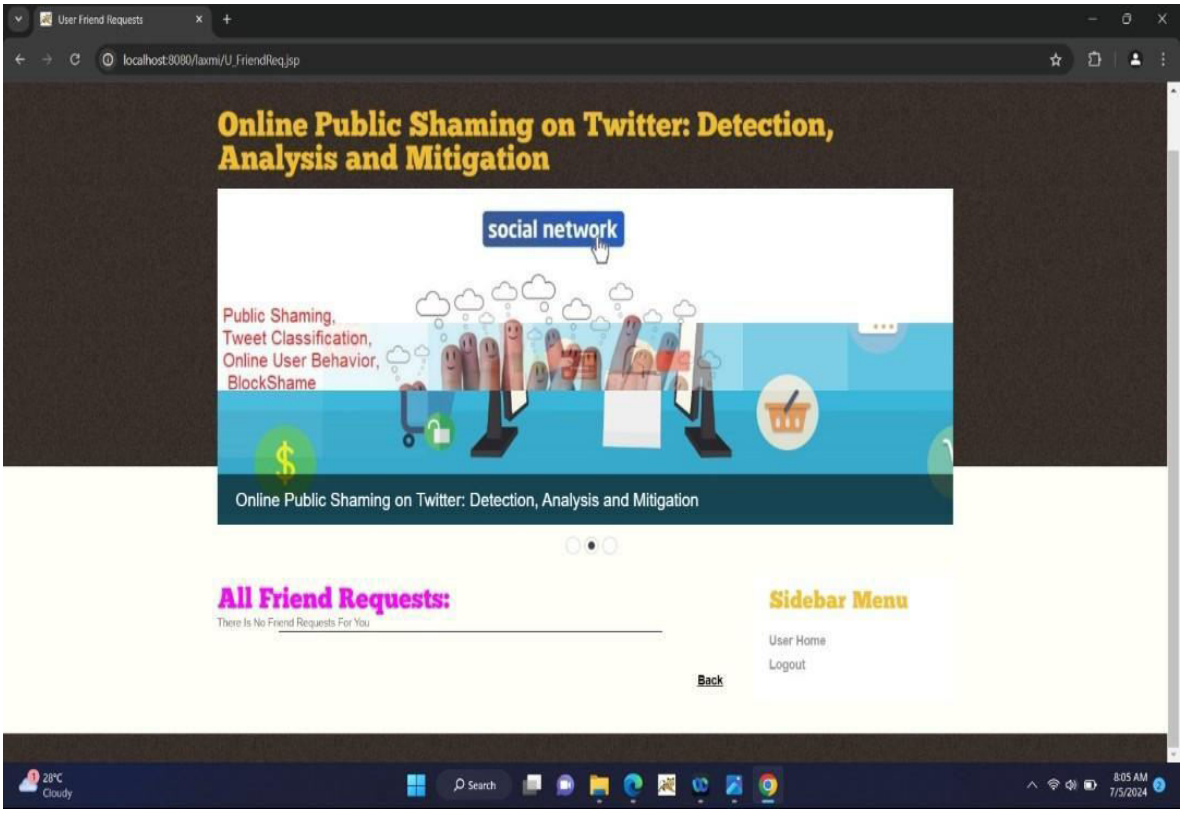

Fig5.6: VIEW FRIEND REQUESTS

ADMIN PAGE

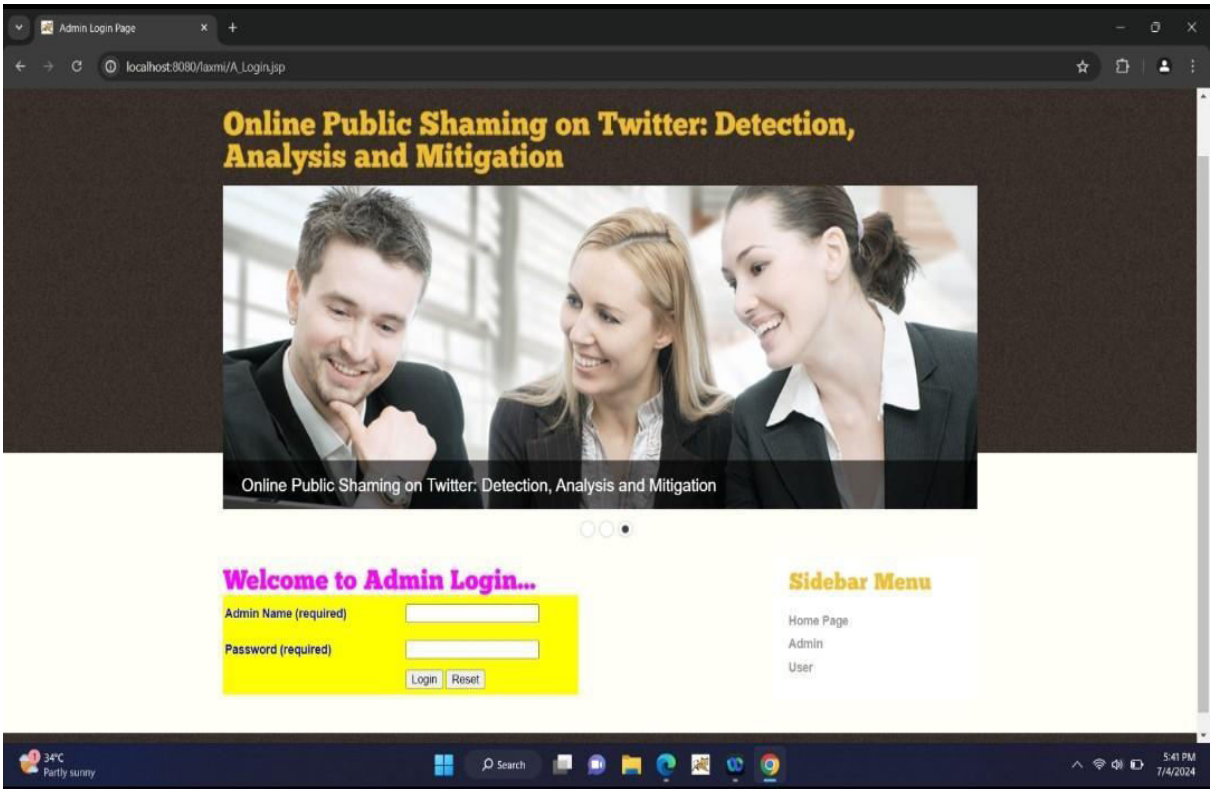

Fig5.7: ADMIN PAGE

ALL FRIEND REQUESTS AND RESPONSE DETAILS

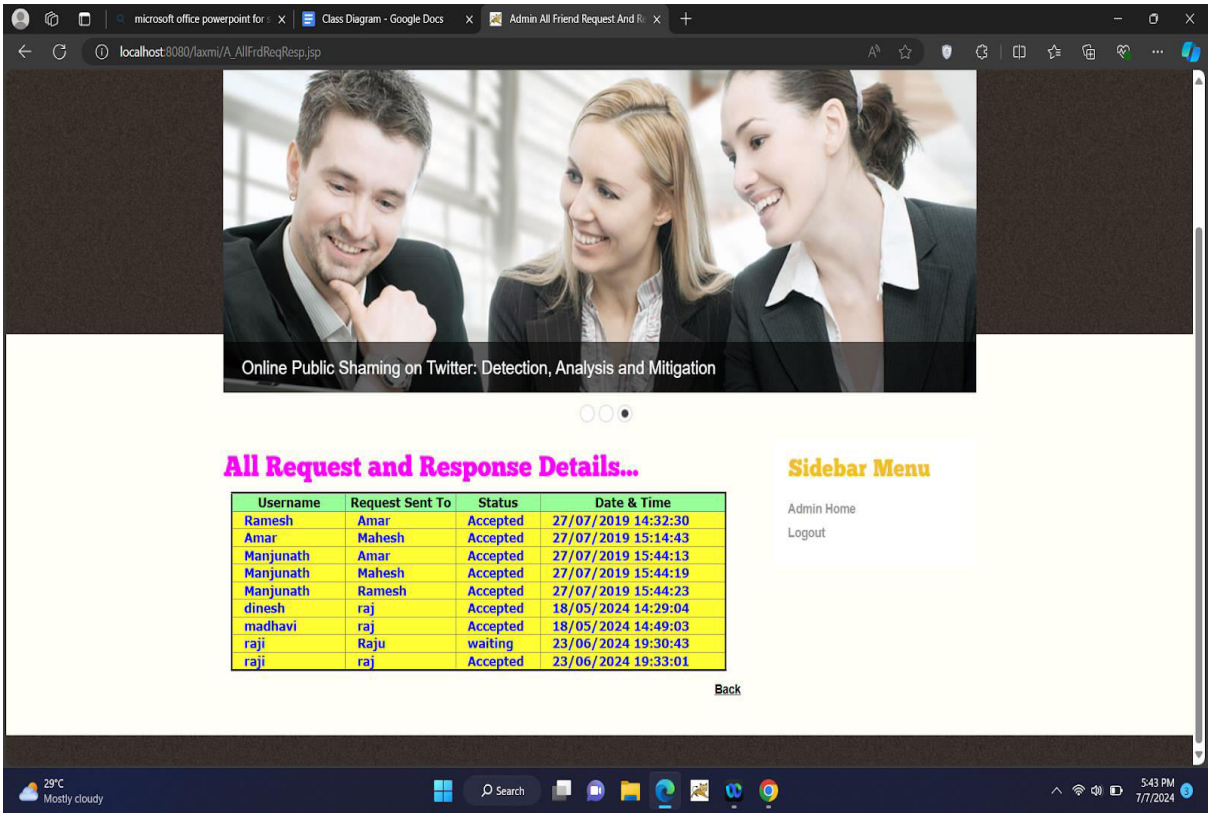

Fig5.8:ALL FRIEND REQUESTS AND RESPONSE DETAILS

6. CONCLUSION AND FUTURE WORK

CONCLUSION

Throughout the research it has been evident that basic details of a criminal activities in an area contains indicators that can be used by machine learning agents to classify a criminal ctivity given a location and date. Even though the learning agent suffers from imbalanced categories of the dataset, it was able to overcome the difficulty by oversampling and under sampling the taset. Through the experiments, it can be seen the imbalanced dataset was benefitted by using ENN under sampling. Using the under sampled data, Adaboost decision tree successfully classified criminal activities based on the time and location. With a accuracy of 81.93%, it was able to outperform other machine learning algorithms. Imbalanced classes are one of the main hurdles to achieve a better result. Though the machine learning agent was able to predictive model out of simply crime data, a demographic dataset would probably help to further improve the result and solidifyit.

REFERENCES

[1] J. Ronson, So You've Been Publicly Shamed. Picador, 2015.

[2] E. Spertus, "Smokey: Automatic recognition of hostile messages," in

AAAI/IAAI, 1997, pp. 1058–1065.
[3] S. Sood, J. Antin, and E. Churchill, “Profanity use in online communities,” in Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. ACM, 2012, pp. 1481–1490.
[4] S. Rojas-Galeano, “On obstructing obscenity obfuscation,” ACM Transactions on the Web (TWEB), vol. 11, no. 2, p. 12, 2017.
[5] E. Wulczyn, N. Thain, and L. Dixon, “Ex machina: Personal attacks seen at scale,” in Proceedings of the 26th International Conference on World Wide Web. International World Wide Web Conferences Steering Committee, 2017, pp. 1391–1399.
[6] A. Schmidt and M. Wiegand, “A survey on hate speech detection using natural language processing,” in Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media. Association for Computational Linguistics, Valencia, Spain, 2017, pp. 1–10.
[7] Hate-Speech, “Oxford dictionaries,” retrieved August 30, 2017 from https://en.oxforddictionaries.com/definition/hate speech.
[8] W. Warner and J. Hirschberg, “Detecting hate speech on the world wide web,” in Proceedings of the Second Workshop on Language in Social Media. Association for Computational Linguistics, 2012, pp. 19–26.
[9] I. Kwok and Y. Wang, “Locate the hate: Detecting tweets against blacks.” in AAAI, 2013.
[10] P. Burnap and M. L. Williams, “Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making,” Policy & Internet, vol. 7, no. 2, pp. 223–242, 2015.
[11] Lee-Rigby, “Lee rigby murder: Map and timeline,” retrieved December 07, 2017 from https://http://www.bbc.com/news/uk-25298580.
[12] Z. Waseem and D. Hovy, “