

CampusQueries – A Community-Based Question and Answer Platform Integrated with Toxic Comment Classifier

Saurav Yadav^{*1}, Ankit Maurya^{*2}, Abhishek Jha^{*3}

^{*1*2*3} Savitribai Phule Pune University, Computer Engineering, D.Y.Patil School Of Engineering Academy, Ambi, Talegaon-Dabhade, Pune, Maharashtra, India

Abstract: When we want any information and answers to our questions related to our college or campus, we search it on the web but we don't always get what we were finding. And also, we discuss or share opinions on social platforms but such activities sometimes encounter threats or harassments which compel people to not express themselves properly. Many social platforms try to find out such harassments or threats in conversations so that such conversations can easily be prevented before it causes any further damage. Toxicity detection is one of such methodologies to find out the different types of conversations that can be classified as toxic in nature. To increase the efficiency in classifying such comments, we can make use of machine learning algorithms to determine the toxicity in comments. This analysis aims in developing a platform that will help us in finding the solution and we named it "CampusQueries". In this proposed system, the authentic users will be able to ask doubts as well as clear others, also we are using a machine learning model for toxic comment classification. In this model, many toxic comments have been fed to build a Bidirectional Long Short-Term Memory (LSTM) Recurrent Neural Network (RNN) model for fulfilling the purpose. So, anyone can fearlessly put their points.

Keywords: Bidirectional Long Short-Term Memory (LSTM), Recurrent Neural Network (RNN), Machine Learning, OTP, Authentication, Toxic comment classifier, GloVe, CNN.

INTRODUCTION

The Internet completely changes our lives. When we want any information and answers to our questions related to our college or campus, what do we do? Just ask Google and hope there's an immediate answer? Filter search results to find the relevant answer? And when we find one, how we get to know that it's the correct one? This analysis aims in developing a platform that will help us in finding the solution and we named it "CampusQueries". It is an engaging site that feels like a duet of search engines and social media networks, but its purpose is to give you the answers and allow you to answer the questions asked by others of subjects in which you're knowledgeable. Introduction in simple words, CampusQueries is a question-and-answer platform for your college campus. It is a platform for students & staff where they can ask questions regarding their campus, related studies, hostel, canteen, fest, and much more. Providing a platform where we can ask easy questions without revealing our identity. Here, we can search for questions that we have and find out their respective answers. And can also post our questions if not already found on the platform, or if we want then we can provide answers to the questions of subjects that we have some knowledge about. But in Community-Based Question and Answer platforms, there is a chance that someone will use Hate Speech or Toxic comments (i.e., comments that are rude, disrespectful, or otherwise likely to make someone leave a discussion) that lead many such communities to limit or completely shut down user comments. To prevent that from happening we are making a machine learning model for toxic comments classification considering various machine learning algorithms and applying the algorithm having the best accuracy and low loss. In this model, many toxic comments have been fed to build a Bidirectional Long Short-Term Memory (LSTM) Recurrent Neural Network (RNN) model for fulfilling the purpose. We are using TensorFlow and Keras libraries for the deep learning concepts and Flask API for connecting our website with our ML model. This model will be working in the background. Here, the post (i.e., either question or answer or comment) will be taken as an input to the model then the model will predict if it's toxic or not and will give output, then its output will be used as an input in the site so that if it is toxic the admin will get the reference to the post to check it then after it the admin will verify that post, the suitable action against the post will be taken. The project principally includes 3 sections particularly website for users,

backend for admin, and ML model. The project is composed of a website in which the users will put on their queries. The ML model will be checking each post in the background for toxicity and the admin will verify the toxic posts sent by the model in the backend. The only requirement to use this project is that the user is supposed to be from the campus with a valid ID number. Because to sign up on the website, the user will require id number provided by the college. After checking the id in the database, it will give two options to the user to choose either from the Phone number or the email id to get an OTP for user authentication. After choosing from the options the user will get an OTP on the phone or mail which they chose earlier that is present in the campus database already. On successfully verifying the OTP the user will be allowed to provide the password and confirm the password for account creation. If a user forgot their password, then they can use forgot password option to get an OTP and then create a new password. For signing in the user will require its ID and Password, after successfully verifying both the user will be able to use the website for posting queries. The user will be able to post questions, answers, or comments and will be able to manage its posts. Or the user can manage their passwords also. The user's identity like its name, phone, or email will not be revealed to others, only their id will be visible, it is for the privacy of the user. The admin will have the user's information like Name, Phone, Email, Id in its database, which will be hidden from the users. The admin can manage the users i.e., the admin can take actions against the account and can also delete the posts. For using this platform, the user must have to go through the authentication process in which the user's id and password will be matched with the authenticated users. If the user is authenticated then only, he/she can log in to the system

II.PROPOSED SYSTEM

A. Problem Definition:

Machine learning models are an excellent solution to many of our problems since they are also used in AI, which is the future as well as current in the technology. Using this technology, we can easily classify the text that is being uploaded online. There is a lot of Hate speech, abusiveness, and toxicity going on the internet nowadays which sometimes leads to mental harassment, and by that someone can quit the conversation, social media and might affect their lives also. For a solution either the platform people will sit and check everything posted online which is a very Time-Consuming process and will require a lot of human effort also or we can simply make a Machine Learning model to process the text and predict itself if it is toxic or not. This technological idea is capable to get rid of toxicity and replace human efforts as well as time consumption to a great extent and is quite efficient as well. For this purpose, we propose a Question and Answering Platform integrated with the Toxic Comment Classification model using various Machine Learning algorithms for our campus so that every student and staff's doubts will be resolved as well as Hate speech and Toxicity will not be present.

B. System Architecture:

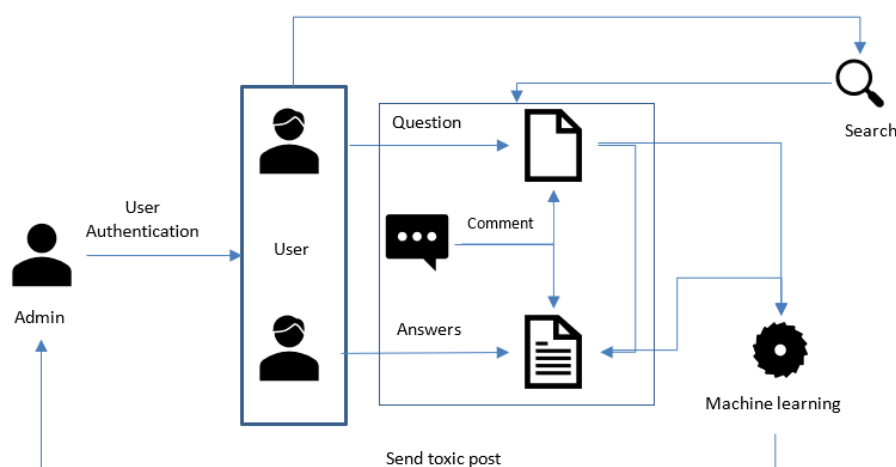


Fig. 2.1. Life cycle of a post in CampusQueries

The life cycle of a Post in CampusQueries site goes through RNN model and send post to the admin if found Toxic as shown in Fig.2.1

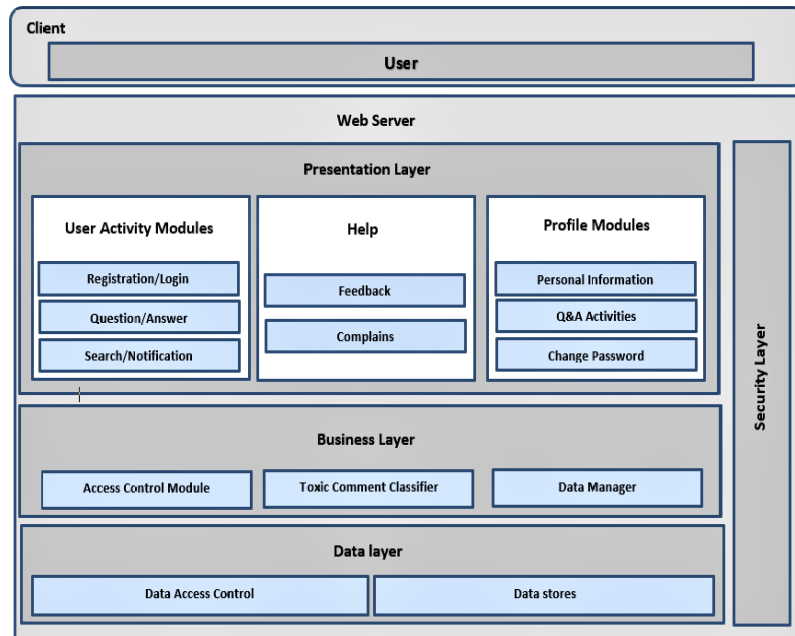


Fig. 2.2. System Architecture

C. Explanation of System Architecture:

a) A common layered architecture has been developed based on the client-server model as shown in Figure 2.2. On the client side, user agents are used for the presentation layer, business layer, and data layer. The security layer is shown as cross-cutting because security issues are mandatory and common to all layers. The security cross-cutting layer of the CampusQueries system supports operations like authorization, authentication, exception management, and validation.

b) Presentation layer

The presentation layer facilitates user interaction and consists of user interface components and presentation logic. This layer consists of three main components, i.e., user activity module, help, and profile module. The user activity module provides basic facilities like user registration, login, answering, and asking a query. The user activity module uses the services of data manager and access control module. CampusQueries site is based on user-generated content which is produced with collaborative efforts. The help module allows users to ask for help from the admin or provide feedback. This module also facilitates users to complain about abusive material. The third component of this layer is the profile module which uses the services of a data manager to display user's information, Q&A activities, and change the password.

c) Business layer

The business layer works as a mediator between the presentation layer and data layer and implements the core functionality of the CampusQueries system. CampusQueries business logic layer commonly consists of components like access control module, toxic comment classifier, and data manager. The access control module is responsible for controlling the access of community members consistent with their level, position, and authority by using predefined access control schemes. The toxic comment classifier is used to distinguish normal contents from bad ones and also, to send the post to the admin if found toxic. The data manager component facilitates the maintenance and retrieval of contents.

d) Data layer

The data layer consists of two components, namely, data access components and data stores. The CampusQueries data like profile information, questions, answers, and other necessary information is stored in data stores. The business layer accesses this data by exploiting the services of data access components.

D. Explanation of Model Architecture:

There are three stages in our system:

1) Preprocessing:

Preprocessing consists of removing any punctuation made on the reviews. Along with it any stop words, for example, "the", "a", "an", "in", "and" are removed. Any lemmatized words are also removed like say, words "am", "are", "is" will become "be". After all this we move on to the next stage Feature Extract

2) Feature Extract:

In feature extraction, we apply two algorithms, NLP and TOKENIZER. We use NLP to predict if reviews are positive or negative. This is done by a bag of words. TOKENIZER is used in splitting a phrase, sentence, paragraph, or entire text document into smaller units, such as individual words or terms. Each of these smaller units is called tokens.

3) Classifier:

We used RNN for classifying toxic comments because RNNs are a powerful and robust type of neural network, and belong to the most promising algorithms in use because it is the only one with an internal memory and because of their internal memory, RNN's can remember important things about the input they received, which allows them to be very precise in predicting what's coming next. This is why they're the preferred algorithm for sequential data like time series, speech, text, financial data, audio, video, weather and much more.

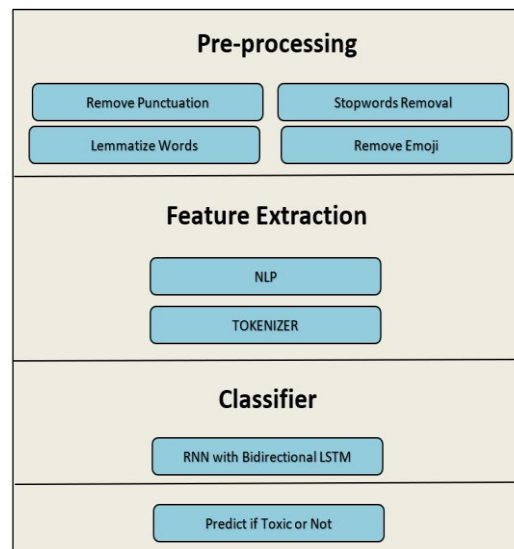


Fig. 2.3. Model Architecture

III.ML MODEL

A. Model Ideology

We discuss or share our opinions on social platforms but such activities sometimes encounter threats or harassments which leads some people to not express themselves properly. Many social platforms try to find out such harassments or threats in conversations so that such conversations can easily be prevented before it causes any further damage. Toxicity detection is one of such methodologies to find out the different types of conversations that can be classified as toxic in nature. To increase the efficiency in classifying such comments, we can make use of machine learning algorithms to determine the toxicity in comments. In this model, many toxic comments have been fed to build a Bidirectional Long Short-Term Memory (LSTM) Recurrent Neural Network (RNN) model for fulfilling the purpose.

B. Model Accuracy and Loss Plots

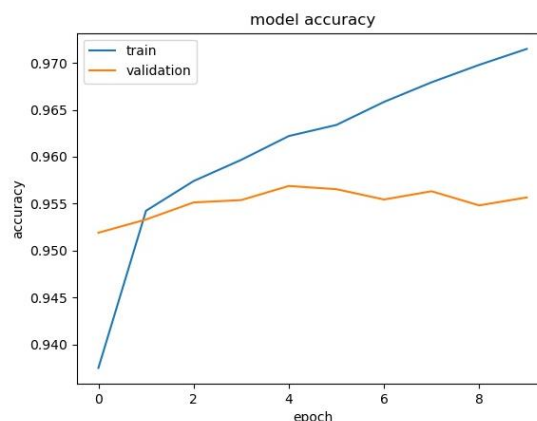


Fig. 3.1. Model Accuracy Plot Graph

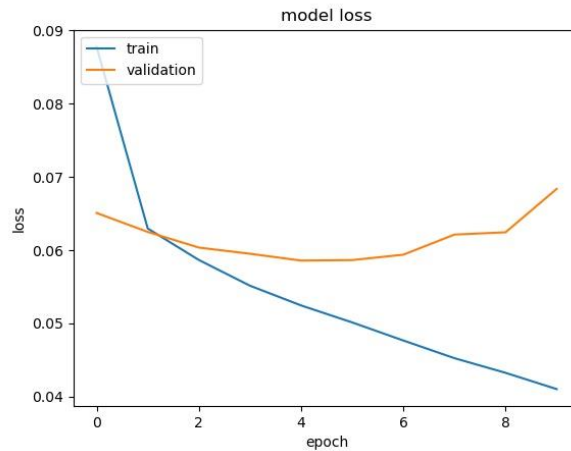


Fig.3.2. Model Loss Plot Graph

C. Algorithm

Input – Text

Output - Predict if toxic or not

- 1) Step 1 - Take the text from the dataset as input.
- 2) Step 2 - Start pre-processing the text by lowering all the text.
- 3) Step 3 - Next, remove all the uncommon signs from the text.
- 4) Step 4 - Expand the abbreviation in the text.
- 5) Step 5 - Correct all the misspelled words.
- 6) Step 6 - Remove the punctuations from the text.
- 7) Step 7 - Remove all the emojis.
- 8) Step 8 - Remove the stopwords, that were downloaded using NLTK.
- 9) Step 9 - Apply the lemmatisation (e.g., make "running"/"ran"/"run" into "run").
- 10) Step 10 - Now, create an embedding vector using Glove.6B.
- 11) Step 11 - Train a Recurrent Neural Network (RNN) with a Bidirectional LSTM layer.
- 12) Step 12 - Produce the output to calculate the model accuracy and loss.

IV.UML DIAGRAM FOR PROPOSED SYSTEM

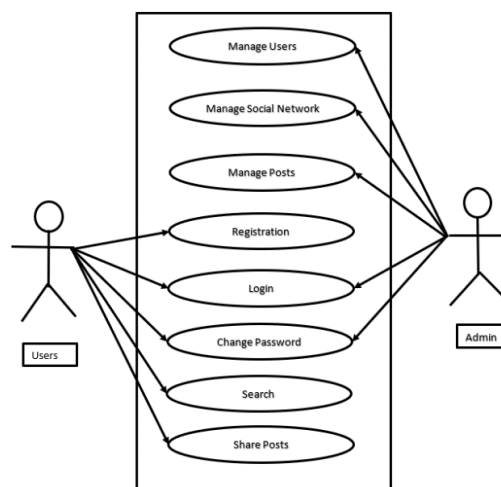


Fig. 3.1 UML DIAGRAM OF SYSTEM

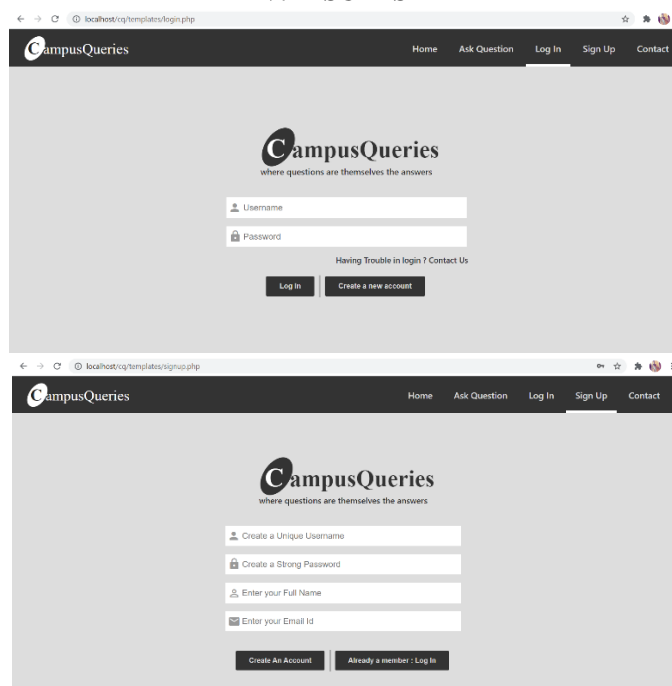
V.RESULTS

Fig. Login and Signup pages.

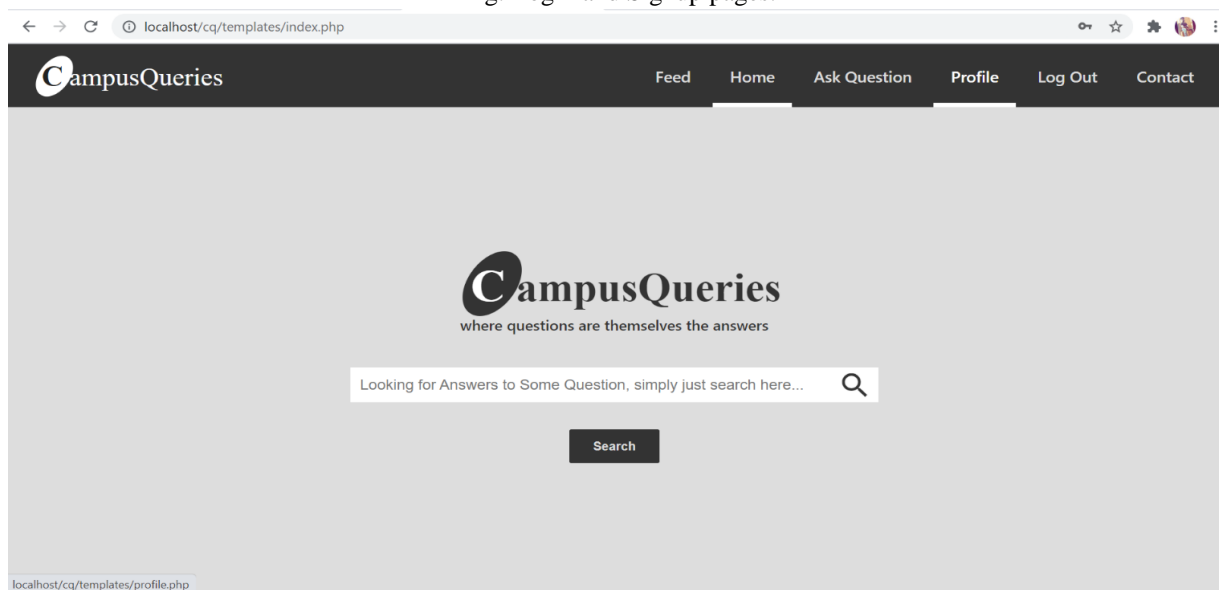


Fig. Search Page

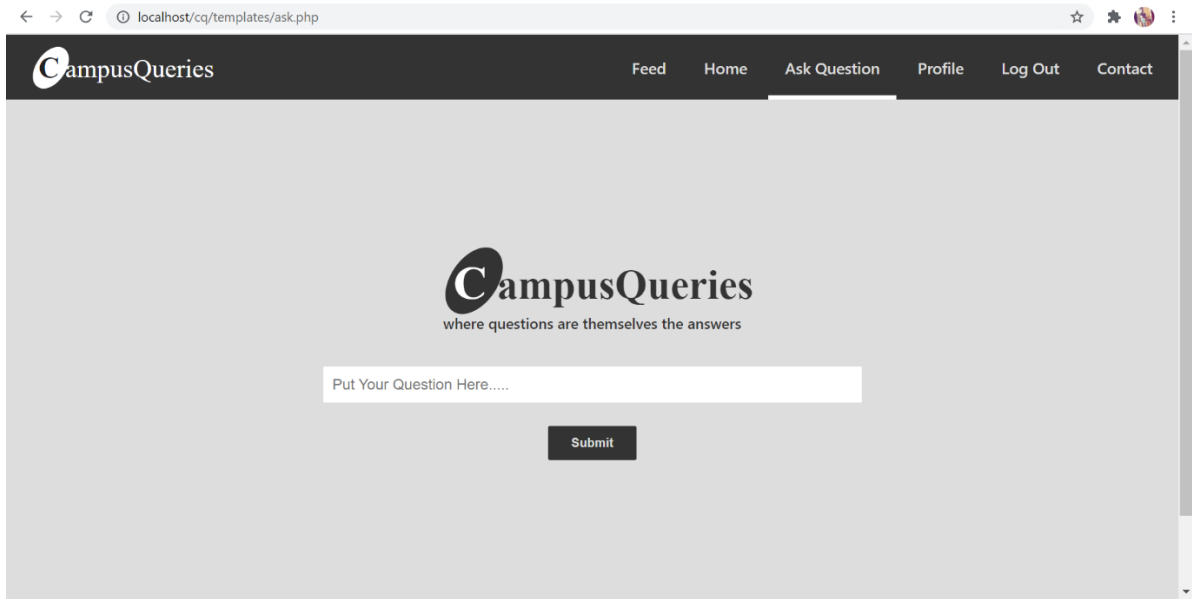


Fig. Ask Question Page

Testing ML Model

Step 1) User posts toxic and insulting answer.

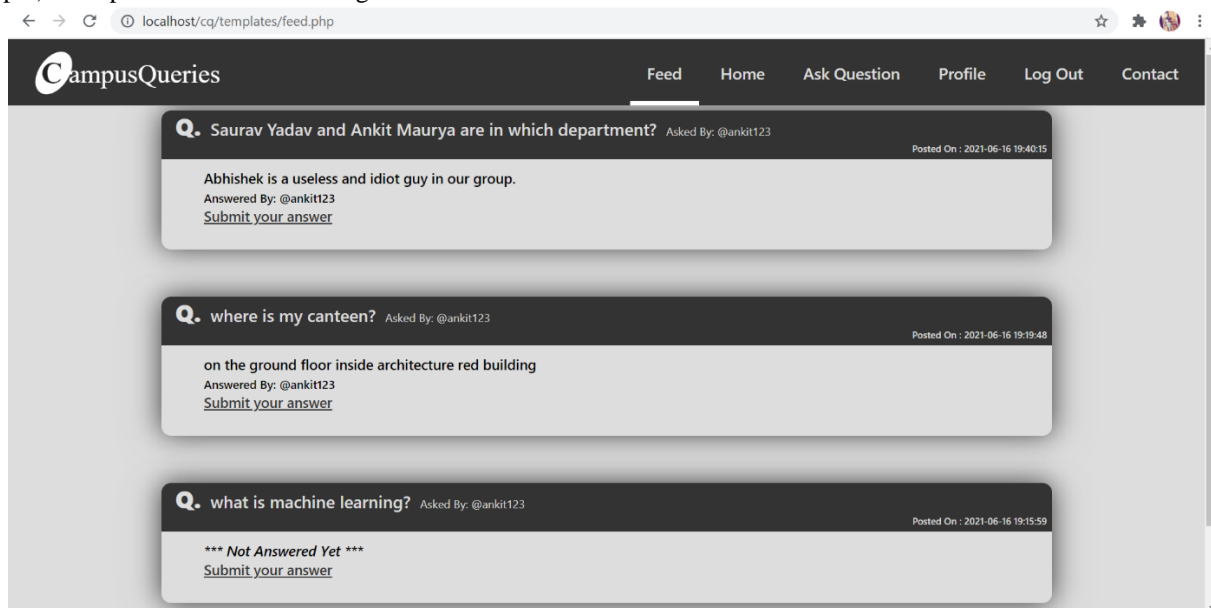


Fig. Feed Page

Step 2) Model checks the answers and questions and finds toxicity in answer and removes it from database.

```
Anaconda Prompt (anaconda3) - python ml.py
Toxic - 0.0837737
Very Toxic - 0.0017115772
Obscene - 0.011031032
Threat - 0.004057944
Insult - 0.015510559
Hate - 0.001955092
Neutral - 0.9137143

Saurav Yadav and Ankit Maurya are in which department?
1/1 [=====] - 0s 22ms/step
Toxic - 0.3026867
Very Toxic - 0.0054028034
Obscene - 0.055701524
Threat - 0.003305316
Insult - 0.08487046
Hate - 0.019432276
Neutral - 0.67637324

Abhishek is a useless and idiot guy in our group.
1/1 [=====] - 0s 24ms/step
Found toxicity in answer, Removing from the Database..!!!!
Toxic - 0.94859195
Very Toxic - 0.011677444
Obscene - 0.39583832
Threat - 0.0025200248
Insult - 0.638315
Hate - 0.011579961
Neutral - 0.03125325
127.0.0.1 - - [17/Jun/2021 21:07:46] "GET / HTTP/1.1" 200 -
```

Fig. Anaconda Prompt

Step 3) Answer is removed from the feed page.

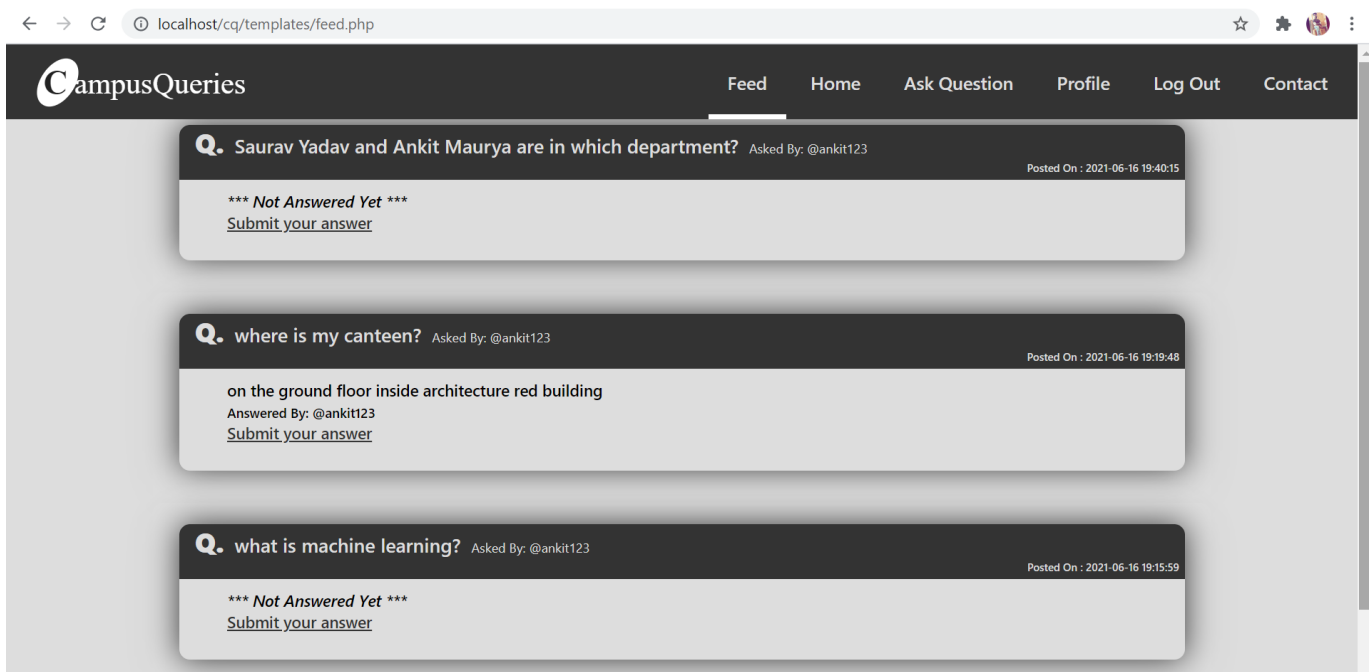


Fig. Feed Page

VI.CONCLUSION

To conclude, we know that nowadays technology is growing day by day, and there are so many Q&A websites and applications available online on which many users ask their doubt or solve others but some only throw hate speech and abusive content, which leads to fights and even some platforms disable their comment sections because of these, so, there is a need to make sure that there is no such content allowed to be posted. Here is where our project will help users resolving the doubts as well as asking the queries with surety of facing no bad content on the platform, so they can freely

put their queries or points without being abused. If anyone posts bad content then the toxic comment classifier will rectify it and if found toxic then it will send it to the admin and the admin will examine the post and will take suitable action.

REFERENCES

- [1] M. Anand and R. Eswari, "Classification of Abusive Comments in Social Media using Deep Learning," *2019 3rd International Conference on Computing Methodologies and Communication (ICCMC)*, Erode, India, 2019, pp. 974-977, doi: 10.1109/ICCMC.2019.8819734.
- [2] Kanagasabai, Thiruthanigesan & Jeyaperatha, T. & V.Balalojanan, & Thiruchelvan, Nagarathnam. (2017). "Secure Login Using Encrypted Password and Email Based Login Approach," *International Journal of Advanced Research in Computer Science & Technology*. 5. 44-48.
- [3] Chandra Sekharan, Sindhu. (2017). "APPLICATION OF SESSION LOGIN AND ONE TIME PASSWORD IN FUND TRANSFER SYSTEM USING RSA ALGORITHM.", *IEEE Conference - International Conference on Electronics, Communication and Aerospace Technology ICECA 2017*.
- [4] Hongyin, Yan & Xuelei, Qi. (2011), "Design and implementation of intranet search engine system.", *2011 International Conference on Mechatronic Science, Electric Engineering and Computer August 19-22, 2011, Jilin, China*.
- [5] H. Shen, G. Liu, H. Wang and N. Vithlani, "SocialQ&A: An Online Social Network Based Question and Answer System," in *IEEE Transactions on Big Data*, vol. 3, no. 1, pp. 91-106, 1 March 2017, doi: 10.1109/TBDATA.2016.2629487.
- [6] B. Mathew et al., "Thou shalt not hate: Countering online hate speech," *Proc. 13th Int. Conf. Web Soc. Media, ICWSM 2019*, no. August, pp. 369-380, 2019.
- [7] Z. Yang, I. Jones, X. Hu, and H. Liu, "Finding the right social media sites for questions," *ASONAM*, 2015.
- [8] Wulczyn, Ellery; Thain, Nithum; Dixon, Lucas: Wikipedia Detox. figshare. doi.org/10.6084/m9.figshare.4054689 (2016)
- [9] 'Sentiment Analysis of Movie Review Using Supervised Machine Learning Techniques' Published By: Gurshobit Singh Brar, Asst. Prof. Ankit Sharma, Published on: Nov 2018
- [10] J. Cheng, C. Danescu-Niculescu-Mizil, and J. Leskovec, "Antisocial behavior in online discussion communities," *Proc. 9th Int. Conf. Web Soc. Media, ICWSM 2015*, pp. 61-70, 2015.
- [11] H. M. Saleem, K. P. Dillon, S. Benesch, and D. Ruths, "A Web of Hate: Tackling Hateful Speech in Online Social Spaces," 2017, [Online]. Available: <http://arxiv.org/abs/1709.10159>.

BIOGRAPHY



Saurav Yadav, Dr. D.Y. Patil School of Engineering Academy, Ambi, Pune, India
Studying in the field of Computer Engineering,



Ankit Maurya, Dr. D.Y. Patil School of Engineering Academy, Ambi, Pune, India
Studying in the field of Computer Engineering,



Abhishek Jha, Dr. D.Y. Patil School of Engineering Academy, Ambi, Pune, India
Studying in the field of Computer Engineering,