

Predicting the Popularity of Reddit Posts

**Prepared By**

Meghna Nair

Sai Harsha Tamada

Chennoufi Mohammed Soheib

Akshith Reddy Kota

Rahul M

Sandeep Yadav

Amiruddin(160961020200)

Lakshmi Girija Mangipudi (1CR15IS050)

Alina Sahoo (1701208003)

# Requirement Analysis:

The project is about predicting score for a particular Reddit post. The application takes URL of the Reddit post as input and displays the score for the post as output. The score for a Reddit post is the combination of upvotes and downvotes. It was identified as a regression problem. The hardware and software specifications were listed for the project for verification in the design phase.

# Feasibility Study:

This phase is usually carried out after closely verifying the requirements of the project. The dataset used for predicting the popularity was well-scraped. Top 10,000 posts were scrapped according to the user’s need using Praw library. Features like Title, Gilded, Over\_18, Upvotes, Downvotes, Number of comments were extracted. The tile was in text format while other features were numbers. So, Sentiment analysis for the title of the post was done using vaderSentiment analyzer. We get 4 columns neg, neu, pos and compound which were combined to one column, Predited\_value, using the value of compound column. Text preprocessing was performed on the title of the post by removing punctuations and stop words, stemming and lemmatization. To convert the title to numeric form, Glove embedding with 100 dimensions was used. This gives a numeric vector for all the unique words in the text. To get one vector representation for each title, weighted average of word vectors method was used. Later one hot encoding for the Predicted\_value and Over\_18 columns was performed. All the unnecessary columns were removed.

# System Design:

After studying through all the phases of feasibility study, we now look for our system specifications for the design phase. The hardware specifications used in this project is: Windows 10, 8 GB RAM, AMD Ryzen 5 Hexa Core Processor. However, the software specifications used in this project is Python 3.8 with Spyder/Jupyter Notebook as it’s environment.

# Implementation:

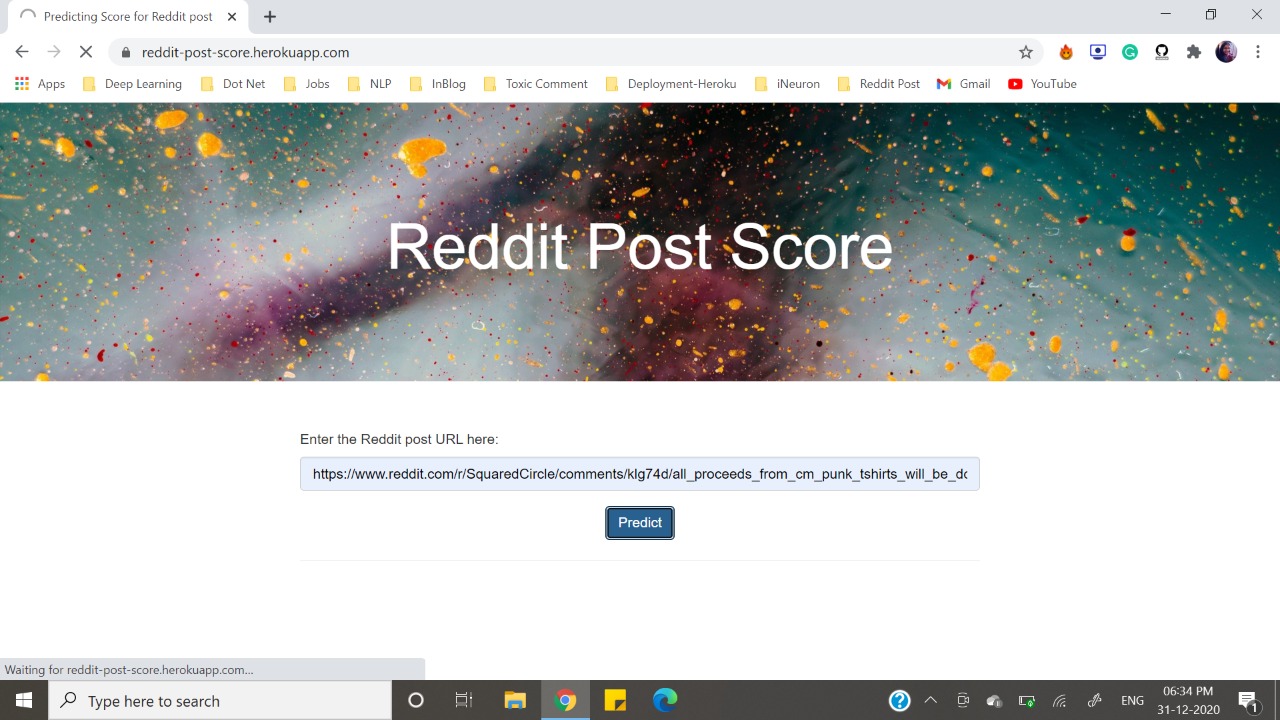
The regression models used are Linear regression, Decision tree regressor, Random forest regressor, KNN regressor, Lasso, Ridge, ElasticNet and XGBoost regressor. Out of all these models, XGBoost regressor and Random forest regressor performed well with the dataset with around 50% accuracy on test dataset.

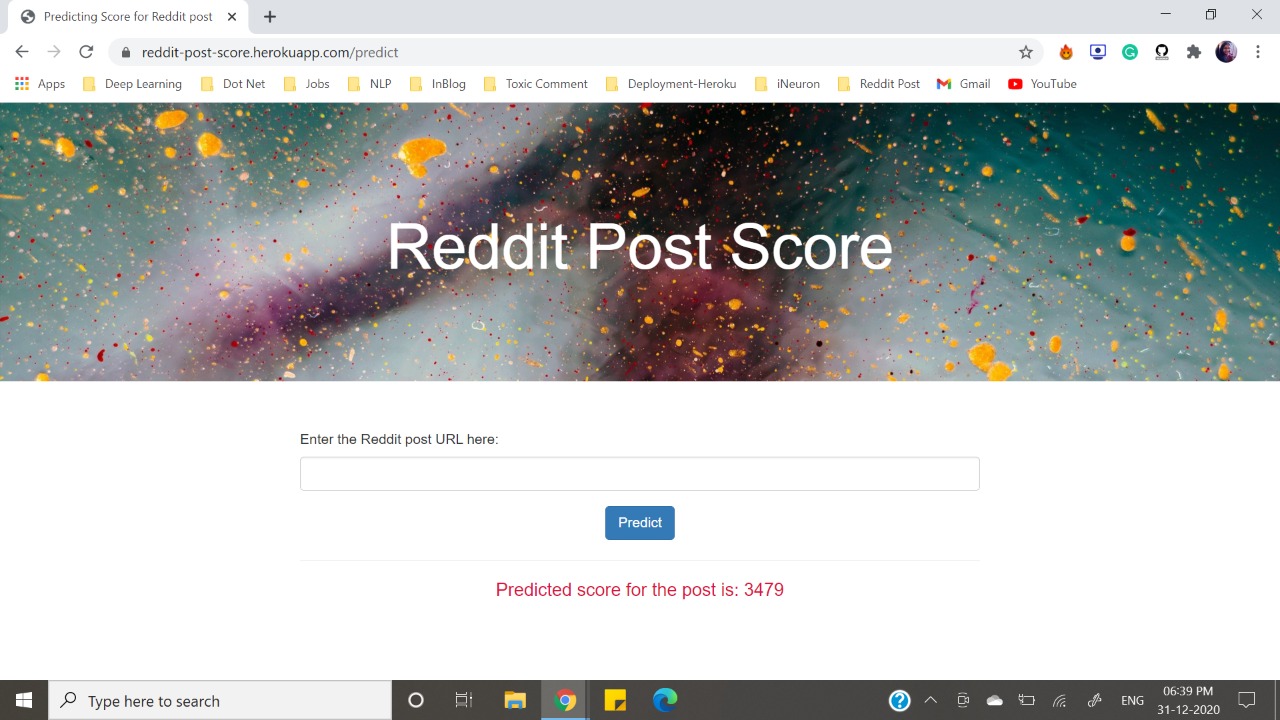
# Testing:

The deployed application was tested with different Reddit post URLs. As the accuracy of the model is around 50%, the predictions were a little different from expected. In future work, we would use more data to train the model to get good accuracy.

# Deployment:

The deployment was done using Heroku. The required HTML files were prepared and the application was deployed on Heroku.





# Maintenance:

The project can be further extended using other features, thus improving the accuracy of the models using ML algorithms. Also, the UI design can be enhanced and made more interactive.