Software documentation

Aryan Verma, Srijan Saini, Tarun Gupta, Kartik Garg

May 30, 2020

Contents

1 Product Documentation

1.1 Product Overview

Our product contains a large dataset of Hindi reviews and a deep learning NLP model for performing sentimental analysis on the dataset. The dataset has been collected by us by scrapping of movie reviews from various Hindi News Websites. We provide the exploratory analysis of Hindi text of the dataset. We have created a website to deploy our NLP model. In the website, we have also added a relevant functionality to search movie reviews.

1.2 Roles and Responsibilities

Table 1

Team member	Responsibilities
Tarun Gupta	Web Scrapping, NLP modeling & testing and data exploratory analysis.
Aryan Verma	Web Scrapping, NLP modeling and UML diagrams.
Kartik Garg	Web Scrapping, Model deployment using FLASK, data exploratory analysis and UML diagrams.
Srijan Saini	Web Scrapping, Website management and model deployment using FLASK.

1.3 Team goals and business objective

- To create the largest Hindi movie review dataset.
- To create an NLP model to obtain state-of-the-art results on the created dataset.

1.4 Background and strategic fit

- Strategic aim for making this product is to facilitate the researchers who wish to pursue their research in the field of natural language analysis of Indic languages such as Hindi.

- Large datasets for sentiment analysis are available for English language. The datasets available for Indic languages are quite small. Our aim is to provide the largest movie review dataset for Hindi language.

2 User Experience Design documentation

2.1 User Stories

Table 2: User Stories Table

Requirements		
User Story Title	Story Description	Priority
Sentiment Analysis Model	A user wants to have a Sentiment Analysis Model along with dataset	Must have
Sentiment Probability	A user also wants to know the probability of the input review belonging to a particular sentiment in the results page	Must have
Overview of Dataset	Along with dataset, users also want to see the basic analysis of dataset(i.e. Exploratory Analysis)	Must have
Search movie reviews	A developer wants a integrate scrapping code with the model to scrape movie reviews on user requirement	Should have

3 System Modelling

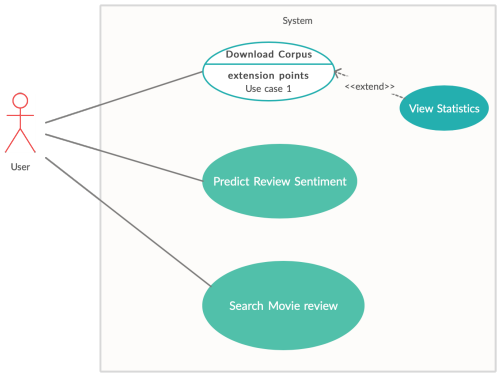


Figure 1: Use Cases.

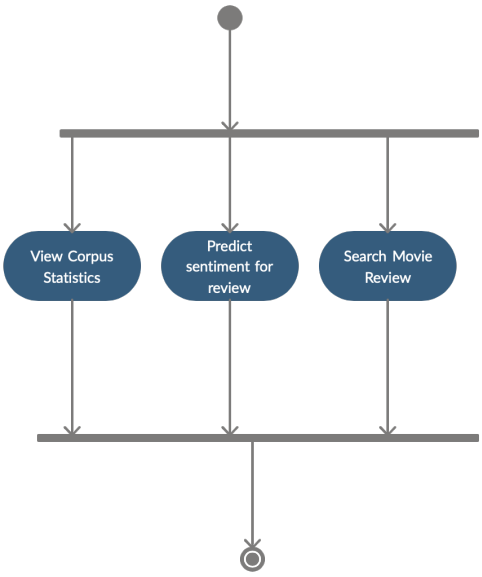


Figure 2: Process diagram.

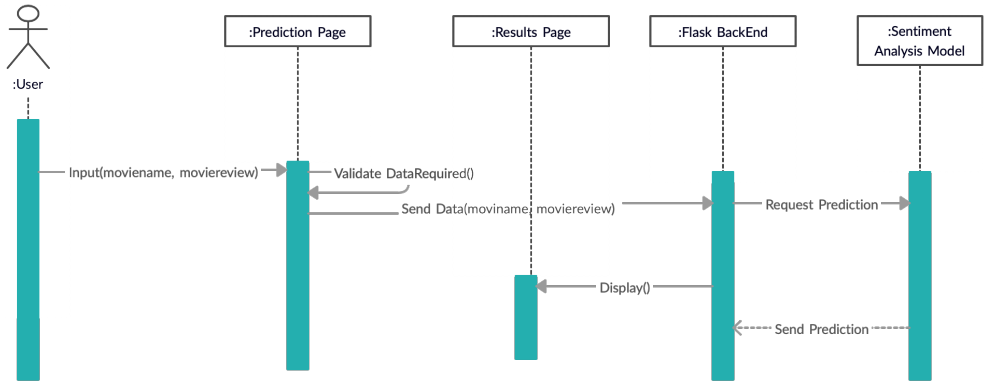


Figure 3: Sequence diagram for use case: predicting review sentiment.

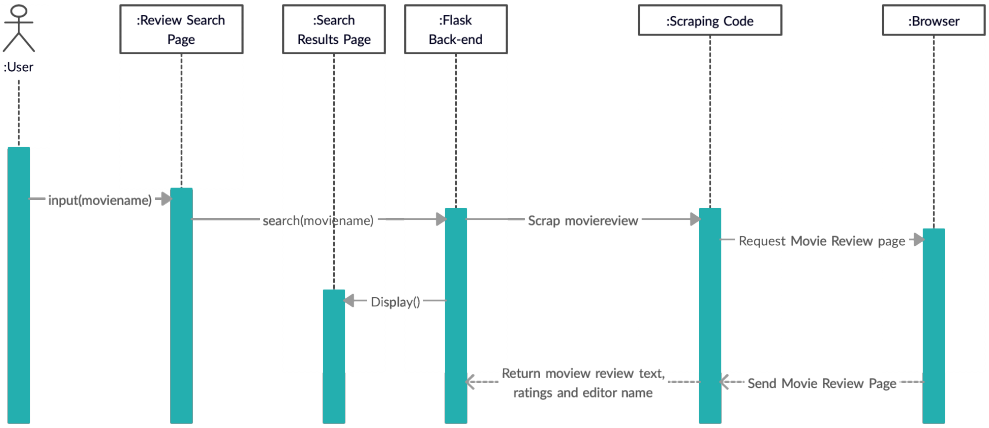


Figure 4: Sequence diagram for use case: searching movie review.

4 Software Architecture

4.1 Dataset Design

We provide a total 12 datasets, all stored in Comma-Separated Values (CSV) format:

- "1_AajTakDataset.csv" contains reviews taken from AajTak website.
- "2_FilmiBeatDataset.csv" contains reviews taken from FilmiBeat website.
- "3_JagranDataset.csv" contains reviews taken from Jagran website.
- "4_NavBharatDataset.csv" contains reviews taken from Navbharat website.
- "5_NavBharatDataset_Cleaned.csv" is the cleaned up version of reviews taken from NavBharat website (removing reviews without any rating).
- "6_Final_dataset.csv" - Vertical stacking of reviews taken from above 4 websites. Reviews having no rating have been assigned -1 rating.
- "7_Final_dataset_2.0.csv" - Same reviews as in "6_Final_dataset.csv", with removal of reviews with -1 rating.
- "8_Final_dataset_3.0.csv" - Reviews in "7_Final_dataset_2.0.csv" with rating ≥ 3 were assigned sentiment 1 (positive), and 0 (negative) sentiment otherwise. Further, removal of stop words and some rows with corrupted reviews was done.
- "9_Final_dataset_3.0_Augmented_only.csv" - For every review in "8_Final_dataset_3.0.csv", a similar review was generated (data augmentation) using iNLTK library.
- "10_Final_dataset_3.0_ExtraInfo.csv" - Contains same data as in "8_Final_dataset_3.0.csv" plus some extra meta-information for each review.
- "train_3.0_Augmented.csv" - First 1500 reviews of "8_Final_dataset_3.0.csv" plus its augmented versions present in "9_Final_dataset_3.0_Augmented_only.csv", used for training of the NLP model.

- "val_3.0.csv" - Last 214 reviews of "8_Final_dataset_3.0.csv", used as validation set for training of the NLP model.

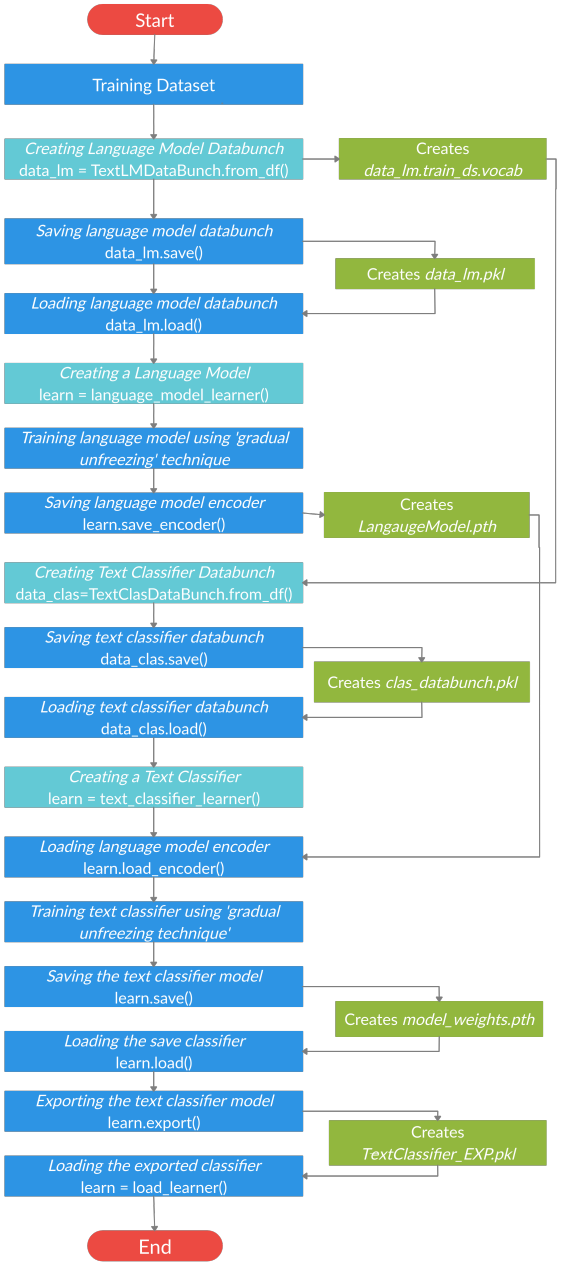
	Review	Sentiment	WordCount	ReviewLen
0	हीरो डबल रोल वाली फिल्में पिछले अरसे कईं रिलीज...	1	476	2662
1	होम फिल्म समीक्षा अक्षय बिना अधूरी सी वेलकम बै...	0	401	2235
2	'मैं मां हूं मां कोई सपने होते।' अश्विनी अय्यर...	1	473	2697
3	चंद्रमोहन शर्मा आमतौर दिवाली पहले वीक ऐसी फिल्...	0	303	1752
4	-अजय ब्रह्मात्मज कलाकार: नीलिमा अजीम, ईशान कौर...	1	282	1658
5	कहानी दिल्ली रहने वाली जोया सोलंकी अपने पिता भ...	1	230	1333
6	रेखा खान इश्क जुनून दीवानगी बॉलिवुड 'डर', 'गुप...	0	257	1497
7	फिल्म इंडस्ट्री ज्यादातर मौकों एक्टर्स काम मुश...	1	196	1066
8	डायरेक्टरः निशिकांत कामतकलाकारः जॉन अब्राहम, द...	0	306	1792
9	चंद्र मोहन शर्माकरीब दस साल पहले बिग बी ऐसी फि...	0	518	2779

Figure 5: Screen shot of "10_Final_dataset_3.0_ExtraInfo" dataset

For illustration, we have provided a screenshot of "10_Final_dataset_3.0_ExtraInfo.csv" dataset in Fig. 5:

- This dataset has 4 columns as shown in above figure:
  - Review - contains review of a movie in hindi language.
  - Sentiment - 1 represents positive sentiment and 0 represents negative sentiment.
  - WordCount - number of words in the review.
  - ReviewLen - total length of review in terms of number of letters.
- This dataset has total 1714 rows.

4.2 Architectural Design of NLP model



We use Universal Language Model Fine-tuning for Text Classification (ULMFiT) model from fastai library pre-trained on Hindi text. fastai pro-

vides a language model with Average-Stochastic Gradient Descent (SGD) Weight-Dropped LSTM (AWD-LSTM) architecture pre-trained on Hindi-text available to download. We create a learner object that will directly create a model, download the pre-trained weights and be ready for fine-tuning. Then, we train the language model using gradual unfreezing technique. Finally, we save the language model encoder to be able to use it for next tasks.

To build the text classifier, we use our fine-tuned encoder. We train the text classifier using gradual unfreezing technique. Finally, we save and load the text classifier model and export as a pickle file to be deployed.

4.3 Source Code Documentation

The PyTorch model used is `1.5.0+cu101`, and fastai model used is `1.0.61`. These can be installed easily to reproduce the results using the following commands:

- pip install torch==1.5.0+cu101 -f https://download.pytorch.org/whl/torch_stable.html
- pip install fastai==1.0.61

The different files present in the github repository at `https://github.com/KartikGarg19/Hindi-Text-Corpus` is described below:

- File: Model.py -
  - ULMFIT: Universal Language Model Fine-tuning for Text Classification (ULMFIT) is a transfer learning technique useful for many NLP tasks. It uses a 3 layer bi-LSTM with various tuned dropout hyperparameters. High level idea of ULMFIT is to train a language model using a very large corpus, then to take this pretrained model's encoder and combine it with a custom head model, e.g. for classification, and to do the good old fine tuning using discriminative learning rates in multiple stages carefully. ULMFiT consists of three stages-

    1. Training the language model on a general-domain corpus that captures high-level natural language features. (Which is done beforehand by fast.ai, the organisation that developed it, as this step is computationally very expensive.)
    2. Fine-tuning the pre-trained language model on target task data
    3. Fine-tuning the classifier on target task data

- File: augment_data.py -
  - iNLTK: Natural Language Toolkit for Indic Languages (iNLTK) library was used for data augmentation. *get_similar_sentences* function of iNLTK is used to create similar sentences for each review. The library is backed by ULMFiT Language Models which are trained using fastai and Pytorch libraries.

- File: ExploratoryAnalysis.ipynb and ExploratoryAnalysisTFA.ipynb -

  - Matplotlib and Seaborn library have been used to do exploratory analysis of reviews in the dataset.

- File: Flask.py -
  - Flask: Flask in used to deploy our model to the website. Flask is a lightweight WSGI web application framework. Flask offers suggestions, but doesn't enforce any dependencies or project layout. It is up to the developer to choose the tools and libraries they want to use.

- File: movieSearch.py -
  - Requests: Requests is a Python HTTP library, released under the Apache License 2.0. The goal of the project is to make HTTP requests simpler and more human-friendly.
  - Beautiful Soup: Beautiful Soup is a Python package for parsing HTML and XML documents. It creates a parse tree for parsed pages that can be used to extract data from HTML, which is useful for web scraping.

- Keywords:
  - Web Scrapping - Web Scraping (also termed Screen Scraping, Web Data Extraction, Web Harvesting etc.) is a technique employed to extract large amounts of data from websites whereby the data is extracted and saved to a local file in your computer or to a database in table (spreadsheet) format.
  - Transfer Learning - Transfer learning refers to the use of a model that has been trained to solve one problem as the basis to solve some other somewhat similar problem. One common way to do this is by fine-tuning the original model. Because the fine-tuned model doesn't have to learn from scratch, it can generally reach higher accuracy with much less data and computation time than models that don't use transfer learning.
  - Gradual Unfreezing - Gradual unfreezing is that all layers except last layer are frozen in the first epoch and only the last layer is fine-tuned. In the next epoch, last frozen layer will be unfrozen

and fine-tuning all unfrozen layer. More and more layer will be unfrozen in coming epoch.

– Data Augmentation - Data augmentation is a strategy that enables practitioners to significantly increase the diversity of data available for training models, without actually collecting new data. Data augmentation techniques such as cropping, padding, and horizontal flipping are commonly used to train large neural networks.

5 Quality Assurance Documentation

5.1 NLP Model

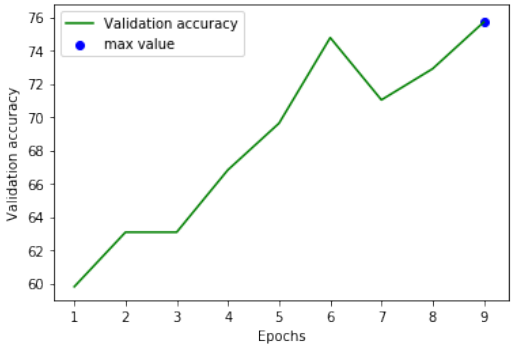


Figure 6: Validation accuracy plot

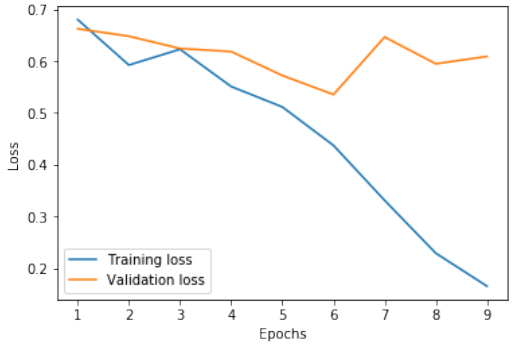


Figure 7: Loss plots

The model is trained on total 3000 reviews - 1500 actual reviews and 1500 similar reviews generated using iNLTK (data augmentation). There are total 1658 positive sentiment reviews and 1342 negative sentiment reviews in the training set. The validation set contains 214 reviews, with 114 positive sentiment reviews and 100 negative sentiment reviews. The validation accuracy obtained on it is **75.70%**. The accuracy and loss plots are shown in Fig. 6 and Fig. 7 respectively, We also tested our model on Hindi Movie Polarity Labeled Corpora by Centre For Indian Language Technology, IIT Bombay and obtained an accuracy of **57.95%**.

5.2 Testing Checklist

Table 3: Testing Checklist Table

Product Testing		
Working Status	User Action	Expected Result
✓	'Home' page	Shows home screen
✓	'About Project' page	Shows description of the project
✓	'Prediction' page	Shows Prediction screen
✓	'Search Movie Review' page	Shows Movie Review Search screen
✓	'Meet The Team' page	Shows Team Members screen which shows details of team members
✓	Clicking 'Submit' in Prediction screen	Predicts the polarity of the movie review
✓	Clicking 'Submit' in Movie Review Search screen	Displays name, rating and hindi review of the movie along with its source and writer

6 User Documentation

For our end-users, we are providing a Hindi Text Corpus, an NLP sentiment analysis model and a website to deploy above.

6.1 User guide for dataset

We have provided datasets of hindi film reviews. The datasets and all of its versions have been described in detail in source code documentation subsection. Here we are presenting the exploratory analysis of our dataset for the end users.

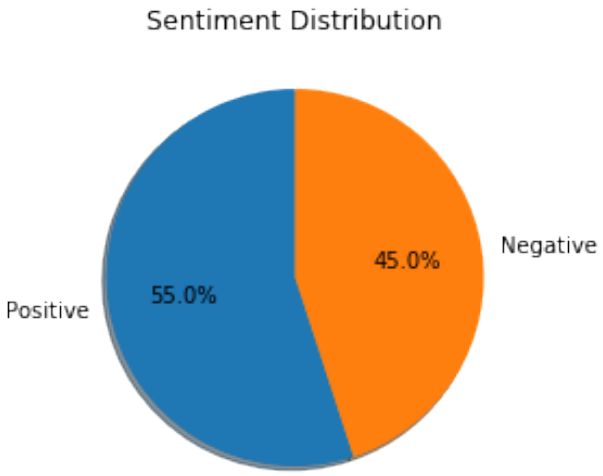
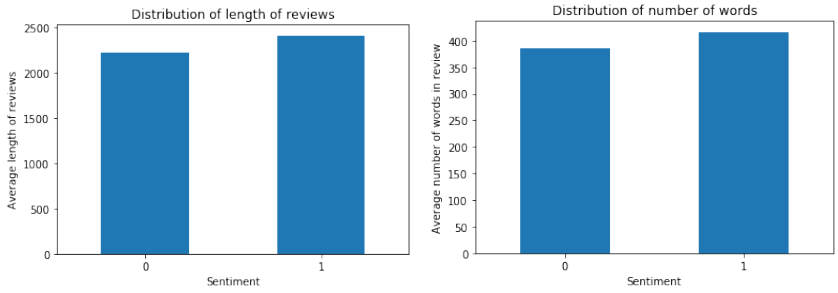


Figure 8: Pie chart showing distribution of reviews in two sentiments - positive and negative. This chart shows that our dataset is well balance



(a) Distribution of length of Reviews

(b) Distribution of number of words

Figure 9: Plots showing that length and number of words in a review are higher for positive sentiment.

Figure 10: Most common words in the form of WordCloud.

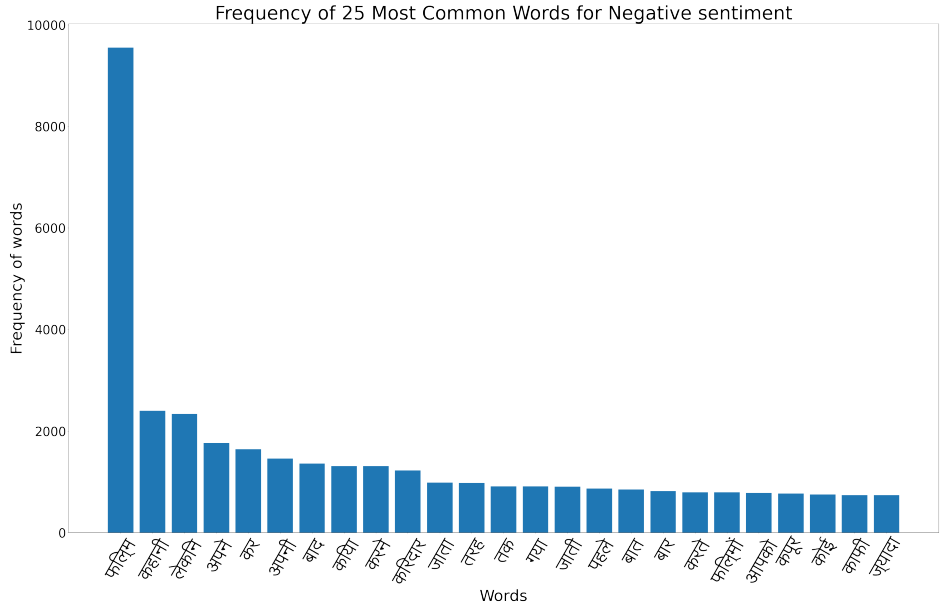


Figure 11: Most common words for negative sentiment.

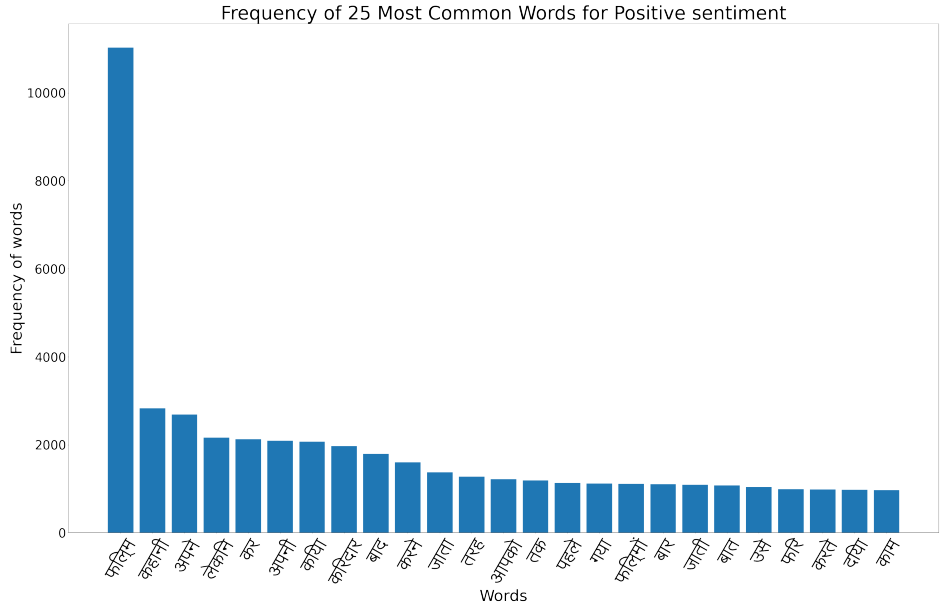


Figure 12: Most common words for positive sentiment.

6.2 User Guide for website

Here is the User Guide to effectively use our website

6.2.1 Predicting Sentiment of your movie review

Figure 13: Home page

Click on the 'Prediction' option in the menu bar at the top of the home page. This will lead you to the Prediction Page of the Website.

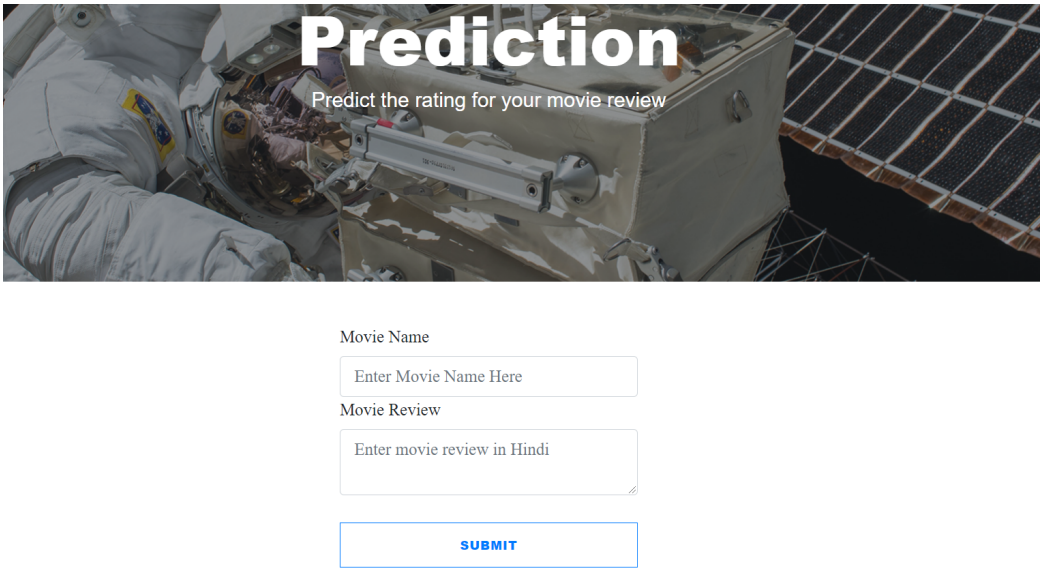


Figure 14: Prediction Page

Here in the 'Movie Review' Bar, type your movie review and click "Submit" button beneath it. In just a moment, it will take you to the result page with the sentiment ('Positive' or 'Negative') with associated probability of your submitted movie review.

6.2.2 Searching rating and review for a given movie

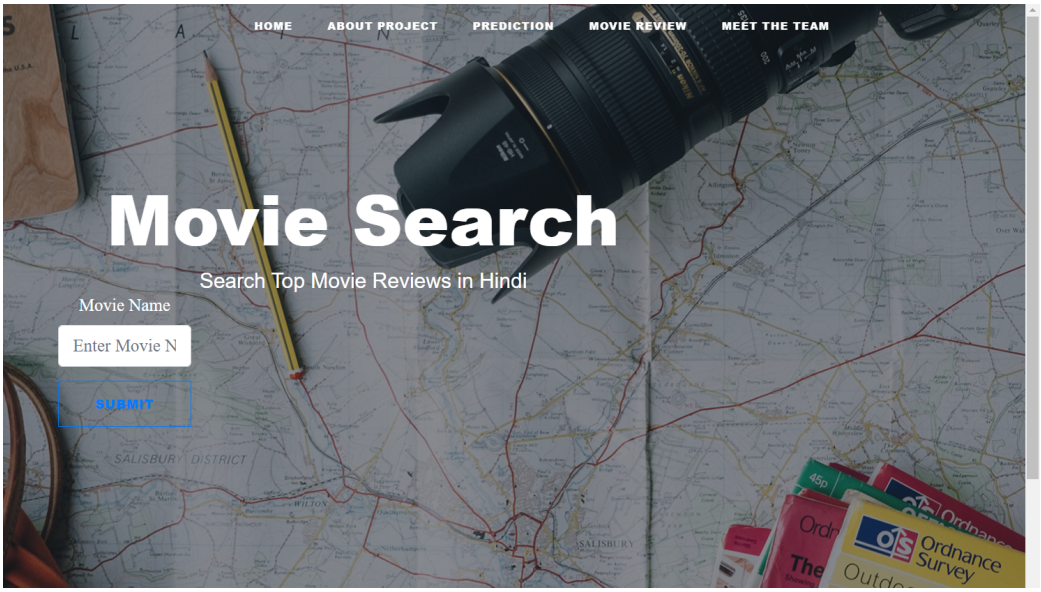


Figure 15: Movie Review Page

Click on the 'Movie Review' option in the menu bar at the top of the home page. This will lead you to the Movie Review page of the Website.

Here in the 'Movie Name' Bar type your movie name and click 'Submit' button beneath it. In Just a Moment, it will take you to the search result page with the movie review and ratings along with the critic's name and its source.

6.2.3 Description of The Project

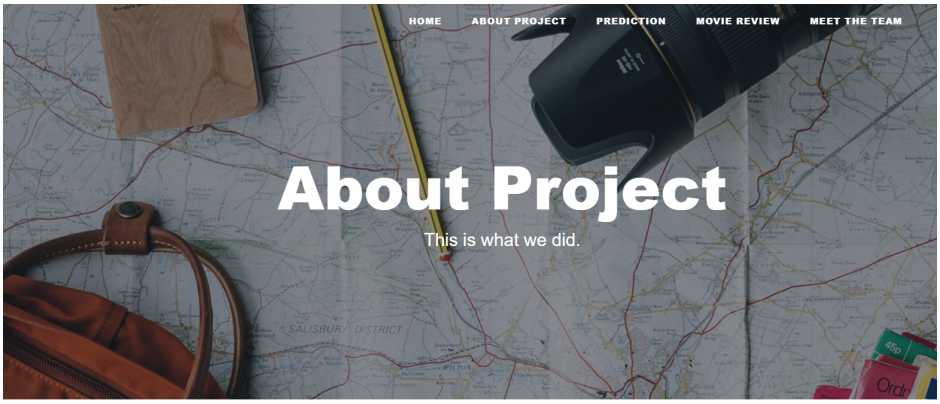


Figure 16: Description Page

Click on the 'About Project' option in the menu bar at the top of the home page. This will lead you to the Description page of the Website.

Here, you will find the information about the functionality of the website. Along with that, the sources of Datasets (sample text required for the Text-Predictor Model) is written with their links.

After scrolling a bit, brief description of the technology used and the Nlp model used in Sentimental Analysis of the Hindi Text is mentioned. For Further reference, the Link of the FAQs is given on the Home Page.

6.2.4 Frequently Asked Questions(FAQs)

FAQ's On The Home Page :-

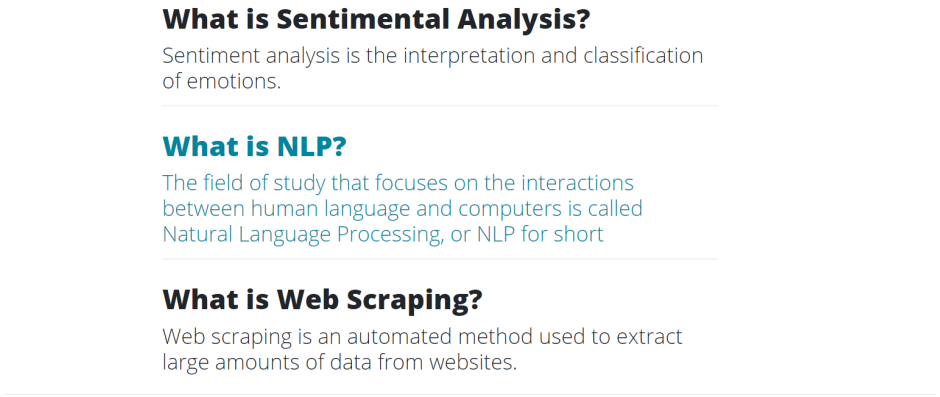


Figure 17: Home Page

After the tour of the website, if you come up with any doubt then just visit the Home Page where basic questions related to 'web scraping', 'NLP' and 'sentimental analysis' are mentioned. Moreover, you can visit the webpage related to the queries you clicked for better understanding of the query.