



Data Science Intern at Data Glacier

Week 5: Cloud and API Deployment

Name: Yash Jayesh Doshi

Batch Code: LISUM19

Date: 01st April 2022

Submitted to: Data Glacier

Table of Contents:

Heading	Page number
Introduction	3
Data Information	3
Building a model in Jupyter notebook	4
Building a Flask Web Application	9
Checking the Flask Web Application	16
API Testing	18
Deployment on Cloud	18

1. Introduction

In this project, I try to predict the sentiment of a movie review with the help of NLP and ML. The overall workflow of the project is shown in the figure below.

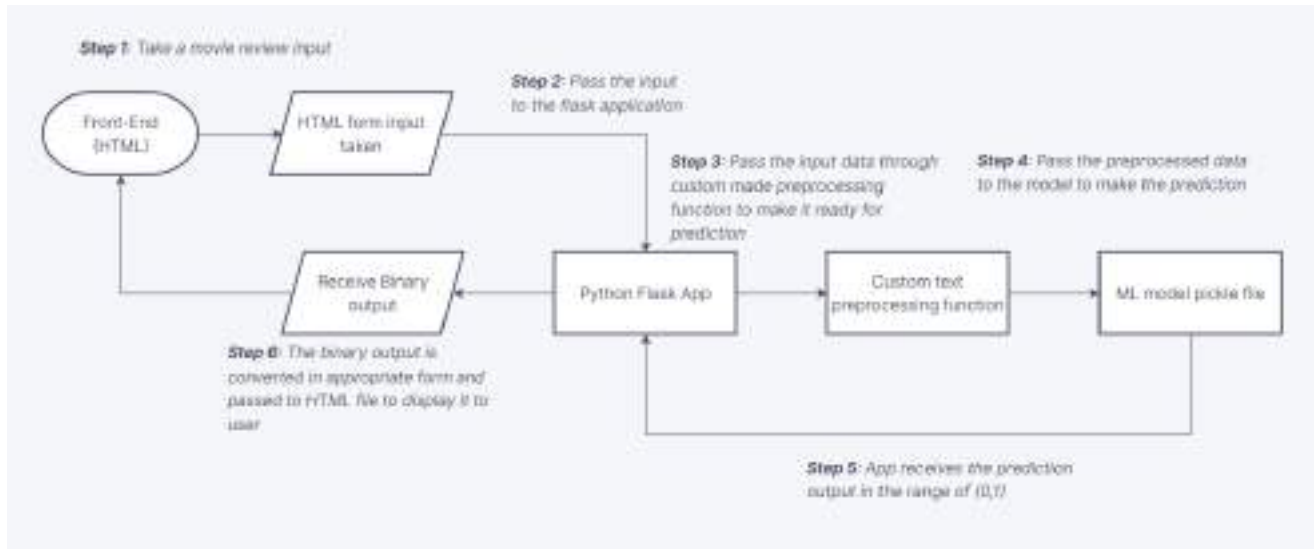


Figure 1.1: Application Workflow

2. Data Information

The data used for this project is IMDb movie reviews dataset ([Link](#))

Source credit:

```
@InProceedings{maas-EtAl:2011:ACL-HLT2011,
  author    = {Maas, Andrew L. and Daly, Raymond E. and Pham, Peter T. and Huang,
  Dan and Ng, Andrew Y. and Potts, Christopher},
  title     = {Learning Word Vectors for Sentiment Analysis},
  booktitle = {Proceedings of the 49th Annual Meeting of the Association for
  Computational Linguistics: Human Language Technologies},
  month     = {June},
  year      = {2011},
  address   = {Portland, Oregon, USA},
  publisher = {Association for Computational Linguistics},
  pages     = {142--150},
  url       = {http://www.aclweb.org/anthology/P11-1015}
}
```

This dataset contains movie reviews along with their associated binary sentiment polarity labels. It is intended to serve as a benchmark for sentiment classification.

The core dataset contains 50,000 reviews split evenly into 25,000 train and 25,000 test sets. The overall distribution of labels is balanced (25,000 pos and 25,000 neg). In the entire collection, no more than 30 reviews are allowed for any given movie because reviews for the same movie tend to have correlated ratings. Further, the train and test sets contain a disjoint set of movies, so no significant performance is obtained by memorizing movie-unique terms and their associated with observed labels. In the labeled

train/test sets, a negative review has a score ≤ 4 out of 10, and a positive review has a score ≥ 7 out of 10. Thus, reviews with more neutral ratings are not included in the train/test sets.

There are two top-level directories [train/, test/] corresponding to the training and test sets. Each contains [pos/, neg/] directories for the reviews with binary labels positive and negative. Within these directories, reviews are stored in text files named following the convention [[id]_[rating].txt] where [id] is a unique id and [rating] is the star rating for that review on a 1-10 scale. For example, the file [test/pos/200_8.txt] is the text for a positive-labeled test set example with unique id 200 and star rating 8/10 from IMDb.

3. Building a Model in Jupyter Notebook

3.1 Importing the Required Libraries and Dataset

```
import numpy as np
import pandas as pd
import warnings
import os
import pickle

from sklearn.model_selection import train_test_split
import re
import nltk
from nltk.corpus import stopwords
from nltk.stem import WordNetLemmatizer
from sklearn.feature_extraction.text import CountVectorizer, TfidfTransformer
from sklearn.metrics import accuracy_score, precision_score, recall_score, f1_score
from sklearn.pipeline import make_pipeline
from sklearn.naive_bayes import MultinomialNB
from sklearn.linear_model import LogisticRegression
from sklearn.svm import SVC
from sklearn.ensemble import RandomForestClassifier
from xgboost import XGBClassifier
from sklearn.metrics import classification_report

warnings.filterwarnings("ignore")
nltk.download('stopwords')
nltk.download('wordnet')
```

Figure 3.1: Importing the required libraries

Loading the data

```
# Load the reviews and labels
# 1 for positive reviews and 0 for negative reviews
reviews = []
labels = []
path="aclImdb/train"
for label_type in ["pos", "neg"]:
    dir_path = os.path.join(path, label_type)
    for filename in os.listdir(dir_path):
        if filename.endswith(".txt"):
            with open(os.path.join(dir_path, filename), "r", encoding="utf-8") as f:
                review = f.read()
                reviews.append(review)
                labels.append(1 if label_type == "pos" else 0)
```

Figure 3.2: Loading the data

3.2 Preprocessing the data

```
# Define a function to perform text preprocessing
def preprocess_text(text):
    # Remove HTML tags
    cleaned_text = re.sub('<[^>]*>', '', text)

    # Remove punctuation
    cleaned_text = re.sub(r'[\W\s]', '', cleaned_text)

    # Convert to lowercase
    cleaned_text = cleaned_text.lower()

    # Lemmatize and remove stopwords
    lemmatizer=WordNetLemmatizer()
    stop_words = set(stopwords.words('english'))
    cleaned_text=' '.join(lemmatizer.lemmatize(word) for word in cleaned_text.split()
                           if word not in stop_words)

    return cleaned_text

# Preprocess the movie reviews
cleaned_reviews = []
for review in reviews:
    cleaned_review = preprocess_text(review)
    cleaned_reviews.append(cleaned_review)
```

Figure 3.3: Preprocessing the reviews

3.3 Splitting the dataset into train, validation and test set

```
# Split the dataset into training, validation, and testing sets
train_reviews, test_reviews, train_labels, test_labels = train_test_split(cleaned_reviews, labels, test_size=0.15,
                                                                           random_state=42, shuffle=True, stratify=labels)
train_reviews, val_reviews, train_labels, val_labels = train_test_split(train_reviews, train_labels, test_size=0.15,
                                                                           random_state=42, shuffle=True, stratify=train_labels)
```

Figure 3.4: Splitting the dataset

3.4 Vectorizing the data and finding the best ML model

```
# Making a pipeline for vectorizing and transforming the reviews
pipeline = make_pipeline(CountVectorizer(), TfidfTransformer())
```

```
X_train = pipeline.fit_transform(train_reviews)
X_val = pipeline.transform(val_reviews)
X_test = pipeline.transform(test_reviews)
```

```
# Defining the models that will be tested for our analysis
models = {
    "MultinomialNB": MultinomialNB(),
    "Logistic Regression": LogisticRegression(),
    "SVM": SVC(kernel="linear"),
    "Random Forest": RandomForestClassifier(),
    "XGBoost Classifier": XGBClassifier()
}
```

Figure 3.5: Transforming the data and defining models dictionary

```

from sklearn.model_selection import GridSearchCV
import time

model_list=[]

# Define the parameter grid for each model
nb_param_grid = {'alpha': [0.1, 0.01, 0.001, 0.0001]}

lr_param_grid = {'C': [0.1, 1, 10, 100],
                  'penalty': ['l1', 'l2']}

svm_param_grid = {'C': [0.1, 1, 10, 100]}

rf_param_grid = {'n_estimators': [50, 100, 200, 300],
                  'max_depth': [None, 5, 10, 20],
                  'min_samples_split': [2, 5, 10]}

xgb_param_grid = {'n_estimators': [50, 100, 200, 300],
                  'max_depth': [3, 5, 10],
                  'learning_rate': [0.1, 0.01, 0.001]}

# Define the parameter grids for all models in a dictionary
param_grids = {'MultinomialNB': nb_param_grid,
                'Logistic Regression': lr_param_grid,
                'SVM': svm_param_grid,
                'Random Forest': rf_param_grid,
                'XGBoost Classifier': xgb_param_grid}

```

Figure 3.6: Importing required libraries and defining hyper-parameters dictionary

```
# Testing all the models and appending the statistics to model_list
for model_name,param in param_grids.items():

    start = time.time()
    algo = models[model_name]
    grid_search = GridSearchCV(estimator=algo,param_grid=param,cv=5,n_jobs=-1,scoring='accuracy')
    grid_search.fit(X_train,train_labels)
    best_para = grid_search.best_params_
    algo.set_params(**best_para)
    algo.fit(X_train,train_labels)
    y_pred = algo.predict(X_val)
    end = time.time()

    print(f"{model_name}")
    print(f"Time taken: {end-start}") #Time is in seconds
    model_list.append([model_name,accuracy_score(val_labels,y_pred),precision_score(val_labels,y_pred),
                    recall_score(val_labels,y_pred),f1_score(val_labels,y_pred),best_para])
    print("-----")
```

```
MultinomialNB
Time taken: 6.968690395355225
-----
```

```
Logistic Regression
Time taken: 20.587644577026367
-----
```

```
SVM
Time taken: 1976.6532769203186
-----
```

```
Random Forest
Time taken: 4465.19069314003
-----
```

```
XGBoost Classifier
Time taken: 5181.1491086483
-----
```

Figure 3.7: Finding the best model and best hyper-parameters

```
models_df = pd.DataFrame(data=model_list,columns=['Model name','Accuracy','Precision','Recall','f1-score','Best Parameters'])
models_df
```

	Model name	Accuracy	Precision	Recall	f1-score	Best Parameters
0	MultinomialNB	0.860100	0.865140	0.853199	0.859128	{alpha: 0.1}
1	Logistic Regression	0.896801	0.888275	0.907779	0.897921	{C: 10, 'penalty': 'l2'}
2	SVM	0.890996	0.890663	0.909661	0.900062	{C: 1}
3	Random Forest	0.861660	0.868842	0.851045	0.860310	{max_depth: None, 'min_samples_split': 5, 'n...
4	XGBoost Classifier	0.863237	0.856527	0.872647	0.864512	{learning_rate: 0.1, 'max_depth': 10, 'n_est...

Figure 3.8: Viewing the results in a dataframe

3.5 Choosing the best model, training it and saving the model and pipeline

Choosing Logistic Regression for our analysis

```
model = LogisticRegression(C=10,penalty='l2')
model.fit(X_train,train_labels)
prediction = model.predict(X_test)
```

```
print(classification_report(prediction,test_labels))
```

	precision	recall	f1-score	support
0	0.86	0.89	0.88	1825
1	0.89	0.87	0.88	1925
accuracy			0.88	3750
macro avg	0.88	0.88	0.88	3750
weighted avg	0.88	0.88	0.88	3750

```
with open('ml_model.pickle', 'wb') as f:
    pickle.dump(model, f)
```

```
with open('preprocessor.pickle', 'wb') as f:
    pickle.dump(pipeline, f)
```

Figure 3.9: Fitting the data to the best model and saving the model

4. Building the Flask Web Application

Now that we have trained and saved the model in a pickle file, we will use it to predict the sentiment of a movie review entered by a user in our Web Application.

4.1 Building utils.py

We will define all the preprocessing operations in a function inside the file utils.py, which we can then use to preprocess the movie review that we get from the user of our application.

```

utils.py > text_preprocessor
1  import numpy as np
2  import re
3  import pickle
4  from nltk.corpus import stopwords
5  from nltk.stem import WordNetLemmatizer
6
7  file_path = 'preprocessor.pickle'
8  with open(file_path, 'rb') as f:
9      preprocessor_pipeline = pickle.load(f)
10
11  def text_preprocessor(text):
12      # Remove HTML tags
13      preprocessed_text = re.sub('<[^>]*>', '', text)
14
15      # Remove punctuation
16      preprocessed_text = re.sub(r'[^\\w\\s]', '', preprocessed_text)
17
18      # Convert to lowercase
19      preprocessed_text = preprocessed_text.lower()
20
21      # Lemmatize and remove stopwords
22      lemmatizer = WordNetLemmatizer()
23      stop_words = set(stopwords.words('english'))
24      preprocessed_text = ' '.join(lemmatizer.lemmatize(word) for word in preprocessed_text.split()
25                                  if word not in stop_words)
26
27      preprocessed_text = preprocessor_pipeline.transform([preprocessed_text])
28
29      return preprocessed_text
30

```

Figure 4.1: *utils.py*

4.2 Building app.py

Now, we will build our Flask application in `app.py`. We will use the default page to take the input from the user, and display the input and the prediction in a separate web page. We will locally store all the inputs the user provides so user can view the predictions of multiple movie reviews.

Additionally, we will define a separate URL to test the app on an API software.

The entire code for `app.py` can be found below.

```

app.py > ...
1  from flask import Flask, render_template, request, Response, url_for, jsonify
2  import tensorflow as tf
3  import pickle
4
5  from utils import text_preprocessor
6
7  file_path = 'ml_model.pickle'
8  with open(file_path, 'rb') as f:
9      model = pickle.load(f)
10
11  app = Flask(__name__)
12
13  # Defining route for home page
14  @app.route('/')
15  def index():
16      return render_template("index.html")
17
18  # Defining route for api
19  @app.route('/predict_api', methods=['GET', 'POST'])
20  def predict_api():
21      if request.method == 'POST':
22
23          data = request.json['data']
24          preprocessed_data = text_preprocessor(data)
25          output = model.predict(preprocessed_data)[0][0]
26
27          return jsonify(output)
28

```

Figure 4.2: app.py (part-1)

```

@app.route('/predict', methods=['GET', 'POST'])
def predict():
    if request.method == 'POST':
        review=request.form['review']
        preprocessed_data = text_preprocessor(review)
        output = model.predict(preprocessed_data)
        if output>0.5:
            sentiment="Positive"
        else:
            sentiment="Negative"
        predictions.append({'review': review, 'sentiment': sentiment})
        return render_template('predictions.html', predictions=predictions)
    return render_template("index.html")

@app.route('/predictions')
def predictions():
    # Render the predictions page with the list of predictions
    return render_template('predictions.html', predictions=predictions)

if __name__ == "__main__":
    predictions=[]
    app.run()

```

Figure 4.3: app.py (part-2)

4.3 Building the HTML and CSS files for the application

Our model includes two HTML files, namely, index.html and predictions.html, and one CSS file to style the two HTML files, namely, style.css

The code for both the HTML files can be found below.

```
templates > index.html > html
1  <!DOCTYPE html>
2
3  <html>
4    <head>
5      <link rel="stylesheet" type="text/css" href="{{ url_for('static', filename='style.css') }}">
6      <title>Movie Review Sentiment Analysis</title>
7    </head>
8    <body>
9      <h1>Movie Review Sentiment Analysis</h1>
10     <form action="/predict" method="post">
11       <label for="review">Enter your movie review:</label>
12       <br>
13       <textarea id="review" name="review" rows="10" cols="50"></textarea>
14       <br>
15       <button type="submit">Submit</button>
16     </form>
17   </body>
18 </html>
```

Figure 4.4: index.html

```
templates > predictions.html > html
1  <!DOCTYPE html>
2  <html>
3    <head>
4      <link rel="stylesheet" type="text/css" href="{{ url_for('static', filename='style.css') }}">
5      <title>Movie Review Sentiment Analysis Predictions</title>
6    </head>
7    <body>
8      <h1>Movie Review Sentiment Analysis Predictions</h1>
9      <table>
10        <thead>
11          <tr>
12            <th>Review</th>
13            <th>Sentiment</th>
14          </tr>
15        </thead>
16        <tbody>
17          {% for prediction in predictions %}
18            <tr>
19              <td class="review">{{ prediction.review }}</td>
20              <td class="{{ 'positive' if prediction.sentiment=='positive' else 'negative' }}">{{ prediction.sentiment }}</td>
21            </tr>
22          {% endfor %}
23        </tbody>
24      </table>
25      <a class="new_predictions" href="/">Make a new prediction</a>
26    </body>
27  </html>
```

Figure 4.5: predictions.html

The code for the CSS file can be found below.

```
1  /* Styles for the index page */
2  body {
3      background-color: #f9f9f9;
4      font-family: 'Helvetica Neue', sans-serif;
5      color: #333;
6  }
7
8  h1 {
9      text-align: center;
10     font-size: 3rem;
11     margin-top: 3rem;
12 }
13
14 form {
15     margin: 2rem auto;
16     max-width: 800px;
17     padding: 2rem;
18     background-color: #fff;
19     border-radius: 10px;
20     box-shadow: 0 0 20px rgba(0, 0, 0, 0.2);
21 }
22
23 label {
24     display: block;
25     font-size: 1.2rem;
26     font-weight: bold;
27     margin-bottom: 1rem;
28 }
29
```

Figure 4.6: style.css (part-1)

```

30  textarea {
31      display: block;
32      width: 100%;
33      font-size: 1.2rem;
34      padding: 0.5rem;
35      border-radius: 5px;
36      border: 1px solid #ccc;
37      resize: none;
38      height: 200px;
39  }
40
41  button {
42      display: block;
43      margin: 1rem auto;
44      padding: 0.5rem 1rem;
45      font-size: 1.2rem;
46      background-color: #007bff;
47      color: #fff;
48      border: none;
49      border-radius: 5px;
50      cursor: pointer;
51  }
52
53  /* Styles for the predictions page */
54  table {
55      margin: 2rem auto;
56      max-width: 1100px;
57      border-collapse: collapse;
58  }
59
60  thead {
61      background-color: #007bff;
62      color: #fff;
63  }
64

```

Figure 4.7: style.css (part-2)


```

65  th, td {
66      padding: 0.5rem;
67      text-align: center;
68      border: 1px solid #ccc;
69  }
70
71  th {
72      font-weight: bold;
73  }
74
75  td.review {
76      padding-top: 20px;
77      padding-bottom: 20px;
78      padding-right: 10px;
79      text-align: left;
80  }
81
82  tbody tr:nth-child(even) {
83      background-color: #f2f2f2;
84  }
85
86  a.New_predictions {
87      display: flex;
88      justify-content: center;
89      margin: 1rem auto;
90      padding: 0.5rem 1rem;
91      font-size: 1.2rem;
92      background-color: #007bff;
93      color: #fff;
94      border: none;
95      border-radius: 5px;
96      text-align: center;
97      text-decoration: none;
98      cursor: pointer;
99  }
100

```

Figure 4.8: style.css (part-3)

```

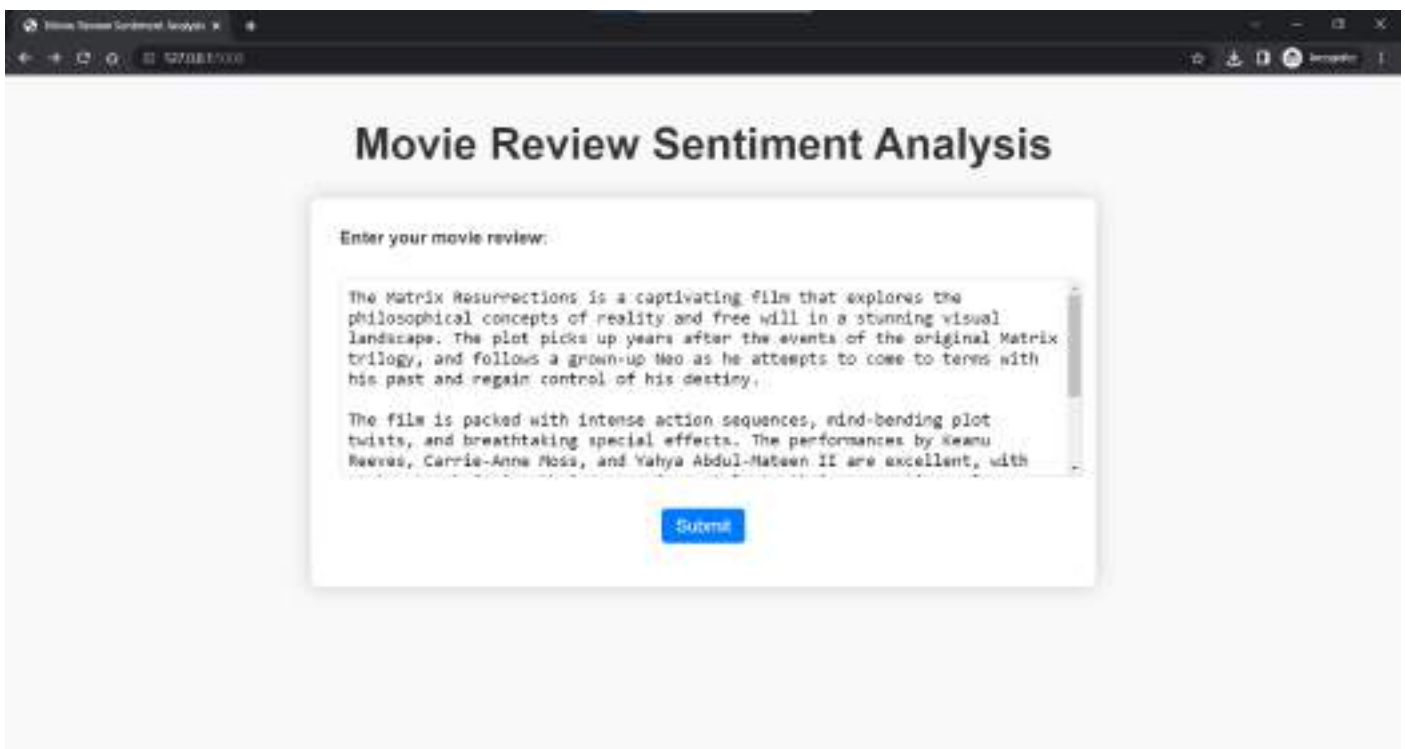
101 a.New_Predictions: hover {
102 | background-color: #0056b3;
103 | }
104
105 .positive {
106 | color: green;
107 | font-weight: bold;
108 | }
109
110 .negative {
111 | color: red;
112 | font-weight: bold;
113 | }
114

```

Figure 4.9: style.css (part-4)

5. Checking the Flask Web Application

The images below depict how the Web Application works.



The screenshot shows a web browser window with the title 'Movie Review Sentiment Analysis'. The main heading on the page is 'Movie Review Sentiment Analysis'. Below the heading is a form with the label 'Enter your movie review:'. The form contains two paragraphs of text: 'The Matrix Resurrections is a captivating film that explores the philosophical concepts of reality and free will in a stunning visual landscape. The plot picks up years after the events of the original Matrix trilogy, and follows a grown-up Neo as he attempts to come to terms with his past and regain control of his destiny.' and 'The film is packed with intense action sequences, mind-bending plot twists, and breathtaking special effects. The performances by Keanu Reeves, Carrie-Anne Moss, and Yahya Abdul-Mateen II are excellent, with'. At the bottom of the form is a blue button labeled 'Submit'.

Figure 5.1: Entering movie review and submitting it

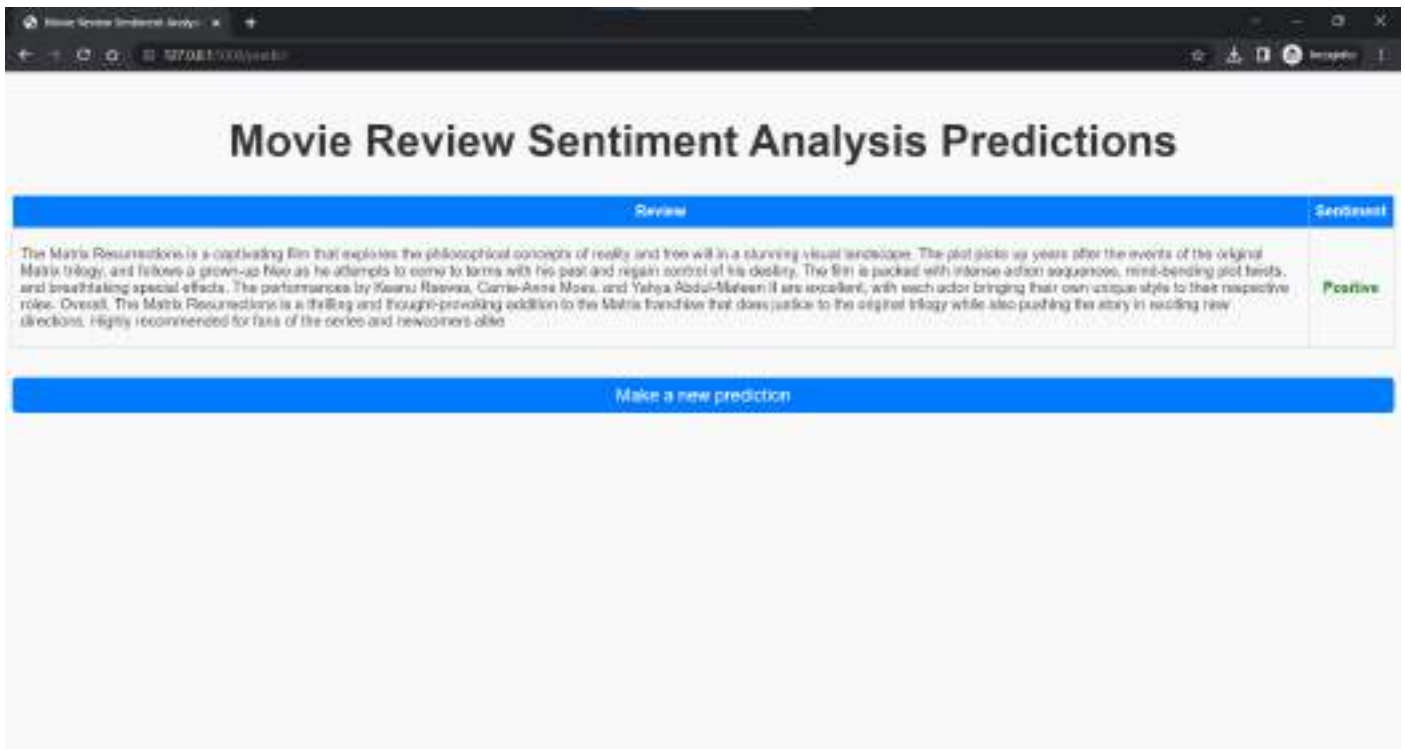


Figure 5.2: Displaying the input and output and clicking 'Make a New Prediction'

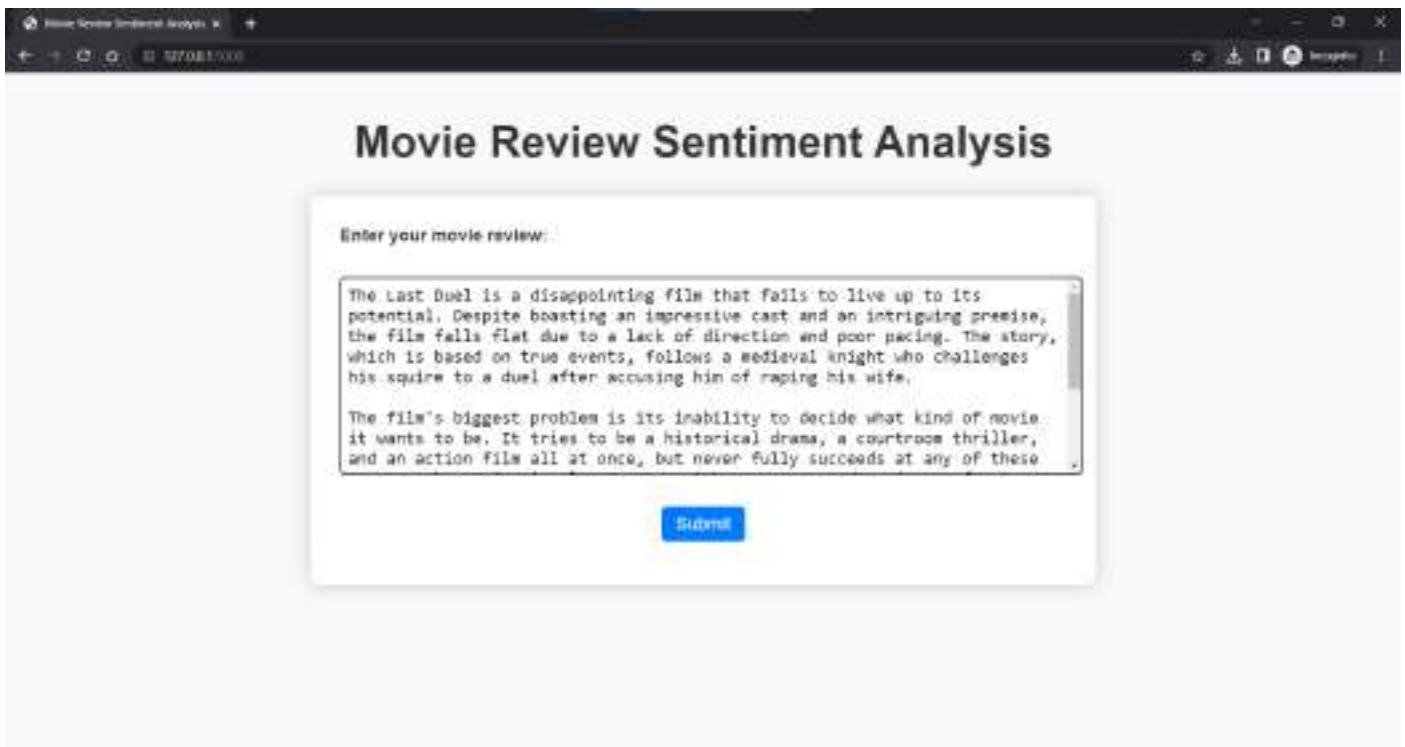


Figure 5.3: Entering second movie review

Review	Sentiment
The Matrix Resurrections is a captivating film that explores the philosophical concepts of reality and free will in a stunning visual landscape. The plot picks up years after the events of the original Matrix trilogy, and follows a grown-up Neo as he attempts to come to terms with his past and regain control of his destiny. The film is packed with intense action sequences, mind-bending plot twists, and breathtaking special effects. The performances by Keanu Reeves, Carrie-Anne Moss, and Yahya Abdul-Mateen II are excellent, with each actor bringing their own unique style to their respective roles. Overall, The Matrix Resurrections is a thrilling and thought-provoking addition to the Matrix franchise that does justice to the original trilogy while also pushing the story in exciting new directions. Highly recommended for fans of the series and newcomers alike.	Positive
The Last Duel is a disappointing film that fails to live up to its potential. Despite boasting an impressive cast and an intriguing premise, the film falls flat due to a lack of direction and poor pacing. The story, which is based on true events, follows a medieval knight who challenges his squire to a duel after accusing him of raping his wife. The film's biggest problem is its inability to decide what kind of movie it wants to be. It tries to be a historical drama, a courtroom thriller, and an action film all at once, but never fully succeeds at any of these genres. The pacing is also uneven, with some scenes dragging on for too long while others are rushed and underdeveloped. The performances are also underwhelming, with Matt Damon, Adam Driver, and Jodie Comer all delivering lackluster performances that fail to connect with the audience. Overall, The Last Duel is a forgettable film that is unlikely to leave a lasting impression on viewers.	Negative

Make a new prediction

Figure 5.4: Displaying input and corresponding output for all the reviews

6. API Testing

We can also test our flask app through an API tester. We will use POSTMAN for our purpose which will use the `"/predict_api"` app-route to make the predictions of the movie review which we will provide. The below image depicts the implementation.



Figure 6.1: API Testing on POSTMAN

7. Deployment on Cloud

Now, we will deploy our flask application to a cloud server. We will use AWS Elastic Beanstalk to deploy our Flask application.

Firstly, we need to create a ‘.ebextensions’ folder in our repository. In that we need to add a python.config file as per below. This will be required by Codepipeline to fetch our application from our GitHub repository and deploy it on Elastic Beanstalk.

```
.ebextensions > ⚙ python.config
1  option_settings:
2    aws:elasticbeanstalk:container:python:
3      WSGIPath: application:application
```

Figure 7.1: python.config code

We will first login to AWS Management Console and open Elastic Beanstalk. Then we will create a sample application with the name ‘sentiment-analysis’.

The screenshot shows the AWS Management Console interface for Elastic Beanstalk. The left sidebar contains navigation links for 'Environments', 'Applications', 'Change history', and 'Recent environments'. The main content area is titled 'Create new application' and contains two sections: 'Application information' and 'Tags'. In the 'Application information' section, the 'Application name' field is populated with 'sentiment-analysis' and is highlighted with a blue border. Below it, the 'Description' field is empty. The 'Tags' section includes a text box with instructions: 'Apply up to 50 tags. You can use tags to group and filter your resources. A tag is a key-value pair. The key must be unique within the resource and is case-sensitive. [Learn more](#)'. Below this, there is a table with two columns: 'Key' and 'Value'. The 'Key' column has an input field, and the 'Value' column has an input field. To the right of the 'Value' input field is a 'Remove tag' button. Below the table is an 'Add tag' button. At the bottom right of the form, there are 'Cancel' and 'Create' buttons. The 'Create' button is orange and highlighted.

Figure 7.2: Create application window on AWS Elastic Beanstalk

After creating the application, we will have to create a new environment for this application. We will click on ‘Create a new environment’ button and then select the Web server environment option. We then need to select Platform (which will be Python for our case) and then click Create environment.

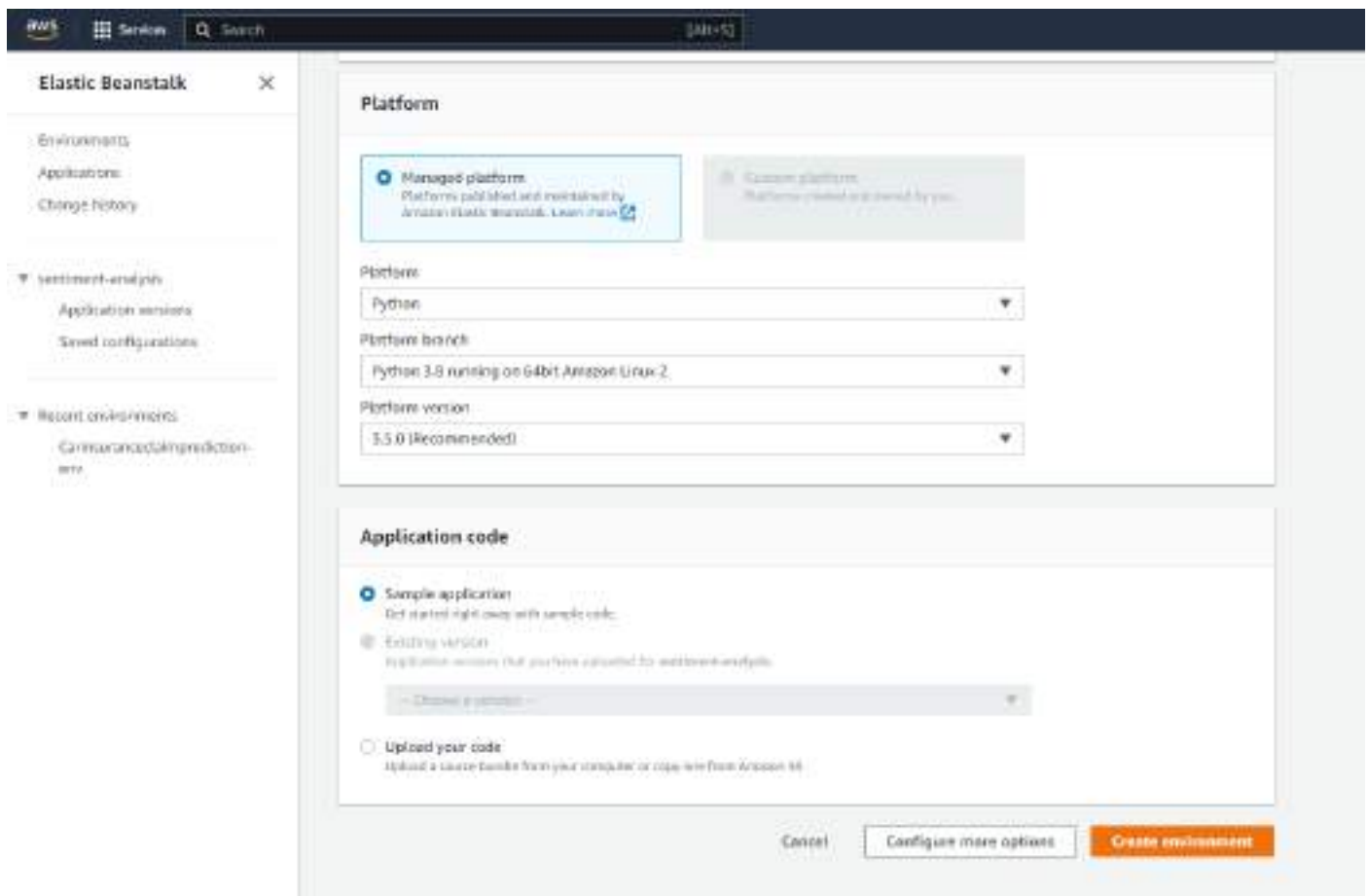


Figure 7.3: Create environment window on AWS Elastic Beanstalk

Creating environment step will take few minutes to complete. Once the environment is created, our application window will look something like this:

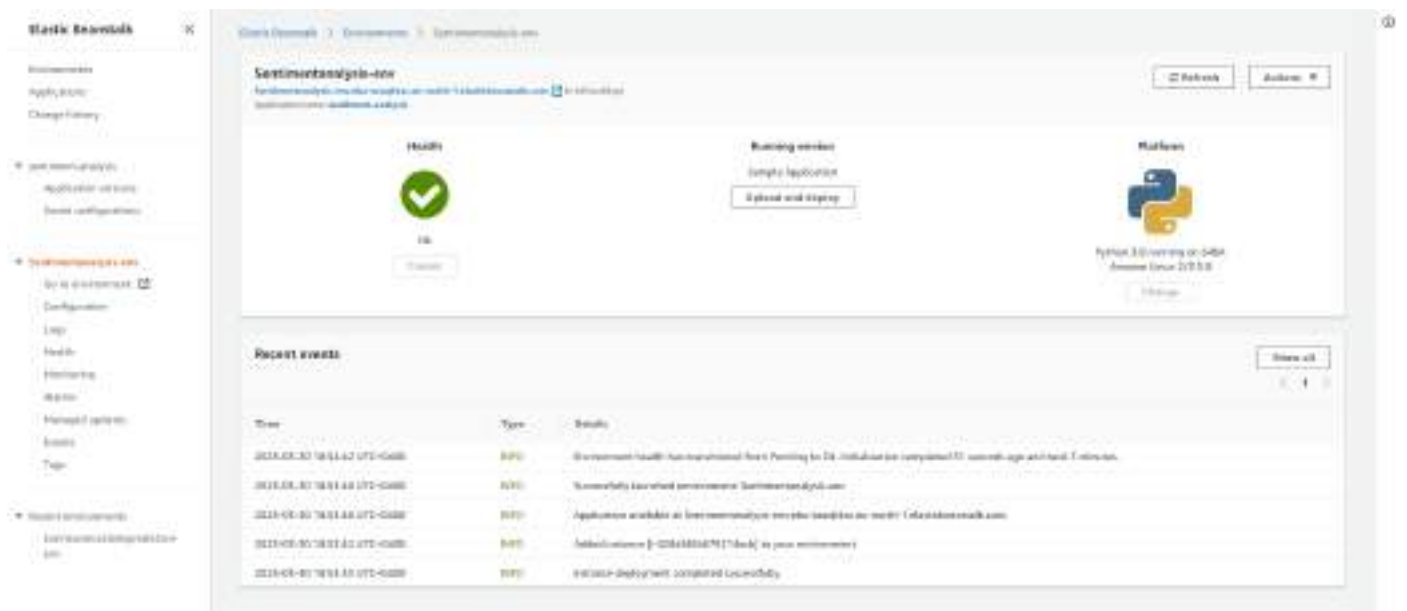


Figure 7.4: New environment window on AWS Elastic Beanstalk

Now, we need to create a pipeline that will link our GitHub repository to this sample application. We will use Codepipeline provided by AWS for this purpose.

Once we go to Codepipeline using AWS Management Console, we need to click on Create Pipeline. Then we need to give a pipeline name (which we will give 'sentiment-pipeline') and hit Next. We then need to select GitHub (version 2) as the source provider and link our repository as per the image below.

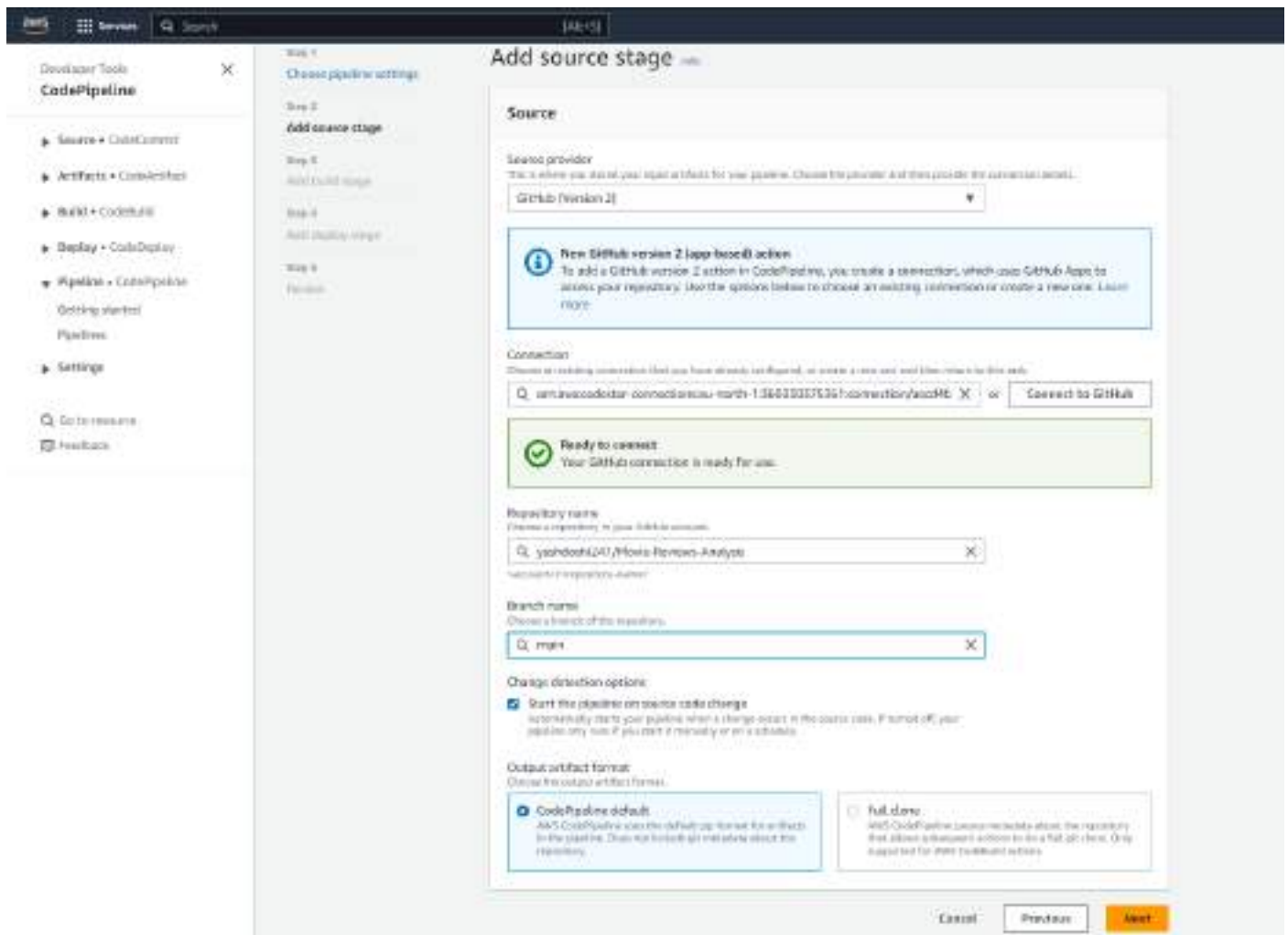


Figure 7.5: Add Source Stage in AWS Codepipeline

After clicking Next, we can skip the 'Add build stage' for our purpose. We then need to add our AWS sample application we created above to the 'Add deploy stage' as per image below.

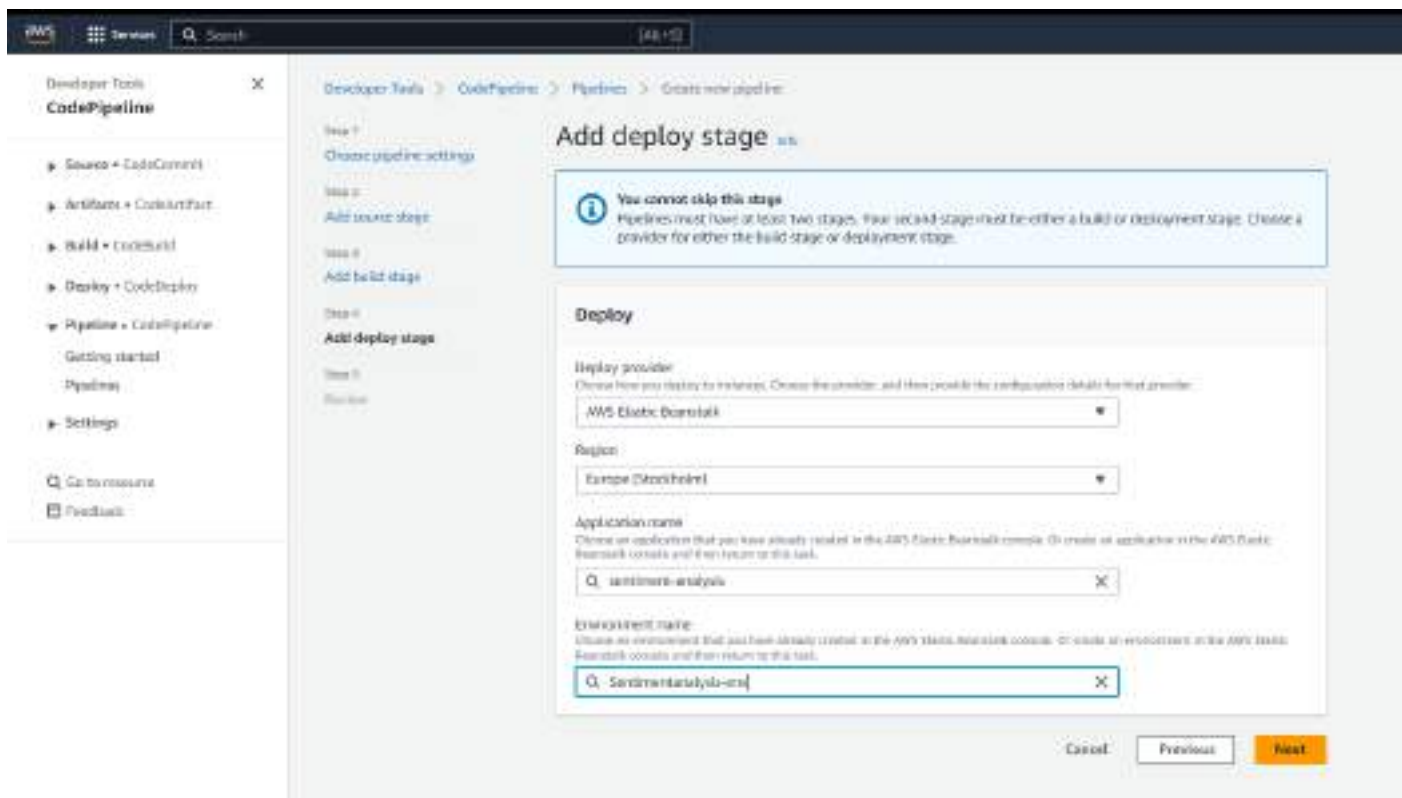


Figure 7.6: Add Deploy Stage in AWS Codepipeline

Then, we can click Next and then click 'Create pipeline'. Once the pipeline gets created, AWS will automatically fetch the latest commit from our GitHub repository and deploy it on AWS Elastic Beanstalk.

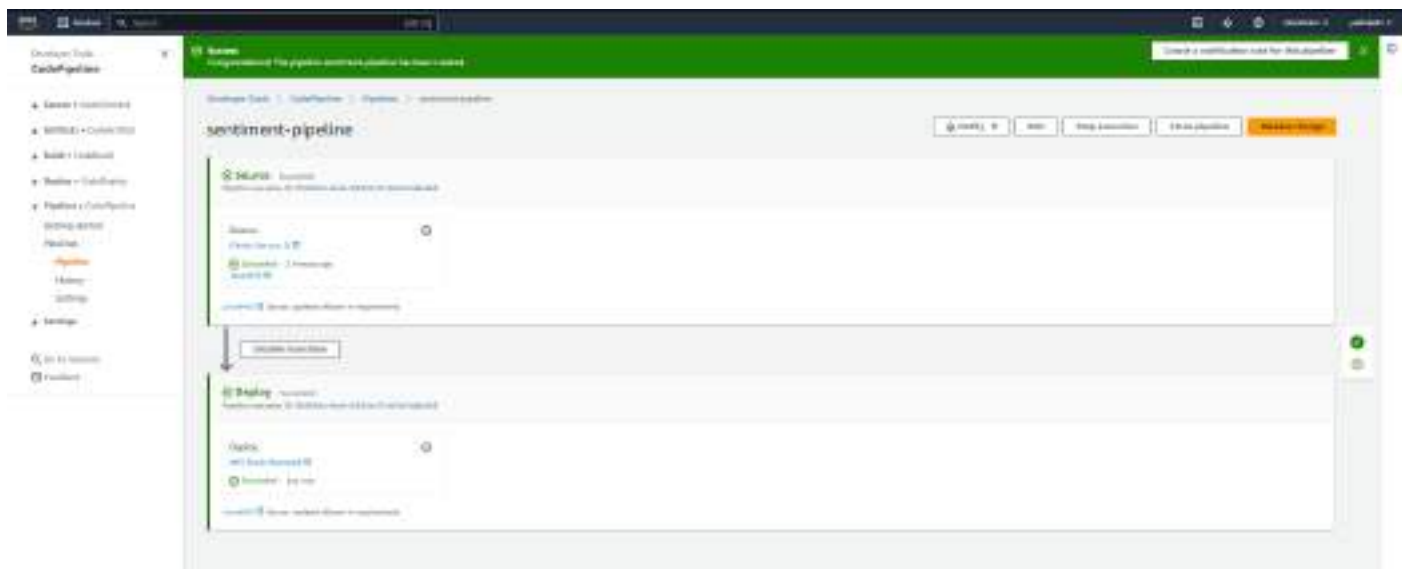


Figure 7.7: Successful creation of AWS Codepipeline

After deployment, this is how the Application pane on Elastic Beanstalk will look,



Figure 7.8: Successful deployment of Flask app on AWS Elastic Beanstalk

Health OK signifies that our application is working properly. If our code had any error, it would show 'Degraded' or 'Severe' instead.

Here's the application URL deployed on AWS Elastic Beanstalk:

<http://sentimentanalysis-env.eba-ezaqkkxc.eu-north-1.elasticbeanstalk.com/>