Architecture Document

News Article Classifier

Revision No: 1.0

Last Date of Revision: 18/11/2023.

Document Version Control-

Date Issued	Version	Description	Author
18/11/2023	1.0	LLD – V- 1.0	Mahesh. A

Contents -

Abstract

By using the pre-trained of News Article classifier user can understand the type of article he is handling.

1 Introduction

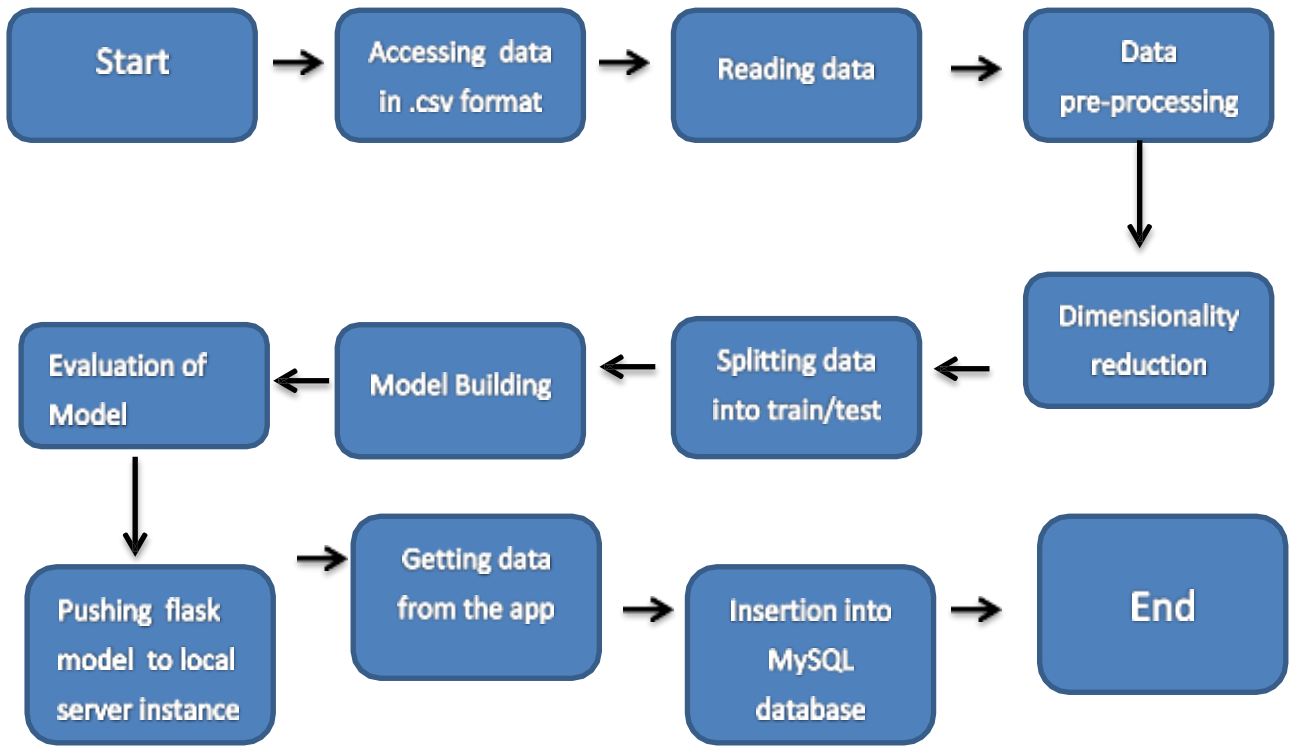
1.1. What is document required?

The goal of Architecture is to give the internal logical design of the actual programmed code for News Article classifier. Describes the class relations with predictor's .It describes the modules so that the programmer can directly code the program from the document.

1.2. Scope

Architecture document is a component-level design process that follows a step-by-stepr process. This process can be used for designing data structures, required software architecture, source code and ultimately, performance algorithms. Overall, the data organization may be defined during requirement analysis and then refined during data design work.

2 Architecture



3 Architecture Description

3.1 Data Accessing

We can access the data in the from the download link as available in the project.

We load the data to the framework using the pandas read function.

Attribute Information: (classes: ['business', 'entertainment', 'politics', 'sport', 'technology'])

1833	worldcom ex-boss launches defence lawyers defending former WorldCom chief bernie ebbers against a battery of fraud charges have called a company whistleblower as their first witness. cynthia cooper worldcom s ex-head of internal accounting alerted directors to irregular accounting practices at the us telecoms giant in 2002. her warnings led to the collapse of the firm following the discovery of an $11bn (Â£5.7bn) accounting fraud. mr ebbers has pleaded not guilty to charges of fraud and conspiracy. prosecution lawyers have argued that mr ebbers orchestrated a series of accounting tricks at worldcom ordering employees to hide expenses and inflate revenues to meet wall street earnings estimates. but ms cooper who now runs her own consulting business told a jury in new york on wednesday that external auditors arthur andersen had approved worldcom s accounting in early 2001 and 2002. she said andersen had given a green light to the procedures and practices used by worldcom. mr ebber s lawyers have said he was unaware of the fraud arguing that auditors did not alert him to any problems. ms cooper also said that during shareholder meetings mr ebbers often passed over technical questions to the company s finance chief giving only brief answers himself. the prosecution s star witness former worldcom financial chief scott sullivan has said that mr ebbers ordered accounting adjustments at the firm telling him to hit our books . however ms cooper said mr sullivan had not mentioned anything uncomfortable about worldcom s accounting during a 2001 audit committee meeting. mr ebbers could face a jail sentence of 85 years if convicted of all the charges he is facing. worldcom emerged from bankruptcy protection in 2004 and is now known as mci. last week mci agreed to a buyout by verizon communications in a deal valued at $6.75bn.	business
154	german business confidence slides german business confidence fell in february knocking hopes of a speedy recovery in europe s largest economy. munich-based research institute ifo said that its confidence index fell to 95.5 in february from 97.5 in january its first decline in three months. the study found that the outlook in both the manufacturing and retail sectors had worsened. observers had been hoping that a more confident business sector would signal that economic activity was picking up. we re surprised that the ifo index has taken such a knock said dz bank economist bernd weidensteiner. the main reason is probably that the domestic economy is still weak particularly in the retail trade. economy and labour minister wolfgang clement called the dip in february s ifo confidence figure a very mild decline . he said that despite the retreat the index remained at a relatively high level and that he expected a modest economic upswing to continue. germany s economy grew 1.6% last year after shrinking in 2003. however the economy contracted by 0.2% during the last three months of 2004 mainly due to the reluctance of consumers to spend. latest indications are that growth is still proving elusive and ifo president hans-werner sinn said any improvement in german domestic demand was sluggish. exports had kept things going during the first half of	business

	2004 but demand for exports was then hit as the value of the euro hit record levels making german products less competitive overseas. on top of that the unemployment rate has been stuck at close to 10% and manufacturing firms including daimlerchrysler siemens and volkswagen have been negotiating with unions over cost cutting measures. analysts said that the ifo figures and germany s continuing problems may delay an interest rate rise by the european central bank. eurozone interest rates are at 2% but comments from senior officials have recently focused on the threat of inflation prompting fears that interest rates may rise.	
1101	bbc poll indicates economic gloom citizens in a majority of nations surveyed in a bbc world service poll believe the world economy is worsening. most respondents also said their national economy was getting worse. but when asked about their own family s financial outlook a majority in 14 countries said they were positive about the future. almost 23 000 people in 22 countries were questioned for the poll which was mostly conducted before the asian tsunami disaster. the poll found that a majority or plurality of people in 13 countries believed the economy was going downhill compared with respondents in nine countries who believed it was improving. those surveyed in three countries were split. in percentage terms an average of 44% of respondents in each country said the world economy was getting worse compared to 34% who said it was improving. similarly 48% were pessimistic about their national economy while 41% were optimistic. and 47% saw their family s economic conditions improving as against 36% who said they were getting worse. the poll of 22 953 people was conducted by the international polling firm globescan together with the program on international policy attitudes (pipa) at the university of maryland. while the world economy has picked up from difficult times just a few years ago people seem to not have fully absorbed this development though they are personally experiencing its effects said pipa director steven kull. people around the world are saying: i m ok but the world isn t . there may be a perception that war terrorism and religious and political divisions are making the world a worse place even though that has not so far been reflected in global economic performance says the bbc s elizabeth blunt. the countries where people were most optimistic both for the world and for their own families were two fast-growing developing economies china and india followed by indonesia. china has seen two decades of blistering economic growth which has led to wealth creation on a huge scale says the bbc s louisa lim in beijing. but the results also may reflect the untrammelled confidence of people who are subject to endless government propaganda about their country s rosy economic future our correspondent says. south korea was the most pessimistic while respondents in italy and mexico were also quite gloomy. the bbc s david willey in rome says one reason for that result is the changeover from the lira to the euro in 2001 which is widely viewed as the biggest reason why their wages and salaries are worth less than they used to be. the philippines was among the most upbeat countries on prospects for respondents families but one of the most pessimistic about the world economy. pipa conducted the poll from 15 november 2004 to 3 january 2005 across 22 countries in face-to-face or telephone interviews. the interviews took place between 15 november 2004 and 5 january 2005. the margin of error is between 2.5 and 4 points depending on the country. in eight of the countries the sample was limited to major metropolitan areas.	business

1976	lifestyle governs mobile choice faster better or funkier hardware alone is not going to help phone firms sell more handsets research suggests. instead phone firms keen to get more out of their customers should not just be pushing the technology for its own sake. consumers are far more interested in how handsets fit in with their lifestyle than they are in screen size onboard memory or the chip inside shows an in-depth study by handset maker ericsson. historically in the industry there has been too much focus on using technology said dr michael bjorn senior advisor on mobile media at ericsson s consumer and enterprise lab. we have to stop saying that these technologies will change their lives he said. we should try to speak to consumers in their own language and help them see how it fits in with what they are doing he told the bbc news website. for the study ericsson interviewed 14 000 mobile phone owners on the ways they use their phone. people s habits remain the same said dr bjorn. they just move the activity into the mobile phone as it s a much more convenient way to do it. one good example of this was diary-writing among younger people he said. while diaries have always been popular a mobile phone -- especially one equipped with a camera -- helps them keep it in a different form. youngsters use of text messages also reflects their desire to chat and keep in contact with friends and again just lets them do it in a slightly changed way. dr bjorn said that although consumers do what they always did but use a phone to do it the sheer variety of what the new handset technologies make possible does gradually drive new habits and lifestyles. ericsson s research has shown that consumers divide into different tribes that use phones in different ways. dr bjorn said groups dubbed pioneers and materialists were most interested in trying new things and were behind the start of many trends in phone use. for instance he said older people are using sms much more than they did five years ago. this was because younger users often the children of ageing mobile owners encouraged older people to try it so they could keep in touch. another factor	tech

3.2 Data Pre-Processing

By the usage of the different data manipulation techniques we will remove unwanted features

Also we use the dimensionality reduction and make all the features in numerical data type

```
import re
def remove_html_tags(text):
    pattern = re.compile('<.*?>')
    return pattern.sub(r'',text)
```

```
df['Text'] = df['Text'].apply(remove_html_tags)
```

```
df['Text'][11]
```

```
'housewives lift channel 4 ratings the debut of us television hit desperate housewives has helped lift channel 4 s january audience share by
year.  other successes such as celebrity big brother and the simpsons have enabled the broadcaster to surpass bbc2 for the first month since
he channel s share of the audience fell from 11.2% to 9.6% last month in comparison with january 2004. celebrity big brother attracted less
series.  comedy drama desperate housewives managed to pull in five million viewers at one point during its run to date  attracting a quarter
dience. the two main television channels  bbc1 and itv1  have both seen their monthly audience share decline in a year on year comparison fo
s proportion remained the same at a slender 6.3%. digital multi-channel tv is continuing to be the strongest area of growth  with the bbc re
ownership of five million  inclu… '
```

```
def remove_urls(text):
    pattern =re.compile(r'http?://\S+|www\.\S+')
    return pattern.sub(r'',text)
```

```
df['Text'] = df['Text'].apply(remove_urls)
```

```
[ ] def word_corrections(text):
        text = re.sub(r"didn't", "did not", text)
        text = re.sub(r"don't", "do not", text)
        text = re.sub(r"won't", "will not", text)
        text = re.sub(r"can't", "can not", text)
        text = re.sub(r"wasn't", "do not", text)
        text = re.sub(r"should't", "should not", text)
        text = re.sub(r"could't", "could not", text)
        text = re.sub(r"\'ve", " have", text)
        text = re.sub(r"\'m", " am", text)
        text = re.sub(r"\'ll", " will", text)
        text = re.sub(r"\'re", " are", text)
        text = re.sub(r"\'s", " is", text)
        text = re.sub(r"\'d", " would", text)
        text = re.sub(r"\'t", " not", text)
        text = re.sub(r"\'m", " am", text)
        text = re.sub(r"n\'t", " not", text)
        return text
```

```
import nltk
nltk.download('punkt')
nltk.download('wordnet')
nltk.download('stopwords')
from nltk.corpus import stopwords
```

```
[nltk_data] Downloading package punkt to /root/nltk_data...
[nltk_data]   Unzipping tokenizers/punkt.zip.
[nltk_data] Downloading package wordnet to /root/nltk_data...
[nltk_data] Downloading package stopwords to /root/nltk_data...
[nltk_data]   Unzipping corpora/stopwords.zip.
```

```
[ ] stop_words = stopwords.words('english')
    stop_words.remove('no')
    stop_words.remove('not')
    stop_words.remove('nor')
```

```
[ ] def remove_stop_words(text):
      words = []
      for word in text.split():
        if word not in stop_words:
          words.append(word)
      return ' '.join(words)
```

```
    ✓ Download and installation successful
    You can now load the package via spacy.load('en_core_web_lg')
```

```
[ ] nlp = spacy.load('en_core_web_lg')
```

```
[ ] def spacy_tokenisation(text):
      words = []
      doc = nlp(text)
      for word in doc:
        words.append(str(word))
      return ' '.join(words)
```

```
[ ] df['Text'] = df['Text'].apply(spacy_tokenisation)
```

```
[ ] from nltk.stem import PorterStemmer, WordNetLemmatizer
    ps = PorterStemmer()
    lm = WordNetLemmatizer()

    def stemming_data(text):
      stem_words = []
      for word in text.split():
        if word.isnumeric():
          stem_words.append('')
        elif len(word)>2:
          stem_words.append(ps.stem(word))
      return ' '.join([l for l in stem_words])
```

```
[ ] df['Text'] = tqdm(df['Text'].apply(stemming_data))
```

```
[ ] def text_convert(data):
      data = pd.Series(data)
      data = data.apply(remove_html_tags)
      data = data.apply(remove_urls)
      data = data.str.lower()
      data = data.apply(word_corrections)
      data = data.apply(remove_punctions_betterway)
      data = data.apply(remove_stop_words)
      data = data.apply(spacy_tokenisation)
      data = data.apply(stemming_data)

      return data
```

```
[ ] df1['Text'] = df1['Text'].apply(text_convert)
```

```
[ ] df_text = pd.concat([df['Text'],df1['Text']],ignore_index=True)
```

```
[ ] df_text[0]
```

```
'worldcom exboss launch defenc lawyer defend former worldcom chief berni ebber batteri fraud charg call compani whistleblow first wit cynthia cooper worldcom exh
ead intern account alert director irregular account practic telecom giant  warn led collaps firm follow discoveri 11bn 57bn account fraud ebber plead not guilti
charg fraud conspiraci prosecut lawyer argu ebber orchestr seri account trick worldcom order employe hide expens inflat revenu meet wall street earn estim cooper
run consult busi told juri new york wednesday extern auditor arthur andersen approv worldcom account earli  said andersen given green light procedur practic use
worldcom ebber lawyer said unawar fraud argu auditor not alert problem cooper also said sharehold meet ebber often pass technic question compani financ chief giv
e brief answer prosecut star wit former worldcom financi chief scott sullivan said ebber order account adjust firm tell hit book howev cooper said sullivan not m
ention anyth uncomfort worldcom a...'
```

```
[ ] df.shape,df1.shape,df_text.shape
```

```
[ ] from sklearn.preprocessing import LabelEncoder
    from keras.utils import to_categorical
    le = LabelEncoder()
```

```
[ ] le_y = le.fit_transform(df['Category'])
```

```
[ ] ca_y = to_categorical(le_y)
```

```
[ ] from sklearn.feature_extraction.text import CountVectorizer
    cv = CountVectorizer(min_df=5, ngram_range=(1,4) )
```

```
[ ] cv.fit(df_text)
```

```
CountVectorizer(min_df=5, ngram_range=(1, 4))
```

3.3 Splitting the Data

```
from sklearn.model_selection import train_test_split
```

```
[ ] x_train,x_test,y_train,y_test = train_test_split(x,ca_y,test_size = 0.2, stratify=ca_y,random_state =123)
```

```
[ ] x_vector = cv.transform(x)
```

```
[ ] x_train_vector = cv.transform(x_train)
```

```
[ ] x_test_vector  = cv.transform(x_test)
```

```
[ ] x_train.shape,x_test.shape,y_train.shape,y_test.shape
```

```
((5192,), (1298,), (5192, 5), (1298, 5))
```

```
[ ] x_train_vector.shape,x_test_vector.shape,y_train.shape,y_test.shape
```

```
((5192, 351368), (1298, 351368), (5192, 5), (1298, 5))
```

```
[ ] x_train_vector[0]
```

```
<1x351368 sparse matrix of type '<class 'numpy.int64'>'
        with 827 stored elements in Compressed Sparse Row format>
```

We use train test split Sklearn function to split the data for training and validation

```
[ ] from sklearn.model_selection import GridSearchCV
    from sklearn.ensemble import RandomForestClassifier
```

```
from sklearn.ensemble import RandomForestClassifier
model_rf = RandomForestClassifier()
min_samples_split = np.array([20,25])
max_leaf_nodes    = np.array([5,10])
n_estimators      = np.array([100,150])
param_grid = {'min_samples_split':min_samples_split,'max_leaf_nodes':max_leaf_nodes,'n_estimators':n_estimators}
neigh   = GridSearchCV(model_rf,param_grid,scoring= 'roc_auc',cv= 5,return_train_score = True,verbose =2)
neigh.fit(x_vector,ca_y)
```

```
Fitting 5 folds for each of 8 candidates, totalling 40 fits
[CV] END max_leaf_nodes=5, min_samples_split=20, n_estimators=100; total time=   1.9s
[CV] END max_leaf_nodes=5, min_samples_split=20, n_estimators=100; total time=   1.4s
[CV] END max_leaf_nodes=5, min_samples_split=20, n_estimators=100; total time=   1.5s
[CV] END max_leaf_nodes=5, min_samples_split=20, n_estimators=100; total time=   1.7s
[CV] END max_leaf_nodes=5, min_samples_split=20, n_estimators=100; total time=   1.4s
[CV] END max_leaf_nodes=5, min_samples_split=20, n_estimators=150; total time=   1.9s
[CV] END max_leaf_nodes=5, min_samples_split=20, n_estimators=150; total time=   2.7s
[CV] END max_leaf_nodes=5, min_samples_split=20, n_estimators=150; total time=   2.8s
[CV] END max_leaf_nodes=5, min_samples_split=20, n_estimators=150; total time=   2.2s
[CV] END max_leaf_nodes=5, min_samples_split=20, n_estimators=150; total time=   2.0s
[CV] END max_leaf_nodes=5, min_samples_split=25, n_estimators=100; total time=   1.4s
[CV] END max_leaf_nodes=5, min_samples_split=25, n_estimators=100; total time=   1.4s
[CV] END max_leaf_nodes=5, min_samples_split=25, n_estimators=100; total time=   1.8s
[CV] END max_leaf_nodes=5, min_samples_split=25, n_estimators=100; total time=   1.5s
[CV] END max_leaf_nodes=5, min_samples_split=25, n_estimators=100; total time=   1.5s
[CV] END max_leaf_nodes=5, min_samples_split=25, n_estimators=150; total time=   2.1s
[CV] END max_leaf_nodes=5, min_samples_split=25, n_estimators=150; total time=   2.0s
```

3.4 Model Building

We will deploy as many models as possible and fine tune them using GridsearchCV select the model which are performing with highest accuracy and select those models and ensemble them using voting classifier for the best model. Perform the model evaluation.

```
[ ] from sklearn.metrics import classification_report, accuracy_score

[ ] y_pred = model_rf.predict(x_test_vector)

[ ] print(classification_report(y_test,y_pred))

              precision    recall  f1-score   support

           0       1.00      0.98      0.99       302
           1       1.00      0.98      0.99       238
           2       1.00      0.97      0.98       240
           3       1.00      1.00      1.00       284
           4       1.00      0.96      0.98       234

   micro avg       1.00      0.98      0.99      1298
   macro avg       1.00      0.98      0.99      1298
weighted avg       1.00      0.98      0.99      1298
 samples avg       0.98      0.98      0.98      1298

/usr/local/lib/python3.10/dist-packages/sklearn/metrics/_classification.py:1344: UndefinedMetricWarning: Precision and F-score are ill-defined and being set to 0.
  _warn_prf(average, modifier, msg_start, len(result))

[ ] print(accuracy_score(y_pred,y_test))

0.98125
```

3.5 Create Front End User Module using flask

Once the model is created download and save the model and now we create GUI for front end user using the flask incorporated with HTML, CSS. Align and map the user data to the data base created. From user data create the data frame and load it to the model for the prediction the same prediction is send back to the user GUI and well saved in the data base (MySQL).

3.6 Testing the Model

- ✓ Verify whether the application is the loading on the local server instance.
- ✓ Verify whether the user can access the application.
- ✓ Verify the user can access the different fields for selection and can be visible
- ✓ Once the user selection the fields and made the submit
- ✓ Check the user can get the result or prediction.
- ✓ Once he gets the prediction.
- ✓ Check the data form the user and prediction from the model is loaded into the local MySQL

- ✓ Verify whether the application is the loading on the web service instance.
- ✓ Check the database and download the data…

```python
print('Tokenizer loaded...')

model_rf = joblib.load('artifacts/models/model_rf.pkl')
print('Model Loaded..')



app = Flask(__name__)



@app.route('/')
def welcome_user():
    return render_template('index.html')


@app.route('/submit', methods = ['POST','GET'])
def submit():
    back = request.referrer
    if request.method == 'POST':
        data  = request.form['entered_text']
        X =  text_pipeline(data,tokenizer)
        prediction = model_predict_rf(X,model_rf)
        print(prediction[0])
        return render_template('index.html',result = prediction[0].upper())
    return redirect(back)


if __name__ == '__main__':
    app.run(debug=True)
```

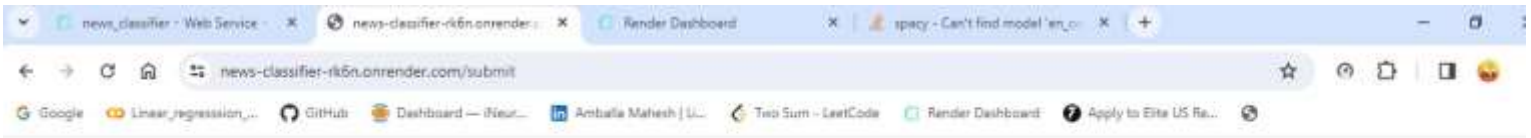
```
      background-color: #45a049;
    }

    div {
      border-radius: 5px;
      background-color: #f2f2f2;
      padding: 20px;
    }
    </style>
  <body>
<h2>News Classifier</h2>



<form action="/submit" method = 'post'>
    <label for="entered_text">Please enter your article -- </label><br>
    <input type= 'text' id = "entered_text" name= "entered_text"><br><br>
    <input type="submit" value="Classify">
</form>

<h3 style = color:blue;>This article is classified as - {{result}}</h3>



</body>
</html>
```

News Classifier

Please enter your article --

Classify

This article is classified as -

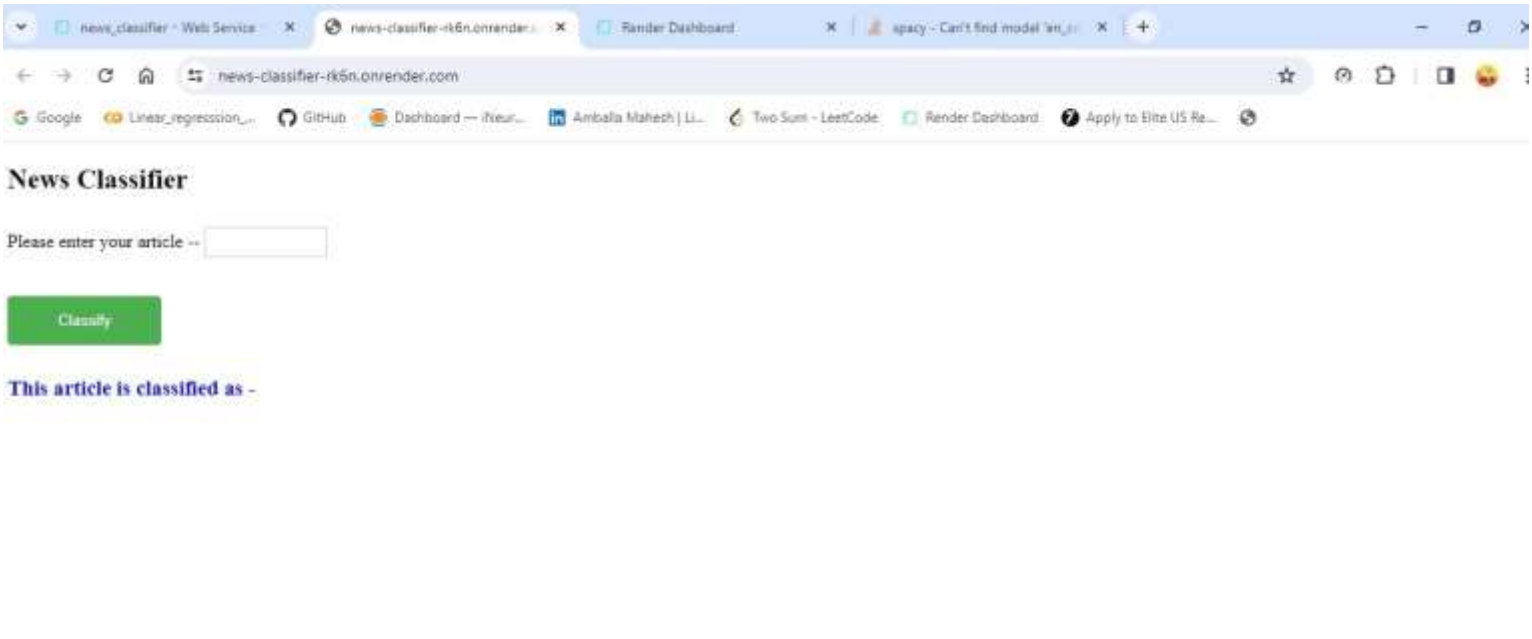


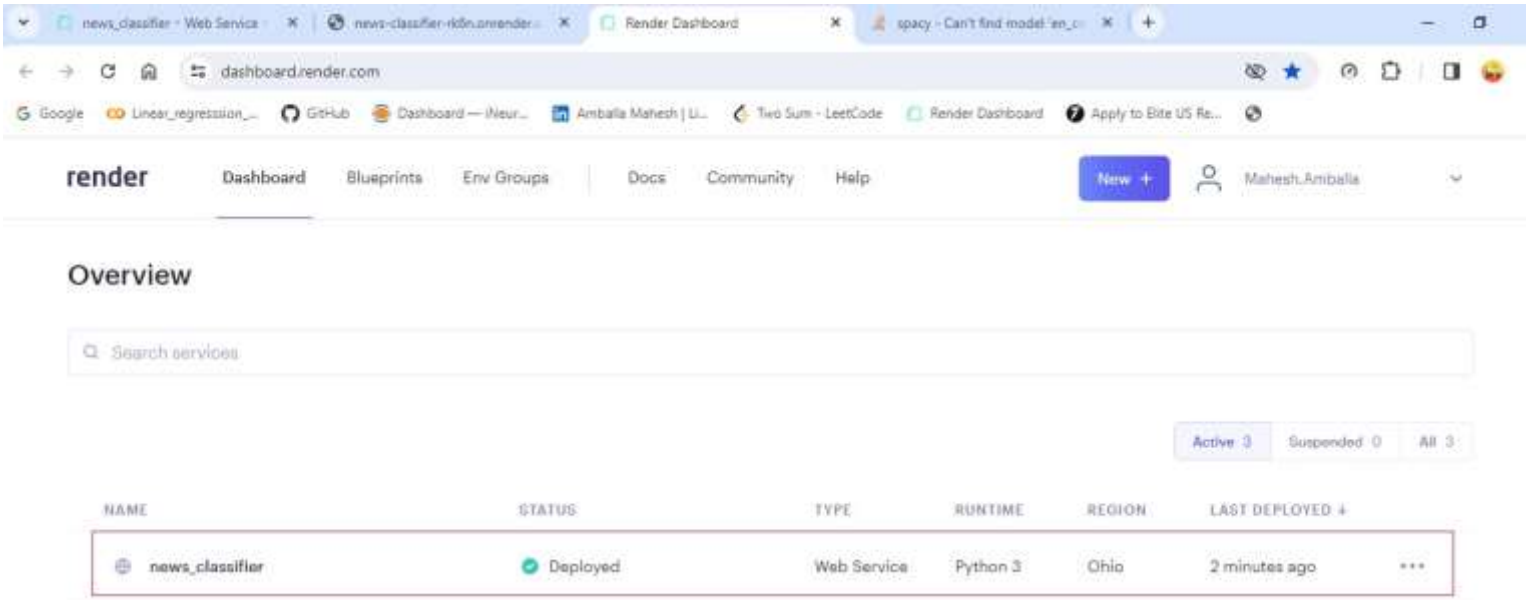
News Classifier

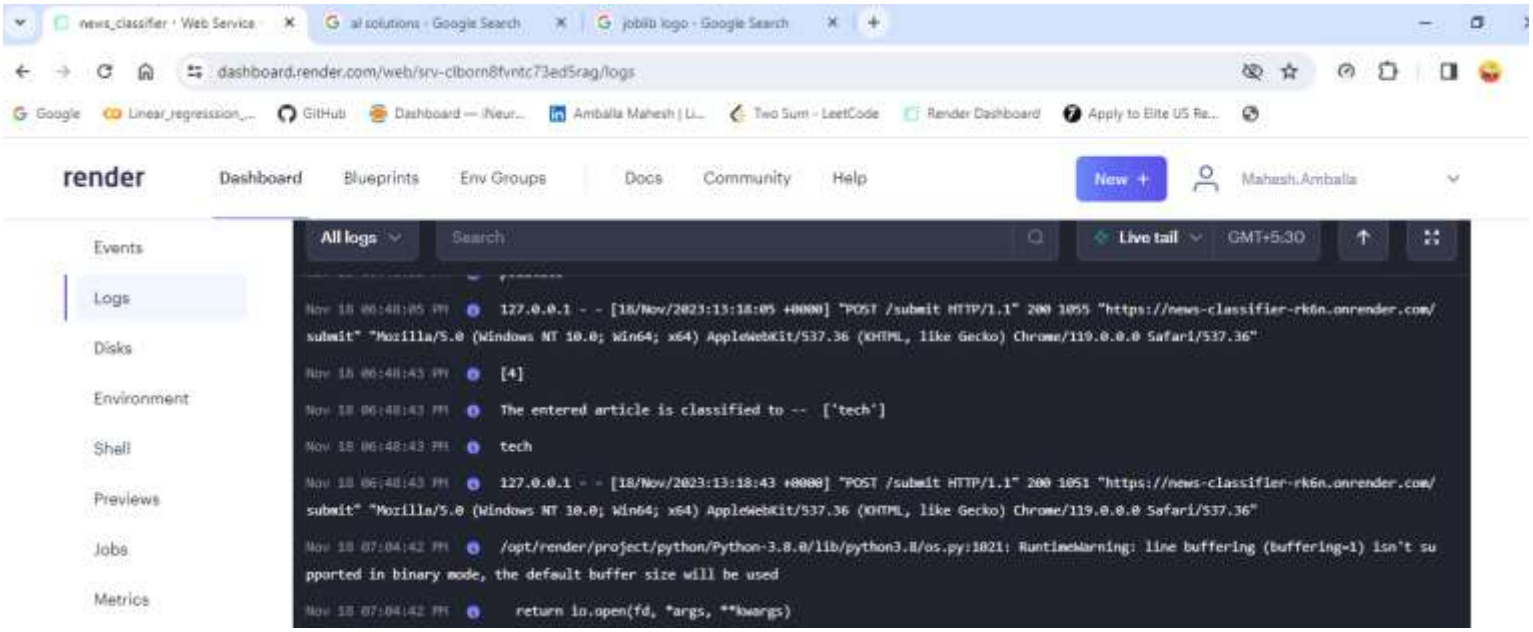
Please enter your article -- rates may rise.

Classify

This article is classified as - BUSINESS


news_classifier - Web Service
news-classifier-rk6n.onrender.com
Render Dashboard
news-classifier-rk6n.onrender.com
Google
Linear_regression_...
GitHub
Dashboard — iNeur...
Amballa Mahesh | Li...
Two Sum - LeetCode
Render Dashboard
Apply to Elite US Re...
News Classifier
Please enter your article --
Classify
This article is classified as -


news_classifier - Web Service
Render Dashboard
dashboard.render.com
render
Dashboard
Blueprints
Env Groups
Docs
Community
Help
New +
Mahesh.Amballa
Overview
Search services
Active 3
Suspended 0
All 3
NAME
STATUS
TYPE
RUNTIME
REGION
LAST DEPLOYED
news_classifier
Deployed
Web Service
Python 3
Ohio
2 minutes ago



HOST WEB ADDRESS: https://news-classifier-rk6n.onrender.com

4. Key performance indicators (KPI)

- Time and work load reduction by using the flask model.
- Compare the accuracy of model using prediction and actual results.
- Check for the wrong predictions
- If found any wrong predictions again train the model with the new data along with previous data
- Retest the model unless the productions attain the good results.