End-to-End Pipeline for News Classification Using Machine Learning Techniques

Major goal of the project is to develop an end-to-end pipeline for the classification of news articles from a news website to different categories. The entire pipeline can be seen in figure 1.

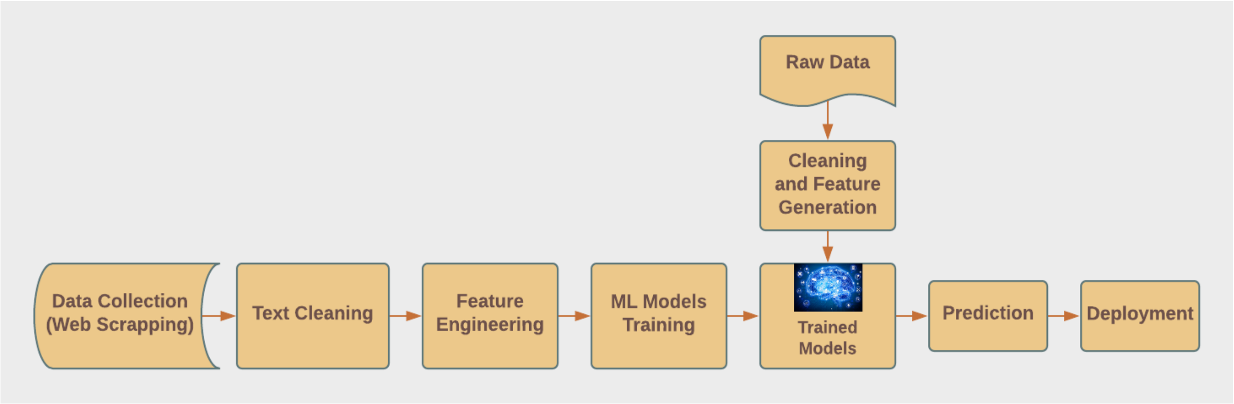


Fig1: Complete pipeline and the workflow of the project

Data Collection, Analysis and Cleaning

Initially data is collected the datasets from 14 different classes including [‘races', 'economy', 'technology', 'international', 'business', 'cricket','society', 'science', 'travel', 'markets','movies', 'entertainment','national', 'other sports'] which is later reduced to seven - **1. Business;** which includes Business, economy, markets, **2. Sports**; which includes cricket, races and other sports, **3. Entertainment**; which includes movies and entertainment **4. Sci-tech;** which includes science and technology **5. National;** 6. **International;** and 7. **Society.** Class travel is dropped because of its low presence .The distribution of 14 class and 7 class can be seen in the figure 2.

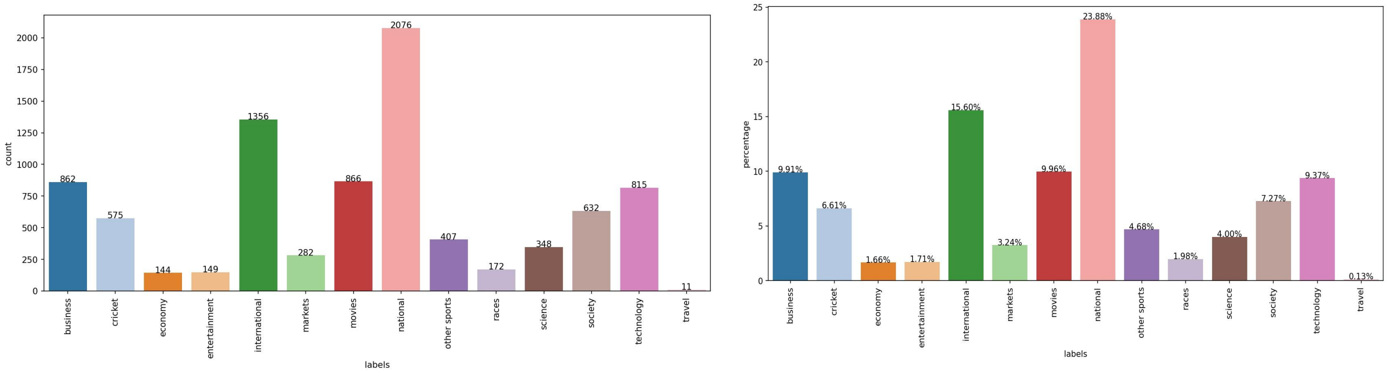


Fig2: Class wise distribution of the 14 class datasets scrapped from the first 60 days of 2020

Looking at very high imbalance, more data from the classes with less no of examples are extracted from next 60-100 and 100-300 days. The distribution of the final datasets with 7-classes with respect to the count of each label can be seen in fig 3.

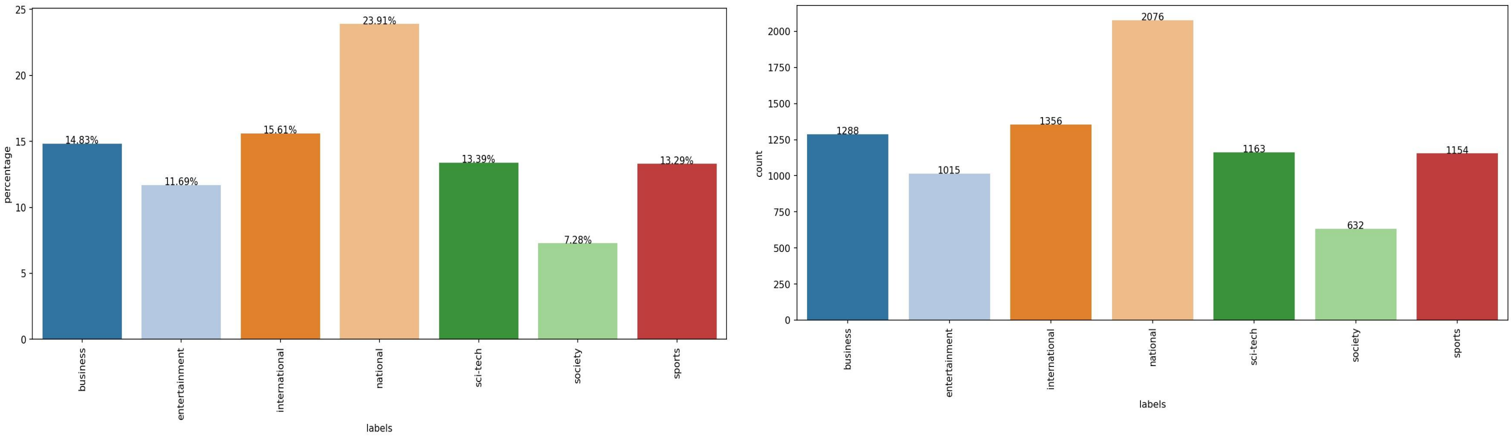


Fig3: Class wise distribution of the complete datasets

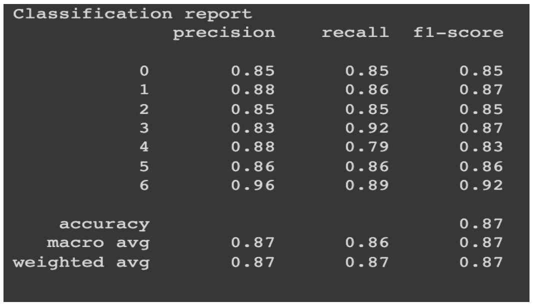
After collecting the datasets, they are cleaned to remove the random text noise, numbers, punctuations, urls, tweets and extra spaces. We also tried removing stopwords and lemmatizing the words but they didn't improve the result so we dropped these ideas.

Feature Engineering, Training and Cross-Validation

The cleaned dataset is then split into train, val, and test. We used TF-IDF to vectorize the text with a maximum feature of 300 and used both unigram and bigram of the words as features. We trained three different algorithms and cross-validated them. Summary of the results of different algorithms can be seen in table 1.

Algorithms	Training Accuracy	Test Accuracy
Random Forests	100%	81.81%
SVM	100%	86.65%
Naive Bayes	77.76%	77.73%

Table 1: Performance of different models



```
Classification report
              precision    recall  f1-score

           0       0.85      0.85      0.85
           1       0.88      0.86      0.87
           2       0.85      0.85      0.85
           3       0.83      0.92      0.87
           4       0.88      0.79      0.83
           5       0.86      0.86      0.86
           6       0.96      0.89      0.92

    accuracy                           0.87
   macro avg       0.87      0.86      0.87
weighted avg       0.87      0.87      0.87
```

Table 2: Classification Report of SVC

Classification reports and Confusion matrix can be shown in figure 4. We found that SVM is performing the best.

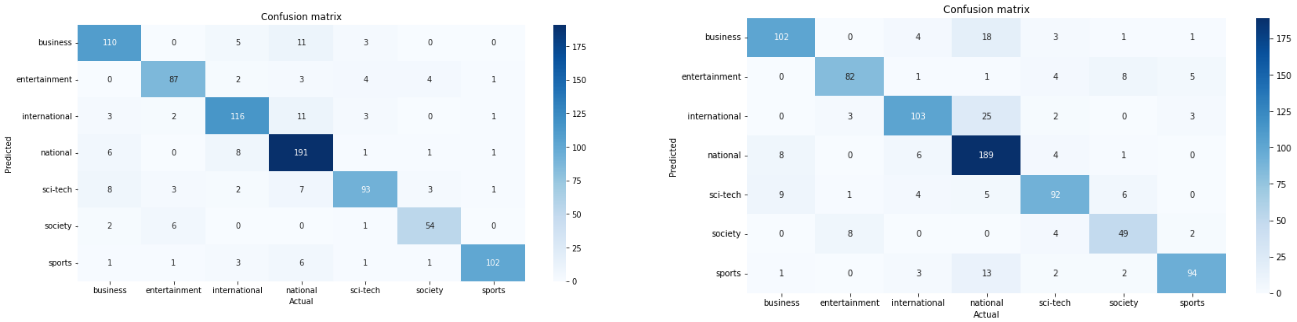
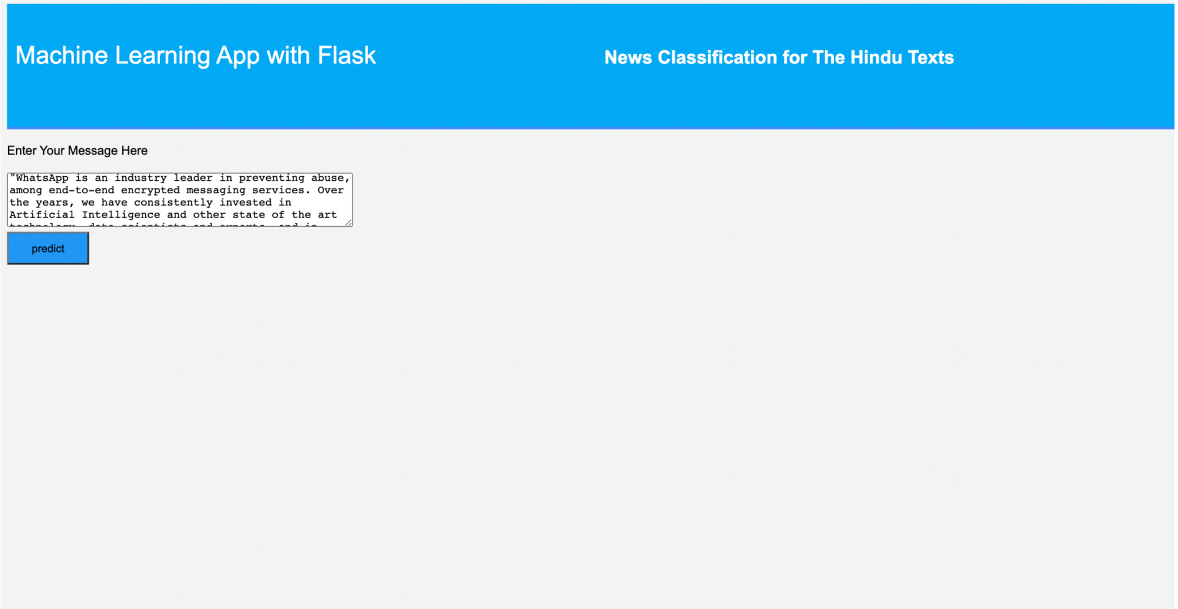


Figure 4: Confusion Matrix for SVM(left) and Random Forest(right) Models on test set

Deployment

Finally we deployed the model using flask as a simple web application where users can copy the news in the form to get the output. The UI can be seen in the figure 5a and 5b.



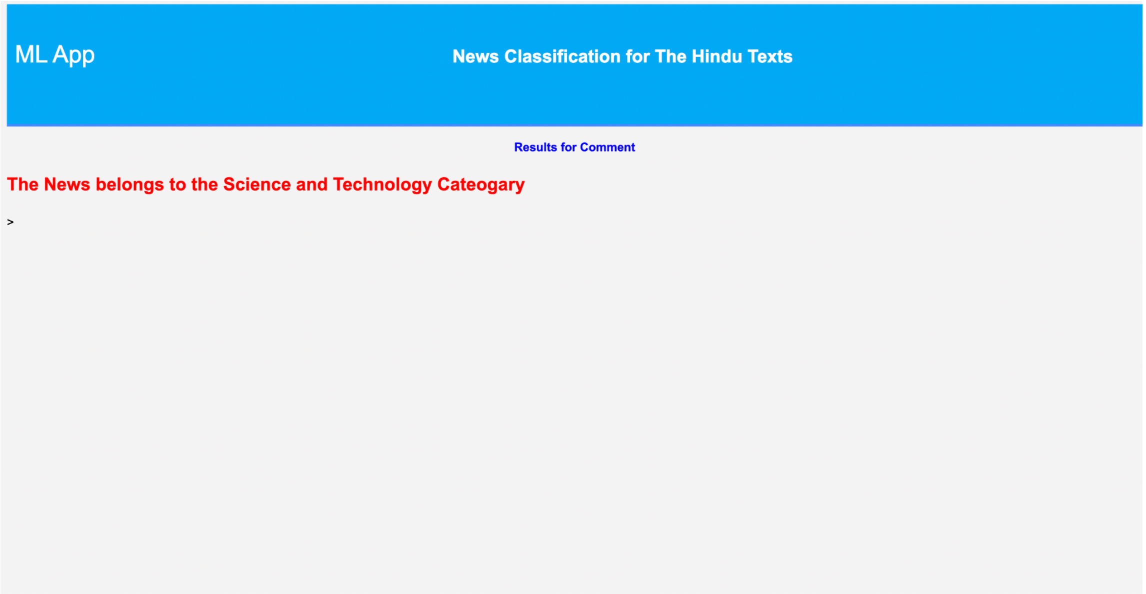
Machine Learning App with Flask

News Classification for The Hindu Texts

Enter Your Message Here

"WhatsApp is an industry leader in preventing abuse, among end-to-end encrypted messaging services. Over the years, we have consistently invested in Artificial Intelligence and other state of the art technology, data scientists and experts, and in

Fig 5a: Home page of the web app



ML App

News Classification for The Hindu Texts

Results for Comment

The News belongs to the Science and Technology Cateogary

>

Fig 5b: Result page of the web app