News Article Sorting

Detailed Project Report

Pauline IC
ineuron.ai

Table of Contents

1. Introduction:

1.1 Objective:

Data is power in today's world. Since news organizations have terabytes of data kept on servers, everyone is trying to find insights that will benefit the company. The classification of news articles stands out among the many examples I could give of how analytics is being utilized to motivate actions. There are many sites that produce a huge amount of news every day today on the Internet. Additionally, user demand for information has been steadily increasing, thus it is critical that the news be classified to enable users to quickly and effectively obtain the information of interest. In this project, the machine learning model for automatic news classification is discussed. This technique could be used to find untracked newspaper articles and/or provide personalized recommendations based on the user's past preferences.

1.2 Results:

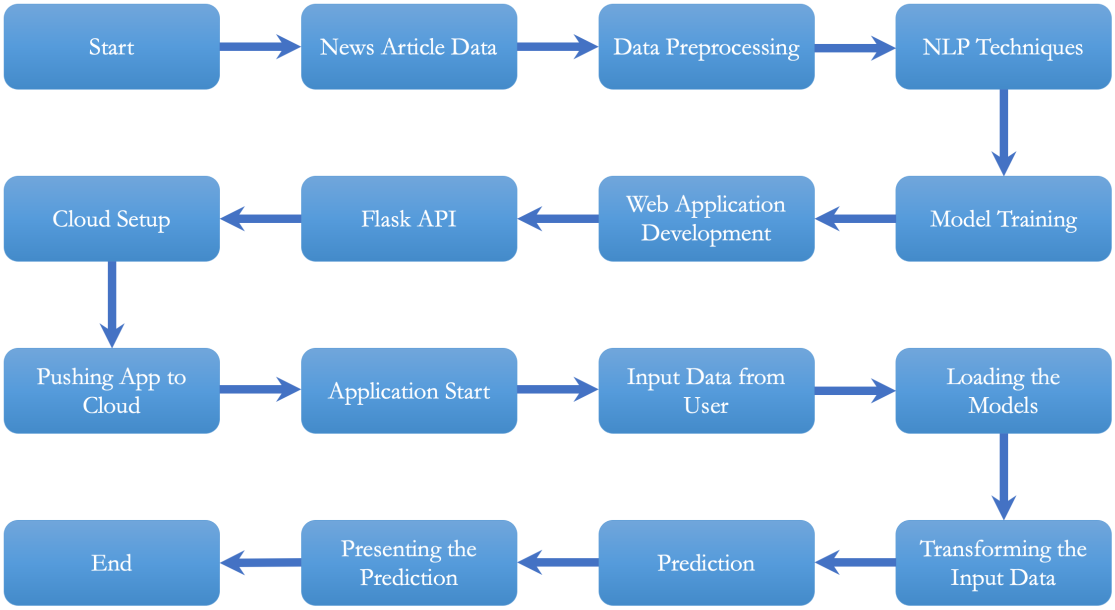
The proposed model organizes news items into categories so that users can get the best recommendations.

2. Project Scope:

The HLD documentation outlines the system's architecture, including the technology architecture, application architecture (layers), application flow, and database architecture. The HLD employs simple to moderately complex terms that system administrators should be able to understand.

3. Project methodology:

3.1 Architectural diagram



3.2 Terminologies

Term	Description
ML	Machine Learning
API	Application Programming Interface
Flask	Framework for deploying the model
AWS	Amazon Web Service
EC2	Elastic Compute Cloud
NLP	Natural Language Processing
SSH	Secure Shell

3.3 System requirements

- Windows 7 and above
- SQL
- PyCharm
- HTML
- CSS
- WinSCP
- AWS Account

3.4 Tools used:

pandas
NLTK
reg[ular]
expr[essio]n
scikit
learn
PC
HTML
5
Flask
Amazon
EC2

3.5 Interfaces

- Input and output news articles are of text format SSH (Secure Shell)
- SSH (Secure Shell) protocol is used to transfer text messages
- Syntax error and logical errors are taken into consideration

3.6 Error Handling

The model handled getting inputs, grammatical errors, vector conversion errors, pickle loading, and data transformations with separate exception handlers.

3.7 Performance

Expected response times

- The system logged every event so that the user knows which process is running internally.
- The system identifies at which step logging required.
- The system logged each and every system flow

3.8 Resource usage

When a task is performed, it used all the processing power available until its work done.

4. Project Execution:

4.1 Data Export:

The accumulated data from database is exported in csv format for model training.

4.2 Data preprocessing:

- Category names are converted to numerical forms
- Duplicate values are removed
- Stop words are removed.

4.3 Word to Vectors:

- Each word is converted to vector using TF-IDF vectorization method.
- Maximum of 5000 features is taken into consideration.

4.4 Train Test Split:

Data are separated for training and testing purpose. For testing purpose 30% of data is used.

4.5 Model Training:

The models used for training are logistic Regression, SVC, linear SVC, Decision Tree Classifier, Random Forest Classifier, K Neighbors Classifier, Multinomial Naïve Bayes Classifier, Gaussian Naïve Bayes, and Bernoulli's Naïve Bayes. Among all these methods Multinomial Naïve Bayes performed well.

4.6 Performance Evaluation:

The model's performance was assessed using accuracy score. A 97% accuracy rate was achieved via the Multinomial Naive Bayes algorithm.

5. Deployment:

The model is deployed using an AWS EC2 instance. Both HTML and CSS were used to design the website. The flask API was used to deploy the model in an AWS EC2 instance.

6. Conclusion:

The document included a thorough description of the News Article Sorting Project. News Article Sorting will classify every news article to different categories. This is done based on the learning made by the model. The model is trained with thousands of news articles with their classification to do better prediction. The model could classify any news article with 97% accuracy. Anyone can utilize the model because it has been installed in an AWS EC2 instance. The project can be enhanced by including user-specific recommendations.