News Classifier Project Report

Objective:

To train machine learning model and classify upcoming news on the fly with good accuracy by building end to end machine learning pipeline.

To make containerized application which is scalable, robust, fault tolerant.

Planning:

We are using agile methodology to build the project. Task level details are mentioned in below gantt chart.

#Sprint: 4

#People: 2

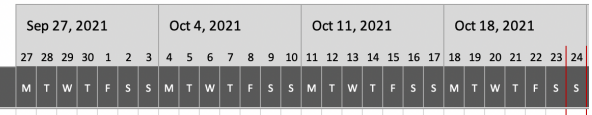
Project members:

Biren Patel, Vaishali khurana

Project Start: Tue, 9/28/2021

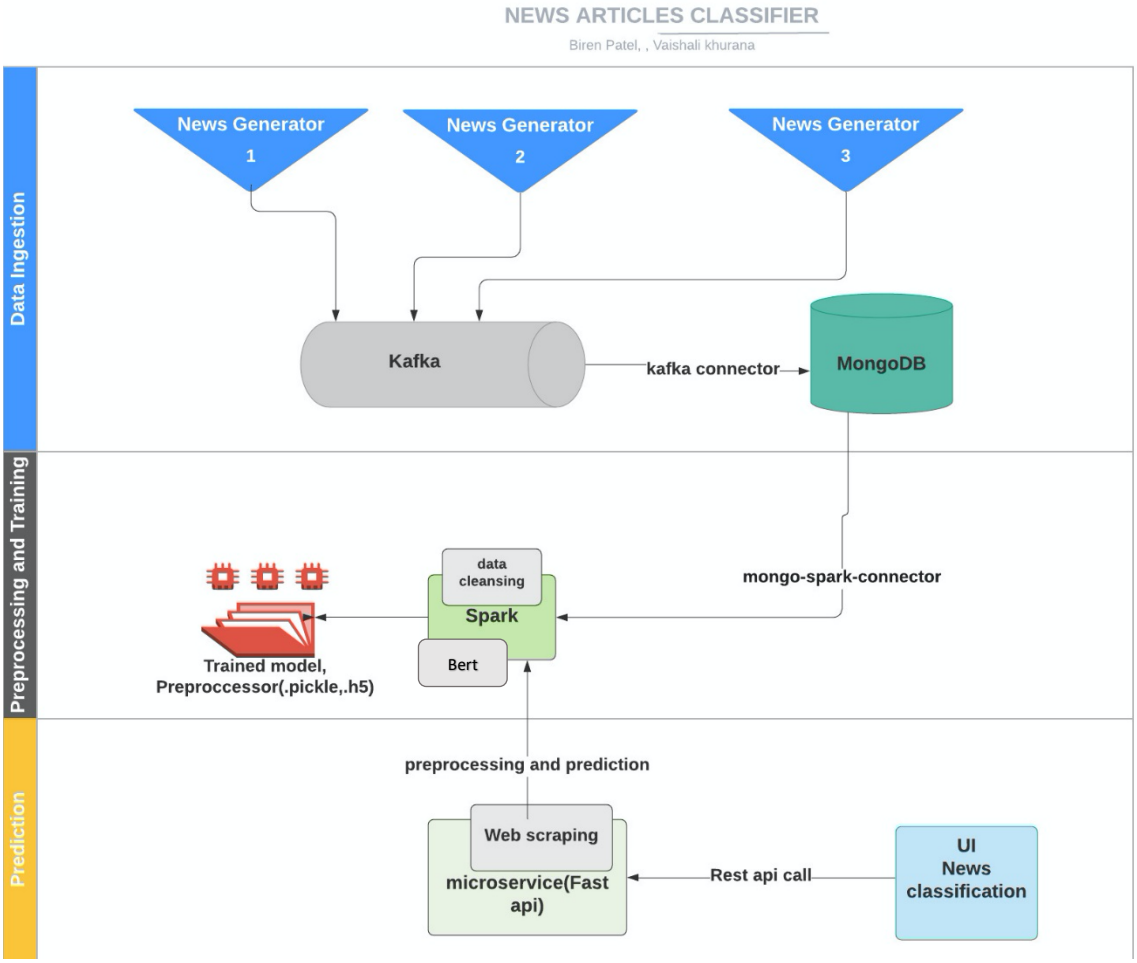
Display Week: 1

TASK	PROGRESS	START	END
Milestone 3: Model Training			
Milestone Planning	100%	10/18/21	10/18/21
Training Spark ml model	100%	10/18/21	10/19/21
Training Pre trained Bert(Transformer) model	100%	10/20/21	10/21/21
Retraining Model	100%	10/22/21	10/22/21
Model saving, loading and deployment	100%	10/22/21	10/22/21
Integration with mlflow	90%	10/23/21	10/24/21
Milestone report	100%	10/24/21	10/24/21



News Classifier Project Report

Architecture:



Components and description:

Component	Description
Data Ingestion	Multithreaded service that is collecting the data from web using rapid api and custom news generator, passing it to kafka queue. Kafka consumer is used to sync the data between kafka and database. Finally, data is dumped into MongoDb database
Preprocessor and trainer	Reading data from mongo db to spark session using mongo-spark-connector. As a part of feature selection we are using category and summary columns.

	Data cleansing including stop word removal, tokenization, tf-idf vectorization using pyspark. This cleaned data is used for training bert model and achieved training accuracy of 87.8% and test accuracy of 82.9%.
Prediction	Scraping the weblink provided by user in UI, clean it and predict the news using rest api.

Since model training is the third milestone we have explained it in detail here.

Model Training:

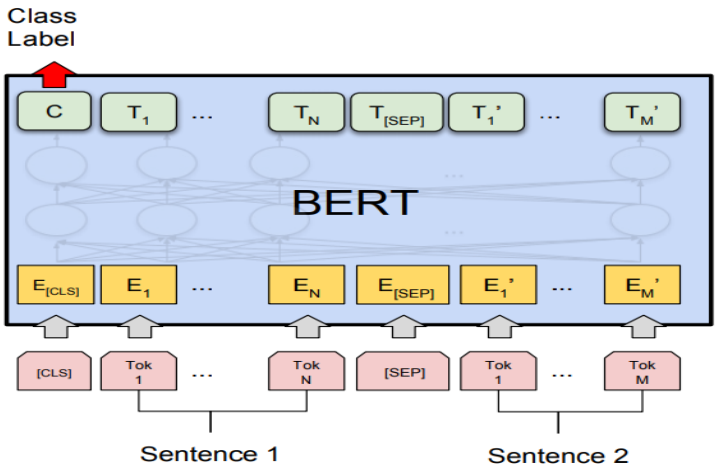
- **Environment details:**

  Docker environment with kafka broker running on 9092 port, zookeeper running on 2181, mongo-db running on 27017, mongo-express running on 8082, spark-master running on 8080, spark-worker running on 8081, producer is connecting to kafka and producing news records to kafka queue, consumer is consuming news records from kafka and dumping to mongo db, preprocessor_trainer is cleansing the news data and training machine learning model on spark node using pyspark.

  Hardware Details: NVIDIA GPU

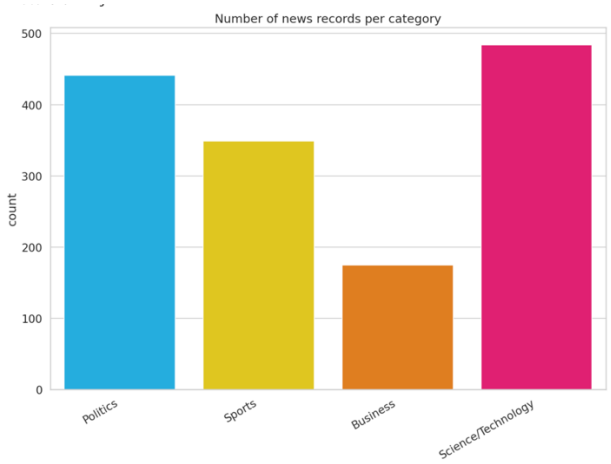
  Tensorflow: 2.0.0

  Pretrained Bert model(Transformer): uncased_L-8_H-512_A-8



News Classifier Project Report

- **What goes in as an input:**
  Spark dataframe having cleaned data.

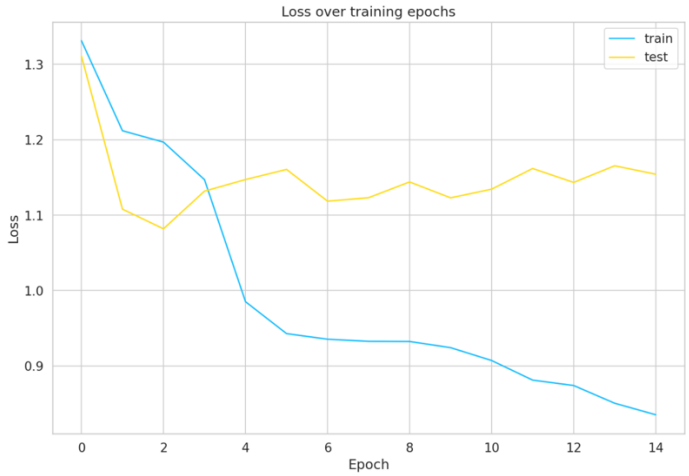


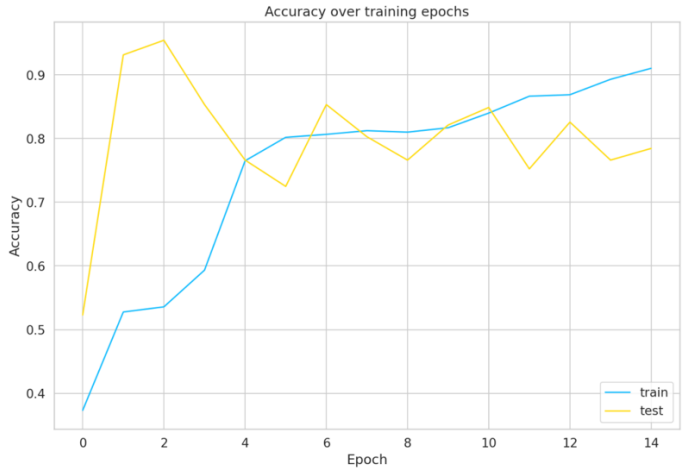
- **How the input is being processed:**
  After the records from mongo db collection is processed and cleaned, these records are further used for training using pyspark over spark node which is connected through mongo_spark_connector.
  Input data is divided into 80% training and 20% test data.
  Pretrained Bert model(uncased_L-8_H-512_A-8) is also used for training for 15 epochs with training accuracy of 87.8% and test accuracy of 82.9%, with following parameters:

```
Model: "model"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
input_ids (InputLayer)       [(None, 512)]             0
_________________________________________________________________
bert (BertModelLayer)        (None, 512, 512)          41109504
_________________________________________________________________
lambda (Lambda)              (None, 512)               0
_________________________________________________________________
dropout (Dropout)            (None, 512)               0
_________________________________________________________________
dense (Dense)                (None, 512)               262656
_________________________________________________________________
dropout_1 (Dropout)          (None, 512)               0
_________________________________________________________________
dense_1 (Dense)              (None, 4)                 2052
=================================================================
Total params: 41,374,212
Trainable params: 41,374,212
Non-trainable params: 0
_________________________________________________________________
```

News Classifier Project Report

- **What comes out as an output:** The trained model and tokenizer is saved in registry.





	precision	recall	f1-score	support
Science/Technology	0.78	0.98	0.87	121
Sports	0.89	0.78	0.83	87
Politics	0.99	0.83	0.90	111
Business	0.53	0.52	0.53	44
accuracy			0.83	363
macro avg	0.80	0.78	0.78	363
weighted avg	0.84	0.83	0.83	363

- **Tools/libraries used**: Docker, Pycharm, kafka, MongoDB, zookeeper, Spark, pyspark, MongoExpress, colab notebook, Tensorflow, seaborn, mlflow.

News Classifier Project Report

Challenges encountered and how we tackled them:

We faced below mentioned challenges:

- Loading model from HDFS
    - Configured HDFS in spark node.
- Python-Tensorflow version compatibility
    - Installed compatible libraries versions
- Configuring mlflow with docker container

Future Scope:

- Further scale optimizations
- Implementing re-training mechanism using feedback feature
- Once we have large volume of labeled data, we will train our own model in place of transfer learning
- Adding multi language support

Github link:

https://github.com/biren162/Capstone