

News Classifier Project Report

Objective:

To train machine learning or deep 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

Sprint:1

Project members:

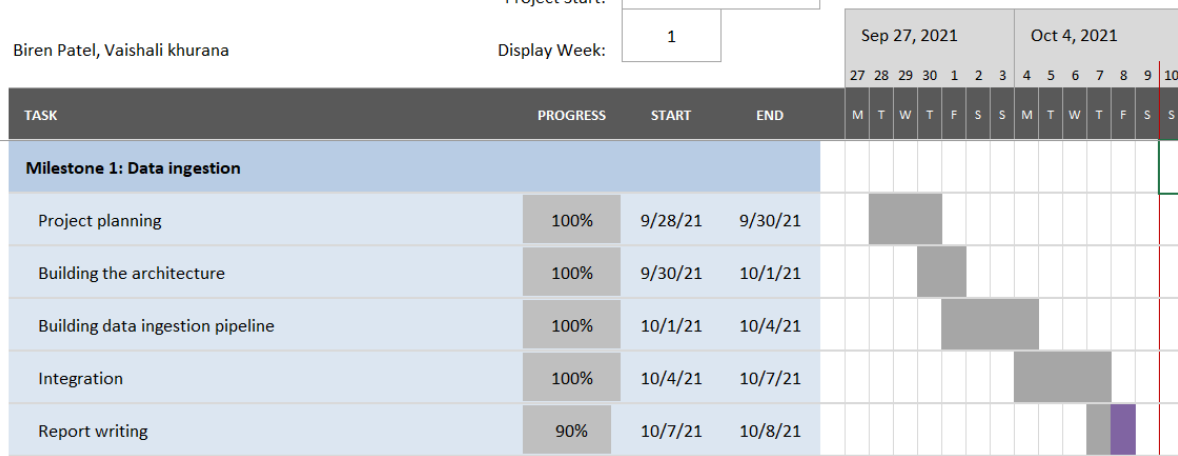
Biren Patel, Vaishali khurana

Project Start:

Tue, 9/28/2021

Display Week:

1



Sprint:2

Project members:

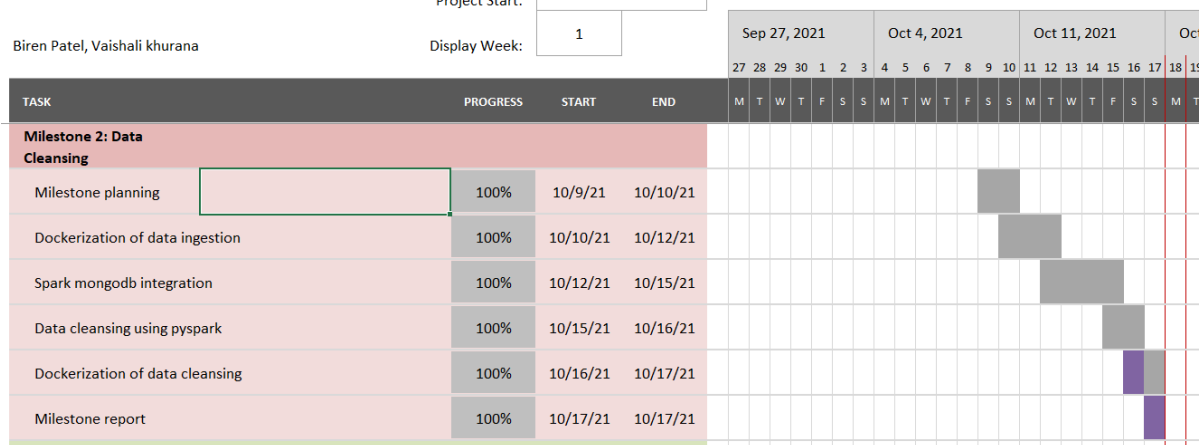
Biren Patel, Vaishali khurana

Project Start:

Tue, 9/28/2021

Display Week:

1



News Classifier Project Report

Sprint:3

Project members:

Project Start:

Tue, 9/28/2021

Biren Patel, Vaishali khurana

Display Week:

1

Biren Patel, Vaishali khurana	Display Week:	1		Sep 27, 2021	Oct 4, 2021	Oct 11, 2021	Oct 18, 2021
TASK	PROGRESS	START	END	M T W T F S S	M T W T F S S	M T W T F S S	M T W T F S S
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				

Sprint:4

Project members:

Project Start:

Tue, 9/28/2021

Biren Patel, Vaishali khurana

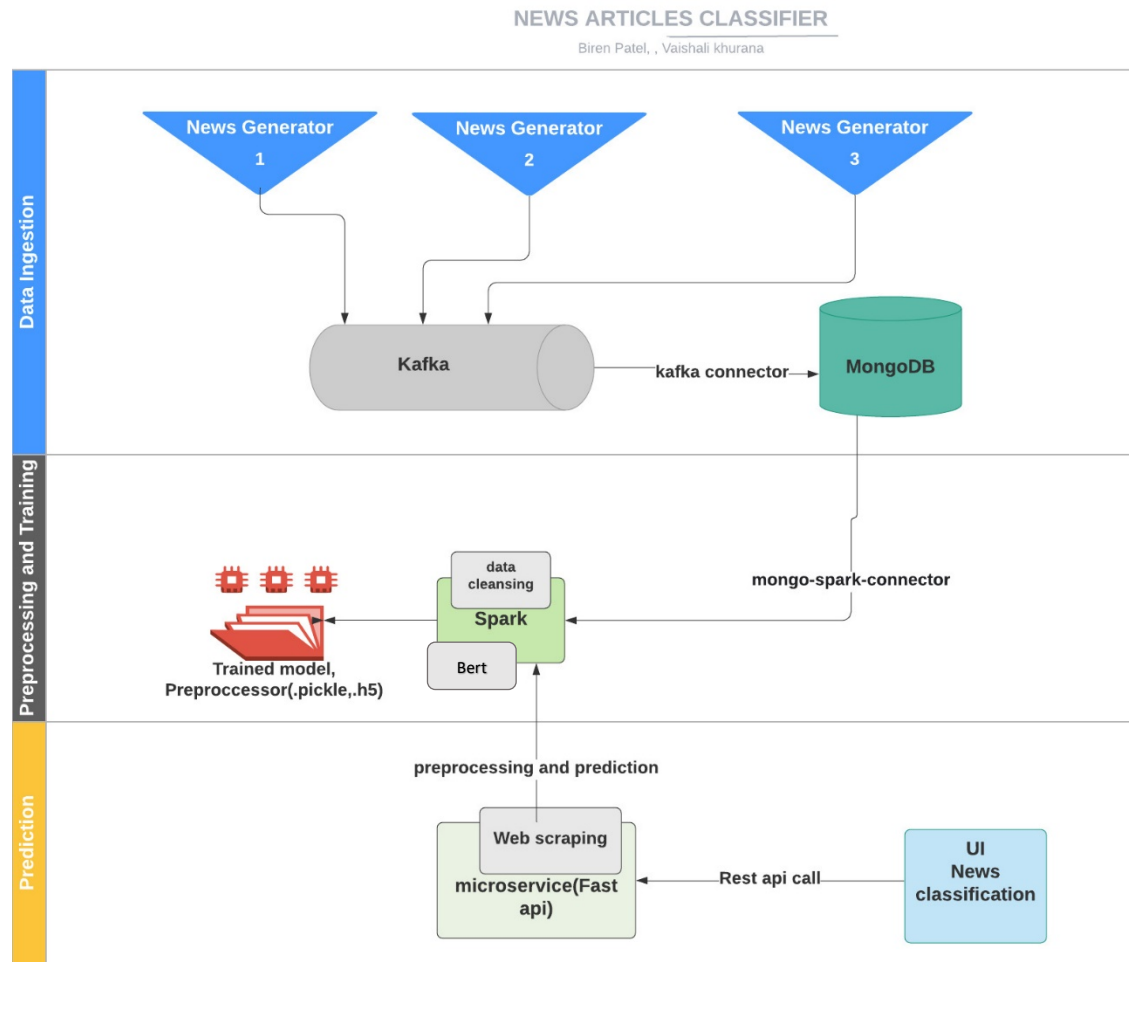
Display Week:

1

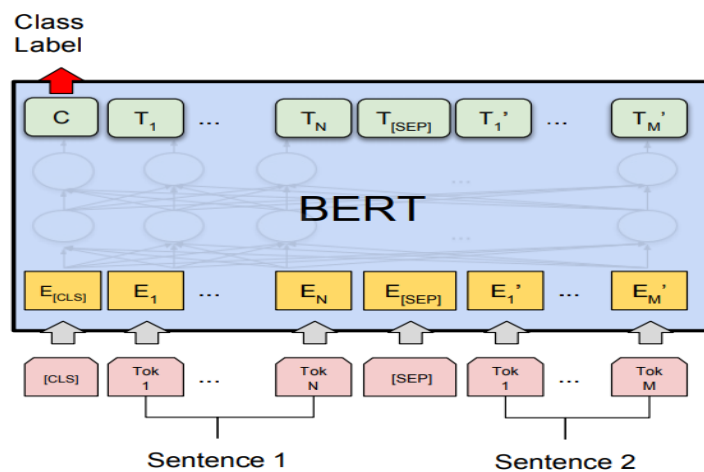
Biren Patel, Vaishali khurana				Display Week: 1				Sep 27, 2021			Oct 4, 2021			Oct 11, 2021			Oct 18, 2021			Oct 25, 2021																						
								27	28	29	30	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	
Task		Progress		Start		End		M	T	W	T	F	S	S	M	T	W	T	F	S	S	M	T	W	T	F	S	S	M	T	W	T	F	S	S	M	T	W	T	F	S	S
Milestone 4: Model Prediction																																										
Milestone Planning		100%		10/25/21		10/25/21																																				
Prediction using trained Bert model		100%		10/25/21		10/26/21																																				
Fast API Development and Integration		100%		10/26/21		10/27/21																																				
UI Development and Integration		100%		10/27/21		10/28/21																																				
Milestone Report		100%		10/29/21		10/29/21																																				

News Classifier Project Report

Architecture:



Bert model (Transformer) Architecture:



News Classifier Project Report

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%.
Predictor and UI	Predicting news category based on the news summary entered by user in UI. Clicking “Predict” button calls backend rest API which firstly processes the data using the trained tokenizer and then predict using the trained Bert model.

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, predictor service running on 8888 port.

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 and predictor service has UI integration with Fast API in backend.

- **Hardware Details:**

- NVIDIA GPU
- Memory requirements: Minimum 4GB RAM dedicated for docker

News Classifier Project Report

- **Tools/libraries used:**
 - Docker
 - Pycharm
 - Kafka
 - Zookeeper
 - MongoDB
 - MongoExpress
 - Spark
 - Pyspark
 - colab notebook
 - Tensorflow : 2.0.0
 - Seaborn
 - mlflow
 - Pretrained Bert model(Transformer): uncased_L-8_H-512_A-8
 - Fast API
 - HTML
 - CSS
 - Jinja2Templates

Data Ingestion:

- **What goes in as an input:**

We have used rapid api and custom news generator as data sources.

```
(env) PS C:\Users\v1105800\PG\Capstone\Capstone\kafka to Mysql\news_producer> python news_producer.py
sports
['cricket', 'hockey']
sending... {'title': 'There Were 2 Mass Shootings In Texas Last Week, But Only 1 On TV', 'date': '2018-05-26', 'summary': 'She left her husband. He killed their children. Just another day in America.', 'category': 'CRIME', 'source': 'https://www.huffingtonpost.com/entry/texas-amanda-painter-mass-shooting_us_5b081ab4e4b0802d69caad89'}
sending... {'title': 'PRESS RELEASE: NACON: Announcing Cricket 22: A New Era Of Cricket Games Has Arrived!', 'date': '2021-10-07 06:00:00', 'summary': 'Announcing Cricket 22:\n\n\nA New Era Of Cricket Games Has Arrived!\n\nIncluding The Ashes, Big Bash, The Hundred, Caribbean Premier League, Cricket 22 Is The Biggest Cricket Simulation Ever Made\n\nLesquin, October 7: Big Ant Studios and Nacon are thrilled to announce that the long-awaited Cricket 22: The Official Game of The Ashes will arrive this November. A true next-generation effort that builds on the massive success of Cricket 19, Cricket 22 will deliver the most robust, substantial game of cri', 'category': 'sports', 'source': 'yahoo.com'}
sending... {'title': 'Will Smith Joins Diplo And Nicky Jam For The 2018 World Cup's Official Song', 'date': '2018-05-26', 'summary': 'Of course it has a song.', 'category': 'ENTERTAINMENT', 'source': 'https://www.huffingtonpost.com/entry/will-smith-joins-diplo-and-nicky-jam-for-the-official-2018-world-cup-song_us_5b09726fe4b0fdb2aa541201'}
sending... {'title': 'Cricket 22 Release Date, India Price Announced', 'date': '2021-10-07 07:40:01', 'summary': 'Cricket 22 will be the first cricket game on next-gen consoles, the PlayStation 5 and the Xbox Series S/X. On Thursday, Melbourne-based developer Big Ant Studios announced Cricket 22: The Official Game of The Ashes, the third entry in that series following 2019's Cricket 19. Cricket 22 will be available November 25 on PC, PS4, PS5, Xbox One and Xbox Series S/X, a couple of weeks ahead of the start of the Ashes. Cricket 22 for Nintendo Switch arrives January 2022. In addition to the Ashes, Cricke', 'category': 'sports', 'source': 'ndtv.com'}
sending... {'title': 'Hugh Grant Marries For The First Time At Age 57', 'date': '2018-05-26', 'summary': 'The actor and his longtime girlfriend Anna Eberstein tied the knot in a civil ceremony.', 'category': 'ENTERTAINMENT', 'source': 'https://www.huffingtonpost.com/entry/hugh-grant-marries_us_5b09212ce4b0568a880b9a8c'}
sending... {'title': 'Big Ant Studios Announces Cricket 22', 'date': '2021-10-07 06:05:19', 'summary': 'Big Ant Studios and Nacon have announced the much-awaited follow-up to Cricket 19, called Cricket 22: The Official Game of The Ashes. The developer has revealed the game modes, gameplay enhancements, and release date of Cricket 22: The Official Game of The Ashes. Cricket 22 will not only feature the Ashes competition, but it also brings Australia's Big Bash T20, Caribbean Premier League, and England's The Hundred. The game will come with fully-licensed teams from Australia, England, The West Ind', 'category': 'gaming', 'source': 'ign.com'}
sending... {'title': 'Jim Carrey Blasts 'Castrato' Adam Schiff And Democrats In New Artwork', 'date': '2018-05-26', 'summary': 'The actor gives Dems an ass-kicking for not fighting hard e
```

News Classifier Project Report

- **How the input is being processed:**

Multithreaded application that is collecting the data from web using rapid api and custom news generator, passing it to kafka queue. Kafka consumer is dumping the data into mongo db.

```
ksql> SHOW TOPICS;

Kafka Topic                | Partitions | Partition Replicas |
-----
confluent_rmoff_01ksql_processing_log | 1          | 1                  |
news                             | 1          | 1                  |
-----

ksql> PRINT news FROM BEGINNING LIMIT 5;
Key format: "\_(/)_/" - no data processed
Value format: JSON or KAFKA_STRING
rowtime: 2021/10/10 17:35:37.413 Z, key: <null>, value: {"title": "There Were 2 Mass Shootings In Texas Last Week, But Only 1 On TV", "date": "2018-05-26", "summary": "She left her husband. He killed their children. Just another day in America.", "category": "CRIME", "source": "https://www.huffingtonpost.com/entry/texas-amanda-painter-mass-shooting_us_5b081ab4e4b0802d69cad89"}, partition: 0
rowtime: 2021/10/10 17:35:39.067 Z, key: <null>, value: {"title": "PRESS RELEASE: NACON: Announcing Cricket 22: A New Era Of Cricket Games Has Arrived!", "date": "2021-10-07 06:00:00", "summary": "Announcing Cricket 22:\n\n\nA New Era Of Cricket Games Has Arrived!\n\nIncluding The Ashes, Big Bash, The Hundred, Caribbean Premier League, Cricket 22 Is The Biggest Cricket Simulation Ever Made\n\nLesquin, October 7: Big Ant Studios and Nacon are thrilled to announce that the long-awaited Cricket 22: The Official Game of The Ashes will arrive this November. A true next-generation effort that builds on the massive success of Cricket 19, Cricket 22 will deliver the most robust, substantial game of cri", "category": "sports", "source": "yahoo.com"}, partition: 0
rowtime: 2021/10/10 17:35:42.415 Z, key: <null>, value: {"title": "Will Smith Joins Diplo And Nicky Jam For The 2018 World Cup's Official Song", "date": "2018-05-26", "summary": "Of course it has a song.", "category": "ENTERTAINMENT", "source": "https://www.huffingtonpost.com/entry/will-smith-joins-diplo-and-nicky-jam-for-the-official-2018-world-cup-song_us_5b09726fe4b0fdb2aa541201"}, partition: 0
rowtime: 2021/10/10 17:35:43.069 Z, key: <null>, value: {"title": "Cricket 22 Release Date, India Price Announced", "date": "2021-10-07 07:40:01", "summary": "Cricket 22 will be the first cricket game on next-gen consoles, the PlayStation 5 and the Xbox Series S/X. On Thursday, Melbourne-based developer Big Ant Studios announced Cricket 22: The Official Game of The Ashes, the third entry in that series following 2019's Cricket 19. Cricket 22 will be available November 25 on PC, PS4, PS5, Xbox One and Xbox Series S/X, a couple of weeks ahead of the start o
```

- **What comes out as an output:**

News records saved in news_collection under news database MongoDB.

Viewing Collection: news_collection

[New Document](#) [New Index](#)

Simple

Advanced

[Find](#)

Delete all 139 documents retrieved

← First

← Prev

Next →

Last →

_id	title	date	summary	category	source
616d7fc390f3e115c28e3128	There Were 2 Mass Shootings In Texas Last Week, B...	2018-05-26	She left her husband. He killed their children. J...	CRIME	https://www.huffingtonpost.com/entry/texas-amanda...
616d7fc790f3e115c28e3129	Will Smith Joins Diplo And Nicky Jam For The 2018...	2018-05-26	Of course it has a song.	ENTERTAINMENT	https://www.huffingtonpost.com/entry/will-smith-j...
616d7fca90f3e115c28e312a	Hugh Grant Marries For The First Time At Age 57	2018-05-26	The actor and his longtime girlfriend Anna Eberst...	ENTERTAINMENT	https://www.huffingtonpost.com/entry/hugh-grant-m...

News Classifier Project Report

Data Cleansing:

- **What goes in as an input:**

News records saved in news_collection under news database MongoDB

Viewing Collection: news_collection

New Document

New Index

Simple

Advanced

Key

Value

String

Find

Delete all 139 documents retrieved

← First

← Prev

Next →

Last →

_id	title	date	summary	category	source
<div><div></div><div>616d7fc390f3e115c28e3128</div></div>	There Were 2 Mass Shootings In Texas Last Week, B...	2018-05-26	She left her husband. He killed their children. J...	CRIME	https://www.huffingtonpost.com/entry/texas-amanda...
<div><div></div><div>616d7fc790f3e115c28e3129</div></div>	Will Smith Joins Diplo And Nicky Jam For The 2018...	2018-05-26	Of course it has a song.	ENTERTAINMENT	https://www.huffingtonpost.com/entry/will-smith-j...
<div><div></div><div>616d7fca90f3e115c28e312a</div></div>	Hugh Grant Marries For The First Time At Age 57	2018-05-26	The actor and his longtime girlfriend Anna Eberst...	ENTERTAINMENT	https://www.huffingtonpost.com/entry/hugh-grant-m...

- **How the input is being processed:**

Records from mongo db collection is processed using pyspark over spark node which is connected through 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

```
root
|-- _id: struct (nullable = true)
|   |-- oid: string (nullable = true)
|-- category: string (nullable = true)
|-- date: string (nullable = true)
|-- source: string (nullable = true)
|-- summary: string (nullable = true)
|-- title: string (nullable = true)

cleaning start
['_id', 'category', 'date', 'source', 'summary', 'title']
```

News Classifier Project Report

After feature selection and tf-idf vectorization:

```
root
|-- summary: string (nullable = true)
|-- category: string (nullable = true)
|-- tf: vector (nullable = true)
|-- idf: vector (nullable = true)
|-- label: double (nullable = false)
```

- **What comes out as an output:**

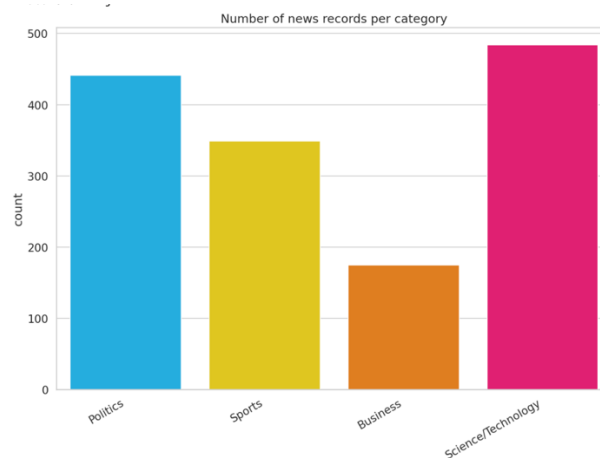
Processed and cleaned data, saved the preprocessed pipeline

```
2021-10-18 14:27:01 INFO DAGScheduler:54 - Job 3 finished: showString at NativeMethodAccessorImpl.java:0, took 0.050267 s
+-----+-----+-----+-----+
| category| summary| tf| idf|label|
+-----+-----+-----+-----+
| CRIME|She left her husb...|(10000,[1662,3562...|(10000,[1662,3562...| 7.0|
| ENTERTAINMENT|Of course it has ...|(10000,[1916,2460...|(10000,[1916,2460...| 1.0|
| ENTERTAINMENT|The actor and his...|(10000,[512,2410,...|(10000,[512,2410,...| 1.0|
| ENTERTAINMENT|The actor gives D...|(10000,[512,1384,...|(10000,[512,1384,...| 1.0|
| ENTERTAINMENT|The "Dietland" ac...|(10000,[2678,3624...|(10000,[2678,3624...| 1.0|
+-----+-----+-----+-----+
only showing top 5 rows
```

Model Training:

- **What goes in as an input:**

Spark dataframe having cleaned data.



News Classifier Project Report

- **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.

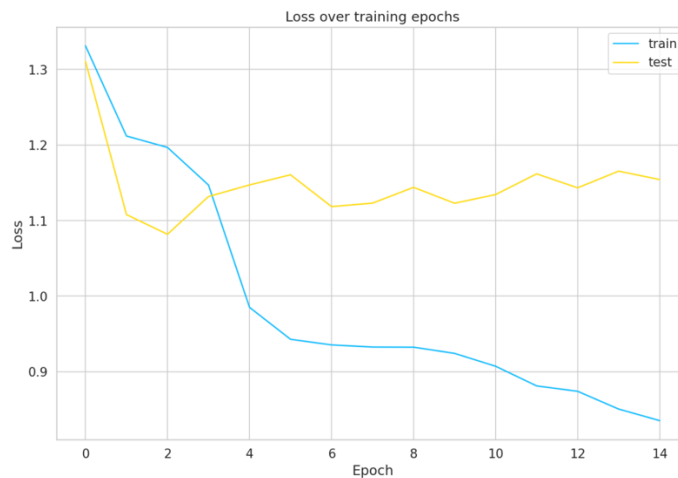
Pretrained Bert model(uncased_L-8_H-512_A-8) is also used for training for 15 epochs with training accuracy of 87.8% and test accuracy of 82.9%, with following parameters:

Model: "model"

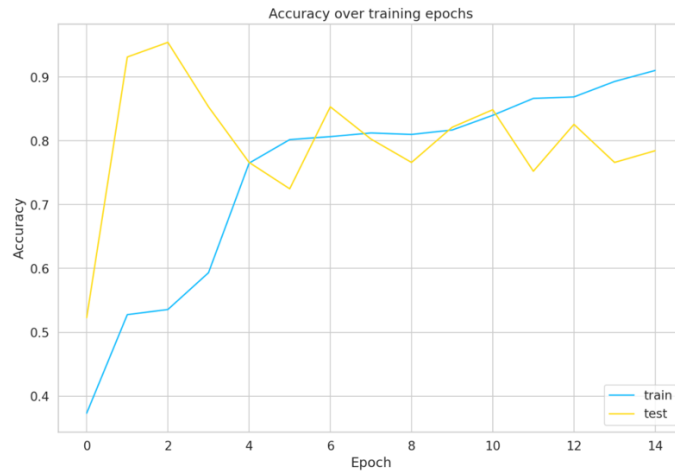
Layer (type)	Output Shape	Param #
input_ids (InputLayer)	[(None, 512)]	0
bert (BertModelLayer)	(None, 512, 512)	41109504
lambda (Lambda)	(None, 512)	0
dropout (Dropout)	(None, 512)	0
dense (Dense)	(None, 512)	262656
dropout_1 (Dropout)	(None, 512)	0
dense_1 (Dense)	(None, 4)	2052
Total params: 41,374,212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

- **What comes out as an output:**

The trained model and tokenizer is saved in registry.



News Classifier Project Report




	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

Model Prediction:

- **What goes in as an input:**
News Summary text entered by user in UI

localhost:8888

 **News Classifier**

[← Go To Home Page](#)

Enter the below feilds to predict the news category.

* All feilds are mandatory

Please enter news text*

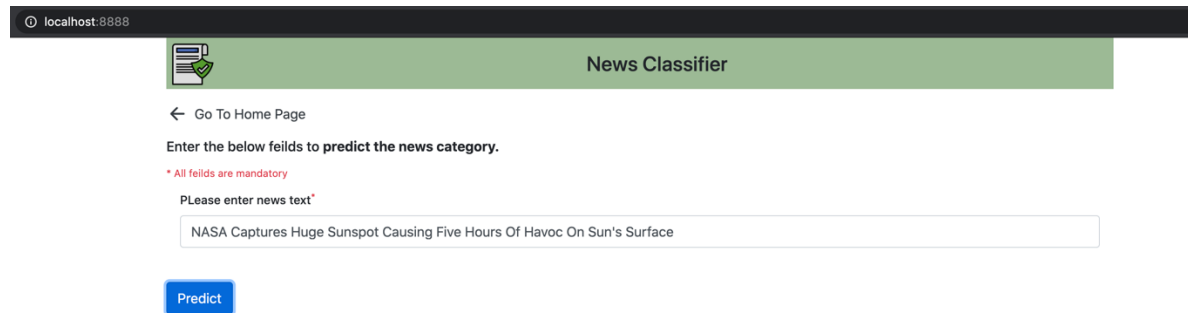
Narendra modi resigned.

Predict

News Classifier Project Report

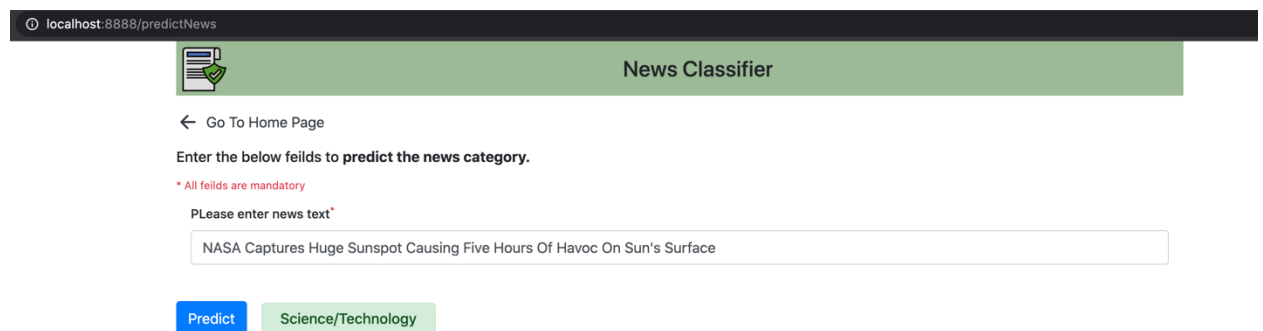
- **How the input is being processed:**

Clicking “Predict” button calls backend rest API which firstly processes the data using the trained tokenizer and then predict using the trained Bert model and returned response is displayed on UI.



- **What comes out as an output:**

Predicted news category is returned to the UI.



Challenges encountered and how we tackled them:

We faced below mentioned challenges:

- Finding better legal data sources
 - o We have used Rapid API and custom dataset.
- Rapid API rate limiting makes the pipeline slow
 - o For training, we have used the dataset dumped into mongodb
- Data Labeling
 - o Annotating correct labels to custom dataset was done manually
- Connecting MongoDB and spark
 - o Tried different ways of connecting spark to mongodb then implemented it with mongo-spark-connector
- Authorization issue in MongoDB connection
 - o Configured appropriate parameters to resolve it

News Classifier Project Report

- Loading model from HDFS
 - o Configured HDFS in spark node.
- Python-Tensorflow version compatibility
 - o Installed compatible libraries versions
- Slow prediction because of Hardware limitation
 - o Increased docker memory allocation and processor cores
- Loading Bert tokenizer in predictor service
 - o Used save and load methods of pytorch
- Responsive UI Design and Integration for mobile support
 - o Used Jinja2Templates

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>