News Data Classification

Student :
Afli Ramzi
Mhamdi Oumayma

Teacher :
Maher Heni

January 12, 2023

Contents

List of Figures

List of Tables

1 Introduction

Today, one of the most crucial aspects of our daily lives is news. People continue to concentrate on the news in order to develop a sense of the national perspective. However, because of the increasing number and variety of topics available, it can be difficult for listeners to focus on specific ones.Because of this, our main objective is to develop a real-time streamer that will be used to categorize news topics using news data. To complete this task, we attempted to implement a kind of seamless chain from the collection of the data to its visualization in the Kibana Dashboard.

2 Producer

In this section we will try to explain how we made a kafka producer that produces(streams)data from an api and store these data inside kafka broker.

The Guardian API, an API that provides real-time news, is our intended data source. The first session of our project involves writing Python scripts to stream the data and store it inside the Kafka broker. We must specify the api-url and api-key in order to retrieve the data from this API (this key is provided after creating account in the Guardian's web page) . After that, we connected our producer to the cloud-based kafka broker using a Python connector.

```
if __name__== "__main__":

    key = '53baf88d-42eb-4df9-bf3e-f30d773b1726'
    fromDate = '2013-01-1'
    toDate = '2022-08-8'

    url = 'http://content.guardianapis.com/search?from-date='+ fromDate +'&to-date='+ toDate
    all_news = getData(url)

    if len(all_news)>0:
        prod = connect_kafka_producer()
        for story in all_news:
            publish_message(prod, 'test', story)
            time.sleep(1)
        if prod is not None:
                prod.close()
```

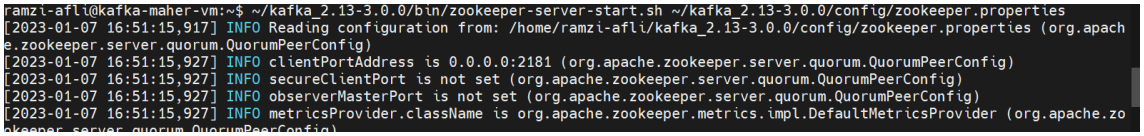
Figure 1: Producer python scripts

```
def connect_kafka_producer():
    _producer = None
    try:
        _producer = KafkaProducer(bootstrap_servers=['20.216.153.173:9092'], api_version=(0, 10
```

Figure 2: Python kafka connector

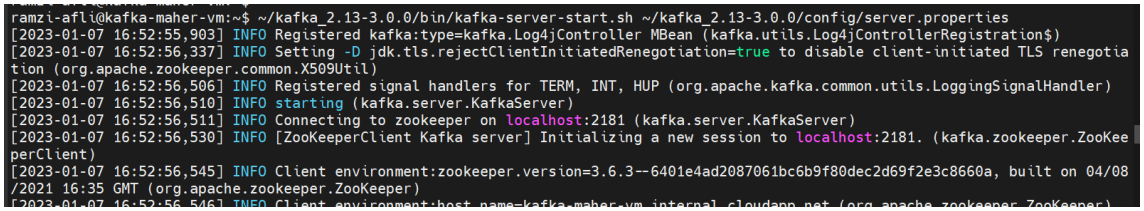
3 Kafka Broker

In order to store these news. Kafka broker will serve as good choice. Our Kafka broker is deployed on Azure cloud in a separate virtual machine.



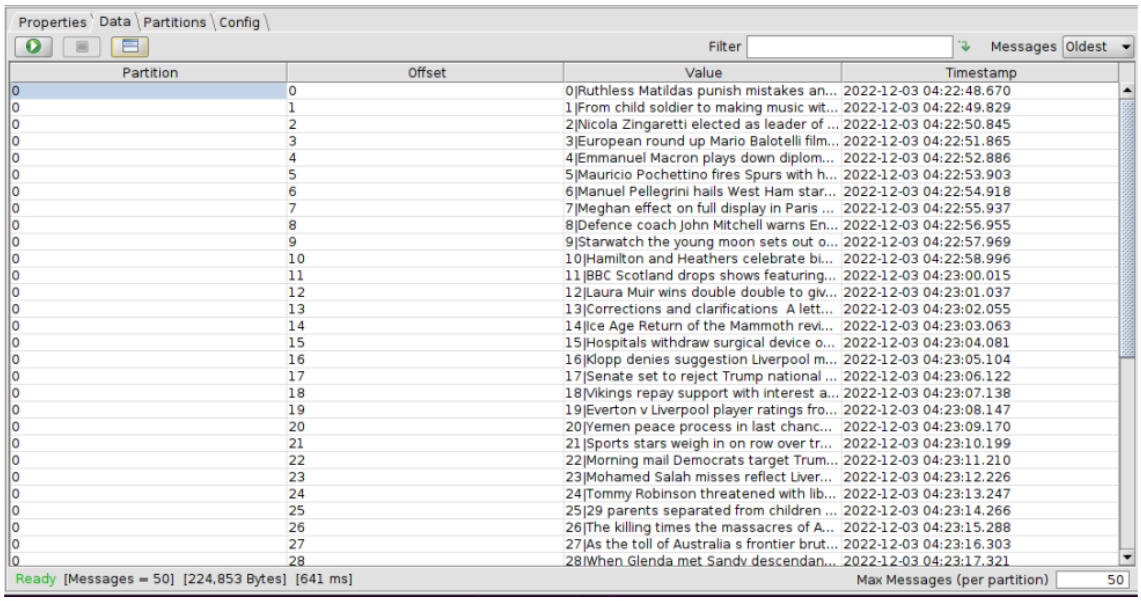
```
ramzi-afli@kafka-maher-vm:~$ ~/kafka_2.13-3.0.0/bin/zookeeper-server-start.sh ~/kafka_2.13-3.0.0/config/zookeeper.properties
[2023-01-07 16:51:15,917] INFO Reading configuration from: /home/ramzi-afli/kafka_2.13-3.0.0/config/zookeeper.properties (org.apach
e.zookeeper.server.quorum.QuorumPeerConfig)
[2023-01-07 16:51:15,927] INFO clientPortAddress is 0.0.0.0:2181 (org.apache.zookeeper.server.quorum.QuorumPeerConfig)
[2023-01-07 16:51:15,927] INFO secureClientPort is not set (org.apache.zookeeper.server.quorum.QuorumPeerConfig)
[2023-01-07 16:51:15,927] INFO observerMasterPort is not set (org.apache.zookeeper.server.quorum.QuorumPeerConfig)
[2023-01-07 16:51:15,927] INFO metricsProvider.className is org.apache.zookeeper.metrics.impl.DefaultMetricsProvider (org.apache.zo
```

Figure 3: Starting Zookeeper



```
ramzi-afli@kafka-maher-vm:~$ ~/kafka_2.13-3.0.0/bin/kafka-server-start.sh ~/kafka_2.13-3.0.0/config/server.properties
[2023-01-07 16:52:55,903] INFO Registered kafka:type=kafka.Log4jController MBean (kafka.utils.Log4jControllerRegistration$)
[2023-01-07 16:52:56,337] INFO Setting -D jdk.tls.rejectClientInitiatedRenegotiation=true to disable client-initiated TLS renegotia
tion (org.apache.zookeeper.common.X509Util)
[2023-01-07 16:52:56,506] INFO Registered signal handlers for TERM, INT, HUP (org.apache.kafka.common.utils.LoggingSignalHandler)
[2023-01-07 16:52:56,510] INFO starting (kafka.server.KafkaServer)
[2023-01-07 16:52:56,511] INFO Connecting to zookeeper on localhost:2181 (kafka.server.KafkaServer)
[2023-01-07 16:52:56,530] INFO [ZooKeeperClient Kafka server] Initializing a new session to localhost:2181. (kafka.zookeeper.ZooKee
perClient)
[2023-01-07 16:52:56,545] INFO Client environment:zookeeper.version=3.6.3--6401e4ad2087061bc6b9f80dec2d69f2e3c8660a, built on 04/08
/2021 16:35 GMT (org.apache.zookeeper.ZooKeeper)
```

Figure 4: Starting kafka



Properties | Data | Partitions | Config

Filter Messages Oldest

Partition	Offset	Value	Timestamp
0	0	0|Ruthless Matildas punish mistakes an...	2022-12-03 04:22:48.670
0	1	1|From child soldier to making music wit...	2022-12-03 04:22:49.829
0	2	2|Nicola Zingaretti elected as leader of ...	2022-12-03 04:22:50.845
0	3	3|European round up Mario Balotelli film...	2022-12-03 04:22:51.865
0	4	4|Emmanuel Macron plays down diplom...	2022-12-03 04:22:52.886
0	5	5|Mauricio Pochettino fires Spurs with h...	2022-12-03 04:22:53.903
0	6	6|Manuel Pellegrini hails West Ham star...	2022-12-03 04:22:54.918
0	7	7|Meghan effect on full display in Paris ...	2022-12-03 04:22:55.937
0	8	8|Defence coach John Mitchell warns En...	2022-12-03 04:22:56.955
0	9	9|Starwatch the young moon sets out o...	2022-12-03 04:22:57.969
0	10	10|Hamilton and Heathers celebrate bi...	2022-12-03 04:22:58.996
0	11	11|BBC Scotland drops shows featuring...	2022-12-03 04:23:00.015
0	12	12|Laura Muir wins double double to giv...	2022-12-03 04:23:01.037
0	13	13|Corrections and clarifications A lett...	2022-12-03 04:23:02.055
0	14	14|Ice Age Return of the Mammoth revi...	2022-12-03 04:23:03.063
0	15	15|Hospitals withdraw surgical device o...	2022-12-03 04:23:04.081
0	16	16|Klopp denies suggestion Liverpool m...	2022-12-03 04:23:05.104
0	17	17|Senate set to reject Trump national ...	2022-12-03 04:23:06.122
0	18	18|Vikings repay support with interest a...	2022-12-03 04:23:07.138
0	19	19|Everton v Liverpool player ratings fro...	2022-12-03 04:23:08.147
0	20	20|Yemen peace process in last chanc...	2022-12-03 04:23:09.170
0	21	21|Sports stars weigh in on row over tr...	2022-12-03 04:23:10.199
0	22	22|Morning mail Democrats target Trum...	2022-12-03 04:23:11.210
0	23	23|Mohamed Salah misses reflect Liver...	2022-12-03 04:23:12.226
0	24	24|Tommy Robinson threatened with lib...	2022-12-03 04:23:13.247
0	25	25|29 parents separated from children ...	2022-12-03 04:23:14.266
0	26	26|The killing times the massacres of A...	2022-12-03 04:23:15.288
0	27	27|As the toll of Australia s frontier brut...	2022-12-03 04:23:16.303
0	28	28|When Glenda met Sandy descendan...	2022-12-03 04:23:17.321

Ready [Messages = 50] [224,853 Bytes] [641 ms] Max Messages (per partition) 50

Figure 5: Kafka Ui

4 Consumer

We will now develop the consumer after first creating the producer that streams data to the Kafka broker. The principal function of the consumer is to retrieve data from the Kafka broker.

after importing all needed libraries , we call the consumer .

```
consumer = KafkaConsumer('test',
                         bootstrap_servers=['20.216.153.173:9092'])
```

```
Command took 0.15 seconds -- by oumayma.mhamdi@studentambassadors.com at 07/01/2023 23:38:18 on Oumayma Mhamdi's
Cluster
```

Figure 6: Calling the consumer

the consumer retrieve data from kafka broker . We can see data received by the consumer in the figure 7 .

```
for text in consumer:
    print(text.value)
```

```
b'29|Plan to axe 91 000 civil servants only possible with cuts to services  Doubts have been cast o
n claims by Boris Johnson that it will be possible to go ahead with plans to axe 91 000 civil serva
nts without harming frontline services The prime minister wrote in May to civil servants justifying
plans for a reduction in headcount of almost 20 saying the government must reduce its costs just as
many families are doing But a review by Steve Barclay Johnson s former chief of staff is reported t
o have caused the Treasury to have second thoughts about the plans given the potential impact on wi
der services Johnson had tasked the cabinet with cutting staff by a fifth telling ministers during
an away day in Stoke on Trent earlier this year that every bit of cash saved on government spending
could be better used elsewhere Trade unions representing civil servants seized on reports about Tre
asury reticence while Labour said it was clear that the disastrous plan would cause historic staff
shortages causing huge delays for people when trying to access appointments and renew passports and
driving licences Now the government has not only belatedly come to the same conclusion but also fou
nd this shoddy proposal would have cost the taxpayer over 1bn in redundancy payments said Rachel Ho
pkins MP Labour s shadow minister in the Cabinet Office The plans have been backed by Liz Truss the
Conservative leadership frontrunner who has been accused of making ludicrous claims as she vowed to
cut civil service salaries and reduce expenditure to recoup 11bn a year in a war on Whitehall waste
However a Whitehall insider who was said to have worked on the plans to axe tens of thousands of ci
vil servants was quoted in the Financial Times as saying that the prime minister had announced the
move without fully thinking through the implications You can only deliver 91 000 cuts by actual cut
s to major frontline services they added There s no way you can get to that number through efficien
cy savings or reductions in HQ staff Another Whitehall source was quoted as suggesting that a figu
Cancelled
```

Figure 7: Receiving Data by the consumer

5 ML Algorithm

To classify News into several categories , we use the Naive Bayes algorithm .
we choose Naive Bayes for many reasons :

- It doesn't require as much training data.
- It handles both continuous and discrete data.
- It is highly scalable with the number of predictors and data points.
- It is fast and can be used to make real-time predictions

To construct the model we firstly import required libraries

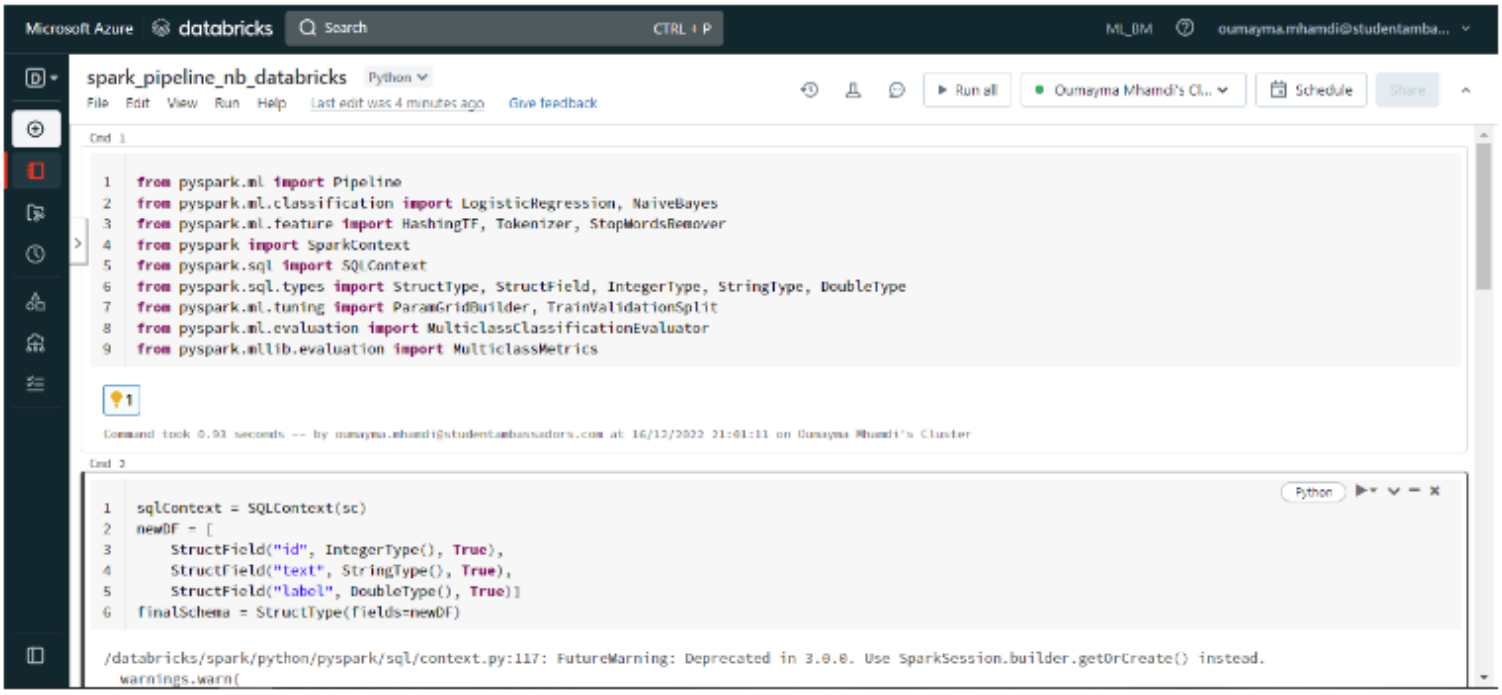


Figure 8: Libraries import

Then we Load the data and split it into training and testing datasets .
The training dataset presents 80 '%óf the data .
The testing dataset presents 20 '%óf the data . After that we applicate the tokenizer then
The stopwords remover and finnally the hashing TF

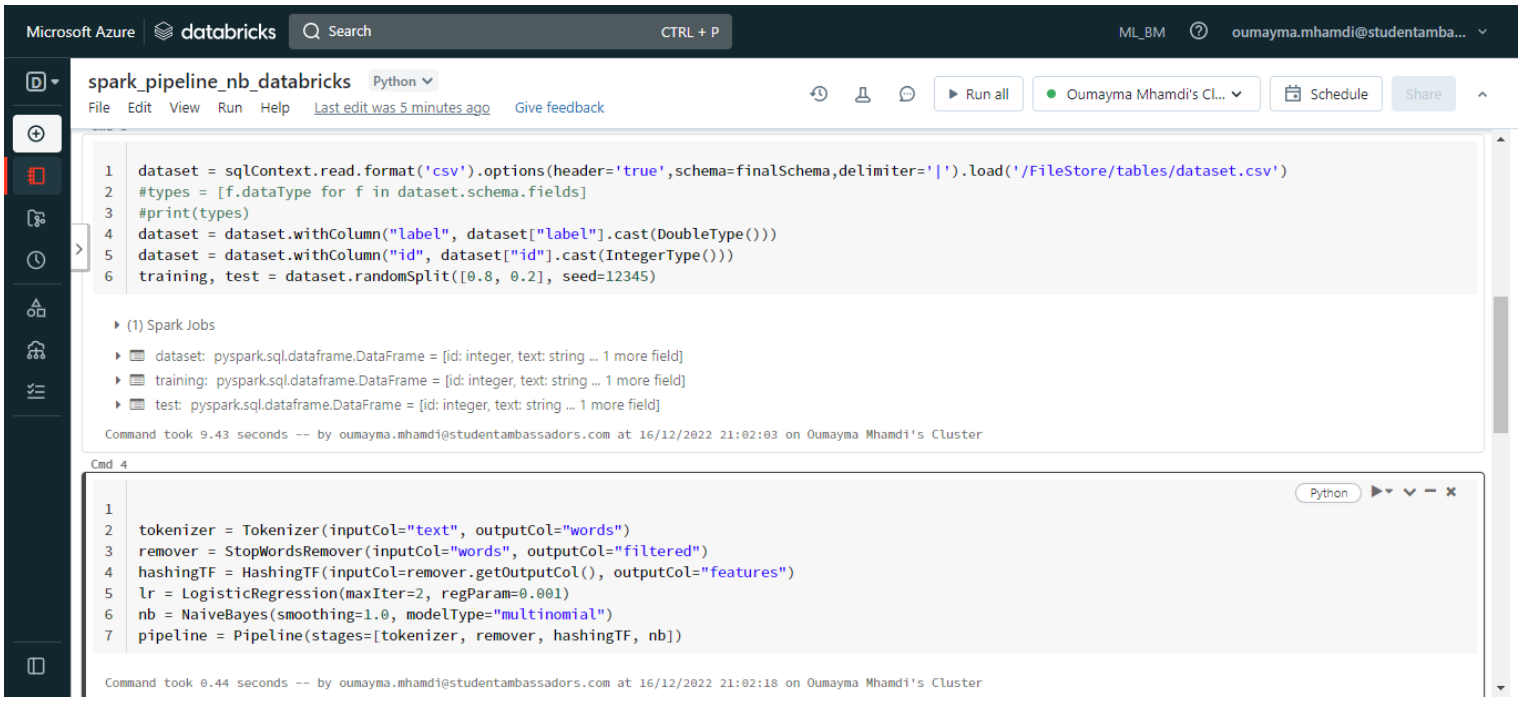


Figure 9: Creating pipeline

The next step is to train the model so that it can be able to classify data correctly .

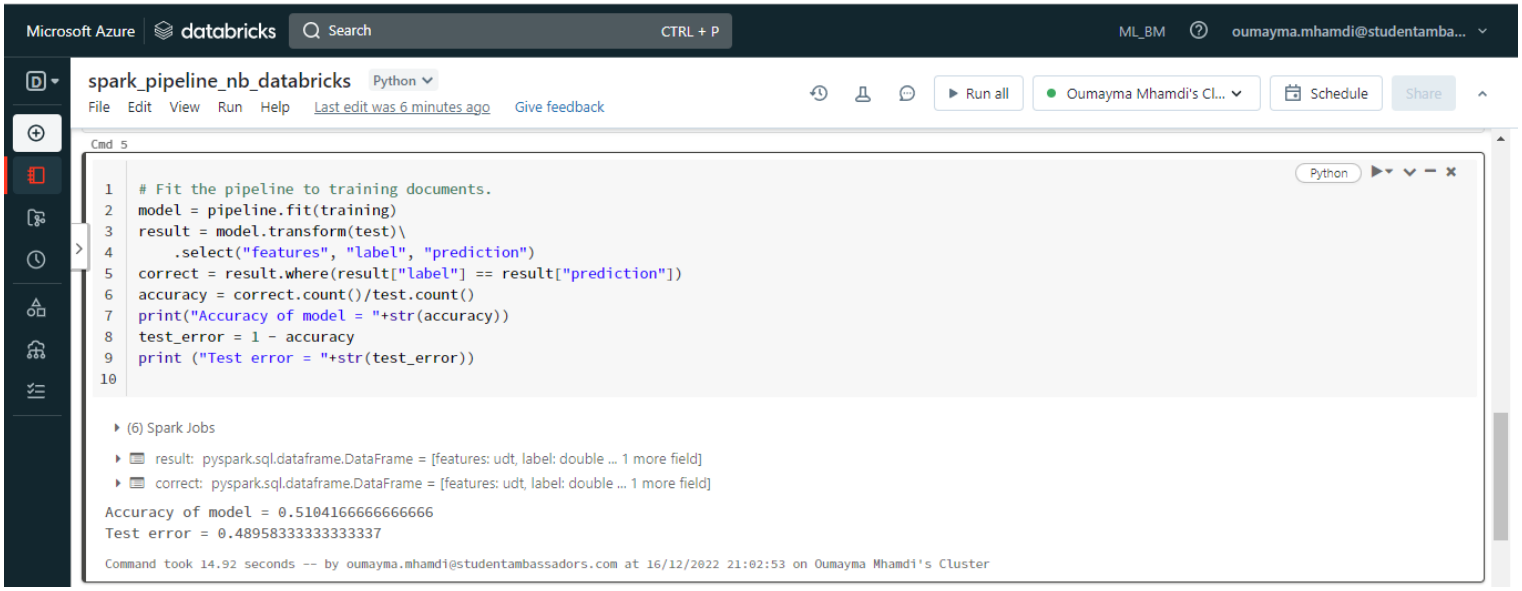


Figure 10: Fitting the model

The final step is to evaluate the performance of the classifier . the metrics of performance used are :

- The F1 Score
- The Recall
- The Precision

We saved the pretrained model to applicate it on testing data .

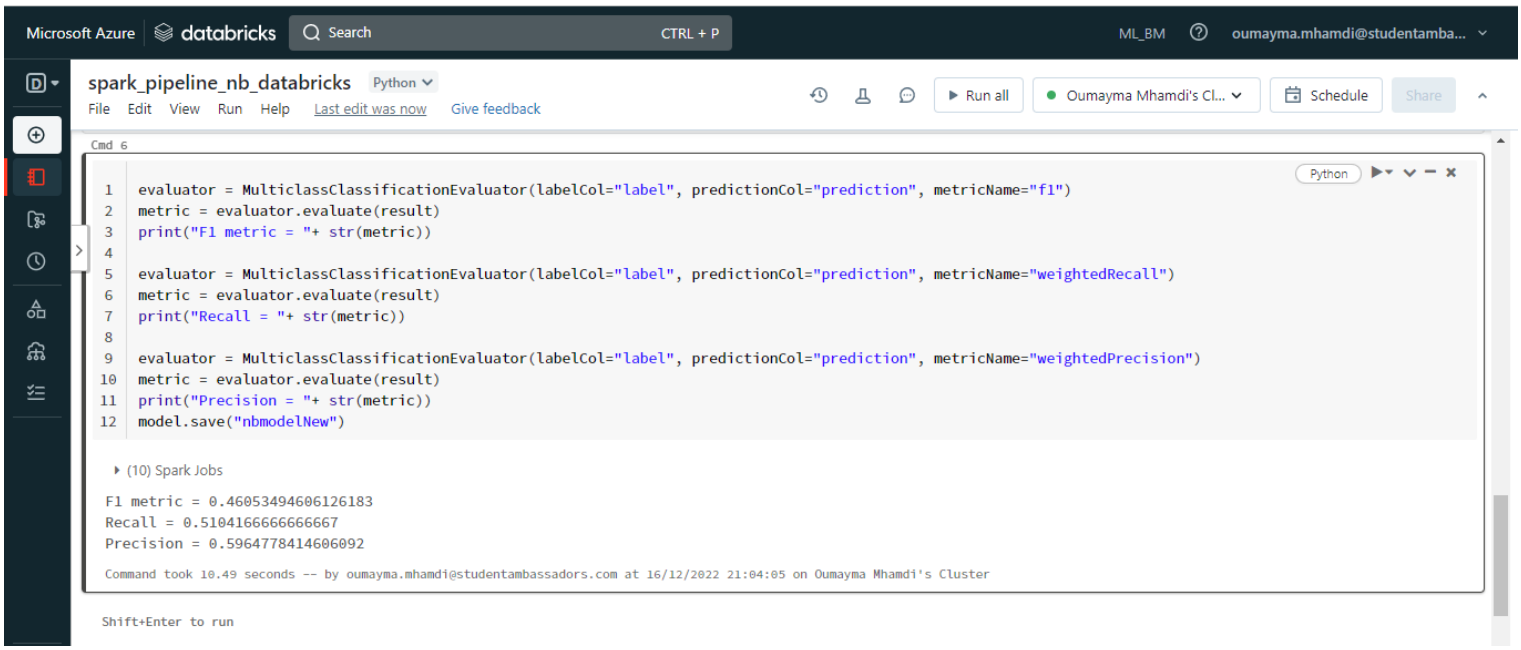


Figure 11: Performance of the model

6 Elastic search & Kibana

In order to index our Data and visualize them .ELK stack is mandatory that's why we were Deploying ELK stack as docker containers using docker compose in the cloud .

```
ramzi-afli@kafka-maher-vm:~/Elk-stack-docker$ vim  docker-compose.yml
version: '3.7'

services:

  # Author Afli Ramzi
  # The 'setup' service runs a one-off script which initializes the
  # 'logstash_internal' and 'kibana_system' users inside Elasticsearch with the
  # values of the passwords defined in the '.env' file.
  #
  # This task is only performed during the *initial* startup of the stack. On all
  # subsequent runs, the service simply returns immediately, without performing
  # any modification to existing users.
  setup:
    build:
      context: setup/
      args:
        ELASTIC_VERSION: ${ELASTIC_VERSION}
    init: true
    volumes:
      - setup:/state:Z
    environment:
      ELASTIC_PASSWORD: ${ELASTIC_PASSWORD:-}
      LOGSTASH_INTERNAL_PASSWORD: ${LOGSTASH_INTERNAL_PASSWORD:-}
      KIBANA_SYSTEM_PASSWORD: ${KIBANA_SYSTEM_PASSWORD:-}
    networks:
      - elk
    depends_on:
      - elasticsearch

  elasticsearch:
    build:
```

Figure 12: Docker compose elasticseach part

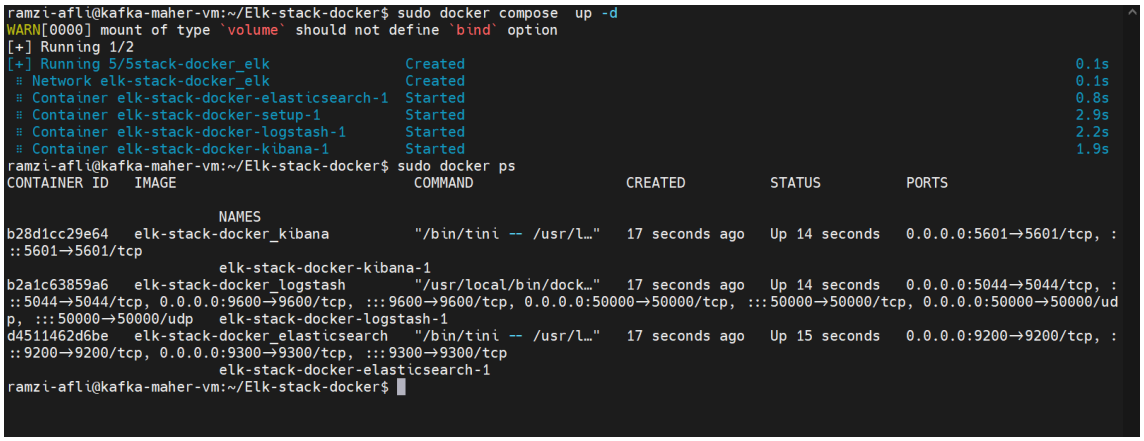
```
  kibana:
    build:
      context: kibana/
      args:
        ELASTIC_VERSION: ${ELASTIC_VERSION}
    volumes:
      - ./kibana/config/kibana.yml:/usr/share/kibana/config/kibana.yml:ro,Z
    ports:
      - "5601:5601"
    environment:
      KIBANA_SYSTEM_PASSWORD: ${KIBANA_SYSTEM_PASSWORD:-}
    networks:
      - elk
    depends_on:
      - elasticsearch

networks:
  elk:
    driver: bridge

volumes:
  setup:
  elasticsearch:
```

Figure 13: Docker compose kibana part

After that we will try to start these containers using docker commands .

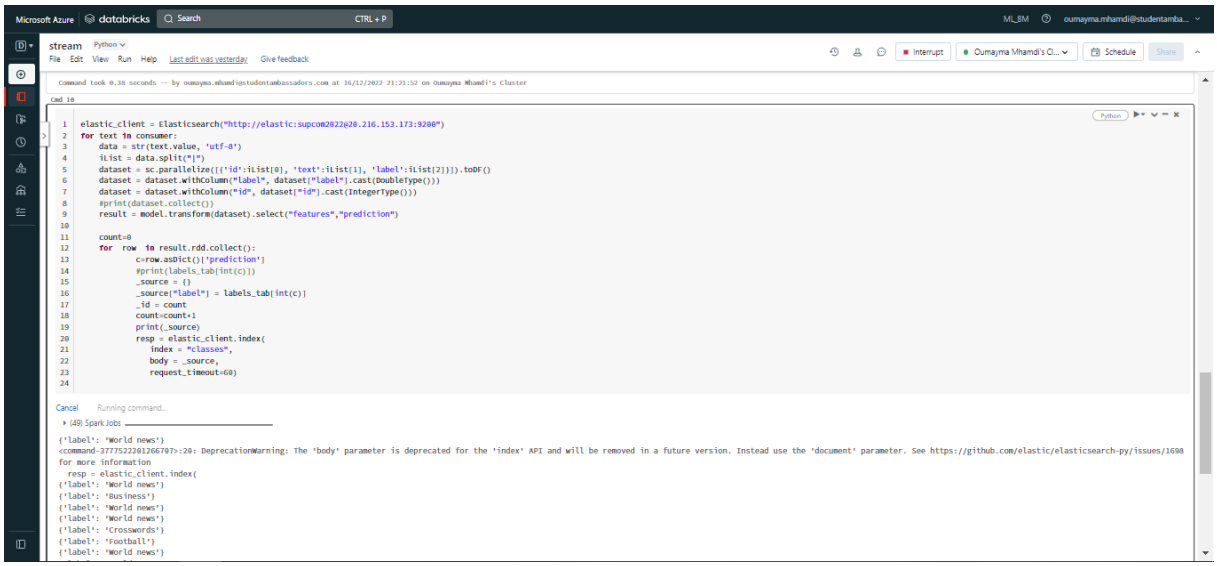


```
ramzi-afli@kafka-maher-vm:~/Elk-stack-docker$ sudo docker compose  up -d
WARN[0000] mount of type `volume` should not define `bind` option
[+] Running 1/2
[+] Running 5/5stack-docker_elk                Created                                                                    0.1s
 ⠿ Network elk-stack-docker_elk                Created                                                                    0.1s
 ⠿ Container elk-stack-docker-elasticsearch-1  Started                                                                    0.8s
 ⠿ Container elk-stack-docker-setup-1          Started                                                                    2.9s
 ⠿ Container elk-stack-docker-logstash-1       Started                                                                    2.2s
 ⠿ Container elk-stack-docker-kibana-1         Started                                                                    1.9s
ramzi-afli@kafka-maher-vm:~/Elk-stack-docker$ sudo docker ps
CONTAINER ID   IMAGE                                COMMAND                  CREATED          STATUS          PORTS

                         NAMES
b28d1cc29e64   elk-stack-docker_kibana              "/bin/tini -- /usr/l…"   17 seconds ago   Up 14 seconds   0.0.0.0:5601→5601/tcp, :
::5601→5601/tcp
                         elk-stack-docker-kibana-1
b2a1c63859a6   elk-stack-docker_logstash            "/usr/local/bin/dock…"   17 seconds ago   Up 14 seconds   0.0.0.0:5044→5044/tcp, :
::5044→5044/tcp, 0.0.0.0:9600→9600/tcp, :::9600→9600/tcp, 0.0.0.0:50000→50000/tcp, :::50000→50000/tcp, 0.0.0.0:50000→50000/ud
p, :::50000→50000/udp   elk-stack-docker-logstash-1
d4511462d6be   elk-stack-docker_elasticsearch       "/bin/tini -- /usr/l…"   17 seconds ago   Up 15 seconds   0.0.0.0:9200→9200/tcp, :
::9200→9200/tcp, 0.0.0.0:9300→9300/tcp, :::9300→9300/tcp
                         elk-stack-docker-elasticsearch-1
ramzi-afli@kafka-maher-vm:~/Elk-stack-docker$
```

Figure 14: Starting docker containers

After receiving news data , In the same notebook as the consumer we applicate the ml algorithm to them then we send them to elastic search



```
elastic_client = Elasticsearch("http://elastic:supcom2022@20.216.153.173:9200")
for text in consumer:
    data = str(text.value, 'utf-8')
    iList = data.split("|")
    dataset = sc.parallelize([{'id':iList[0], 'text':iList[1], 'label':iList[2]}]).toDF()
    dataset = dataset.withColumn("label", dataset["label"].cast(DoubleType()))
    dataset = dataset.withColumn("id", dataset["id"].cast(IntegerType()))
    #print(dataset.collect())
    result = model.transform(dataset).select("features","prediction")

    count=0
    for row in result.rdd.collect():
        c=row.asDict()['prediction']
        #print(labels_tab[int(c)])
        _source = {}
        _source["label"] = labels_tab[int(c)]
        _id = count
        count=count+1
        print(_source)
        resp = elastic_client.index(
            index = "classes",
            body = _source,
            request_timeout=60)
```

```
{'label': 'World news'}
<command-3777522201266707>:20: DeprecationWarning: The 'body' parameter is deprecated for the 'index' API and will be removed in a future version. Instead use the 'document' parameter. See https://github.com/elastic/elasticsearch-py/issues/1698
for more information
  resp = elastic_client.index(
{'label': 'World news'}
{'label': 'Business'}
{'label': 'World news'}
{'label': 'World news'}
{'label': 'Crosswords'}
{'label': 'Football'}
{'label': 'World news'}
```

Figure 15: Indexing data

At the end, we use Elastic Search and Kibana for visualization of our results. Elastic Search is used for storing results into indexes and these indexes are used by kibana for various visualization. Results are given by figure 16 and figure 17

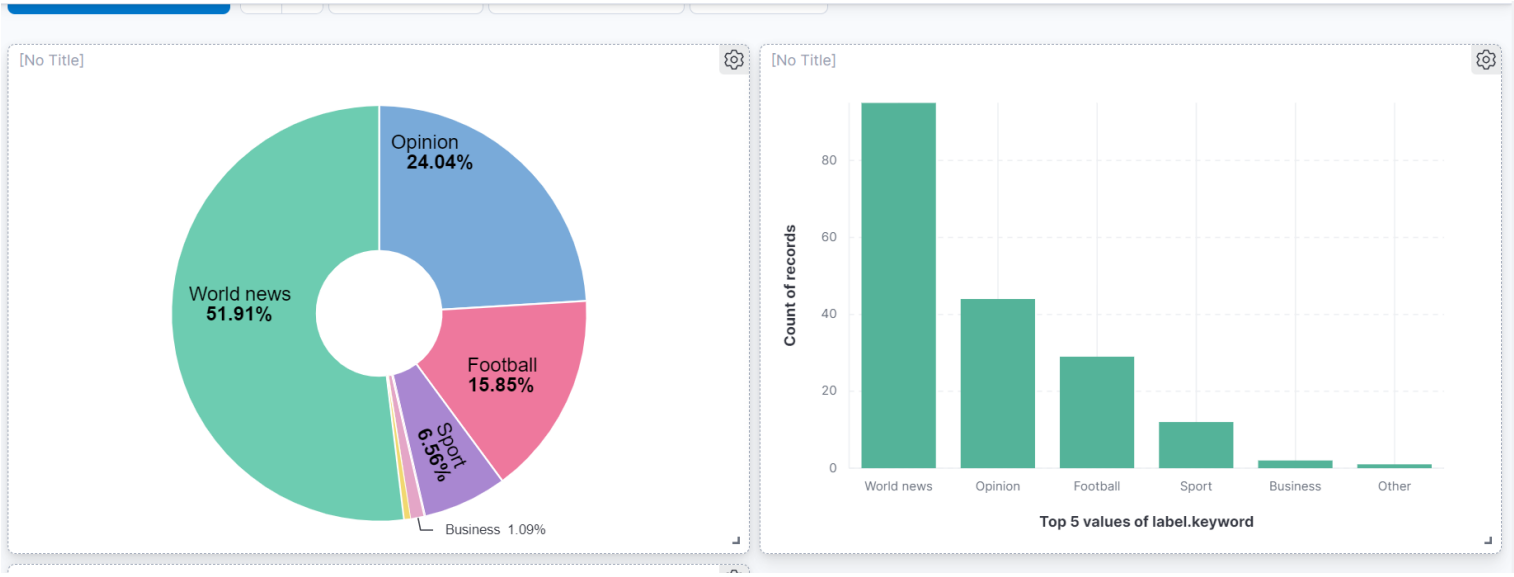


Figure 16: Kibana visualisations

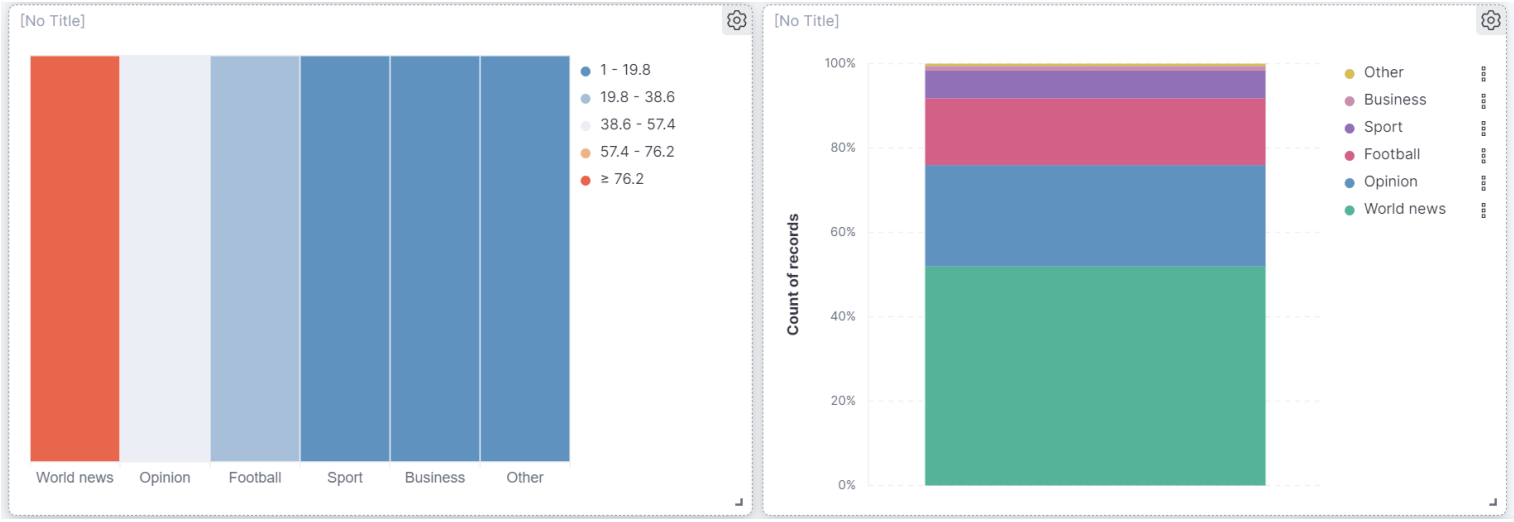


Figure 17: Kibana visualisations

==> These visualisations show the percentage of each news category based on the result given by the ML algorithm .

7 Conclusion

This project, named "News Data Classification", was conducted to create full intelligent chain of streaming , used to stream news in real time ,classify these data using machine learning algorithm , and indexing the news topics inside elastic search and visualize the most common using kibana dashboard .