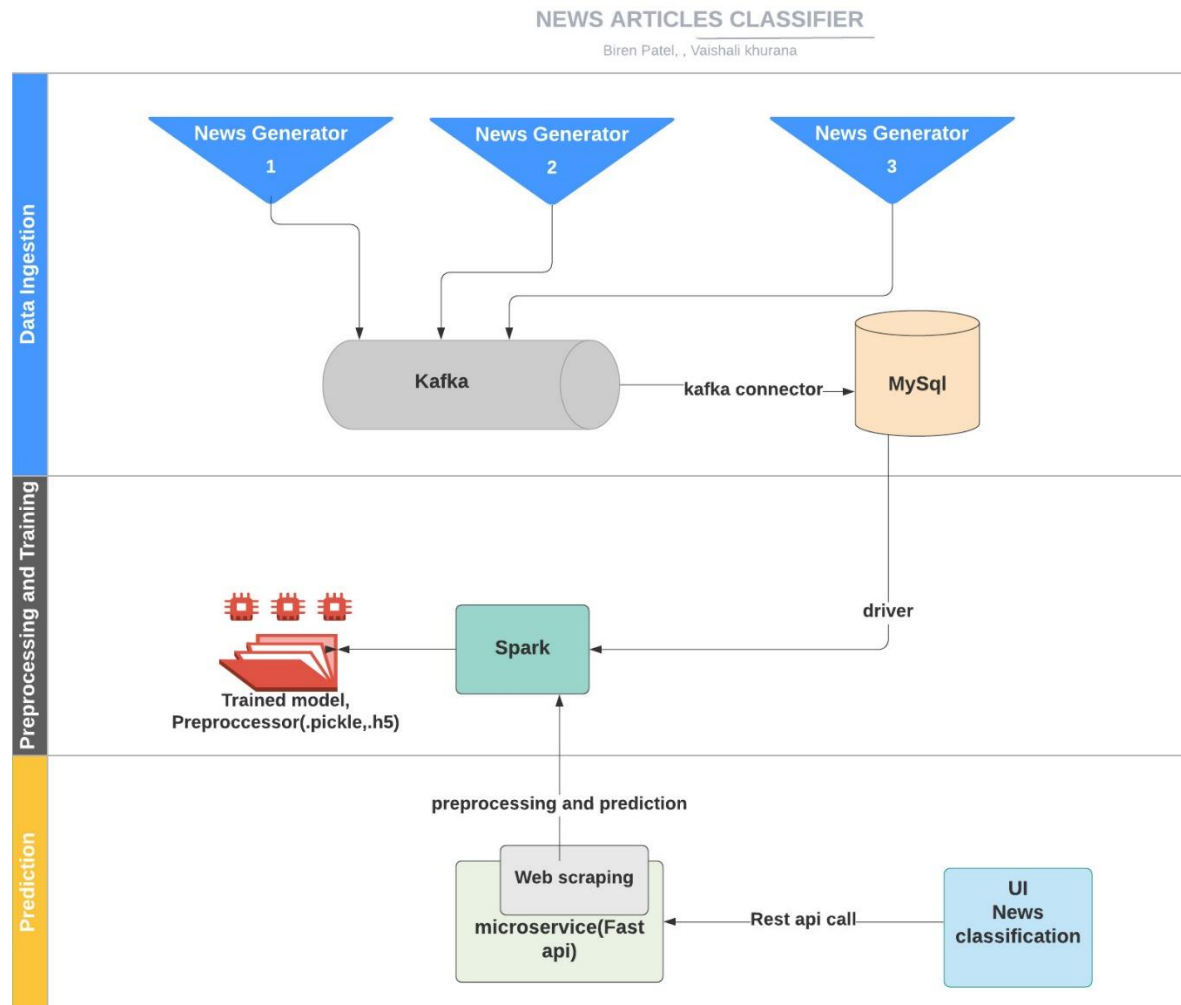


[illegible]

News Classifier Project Report

Architecture:



Note: We have used mysql considering the future scope of this application, Please read the future scope at the end of the document

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 jdbc connector is used to sync the data between kafka and database. Finally, data is dumped into Mysql database
Preprocessor and trainer	Distributed service to preprocess the data stored in database and train ML model using spark

News Classifier Project Report

Prediction	Scraping the weblink provided by user in UI, clean it and predict the news using rest api.
------------	--

Since data ingestion is the first milestone we have explained it in detail here.

Data Ingestion:

- Environment details:**

Docker environment with kafka broker running on 9092 port, zookeeper running on 2181, kafka connect running on 8083, mysql running 3306

- What goes in as an input:**

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

```
(env) PS C:\Users\u1105800\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_5b081ab4e4b882d69caad89'}
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\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_5b09726fe4b9fdb2aa541201'}
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 jdbc connector is used to sync the data between kafka and database. Finally, data is dumped into Mysql database.

```
ksql> SHOW TOPICS;

Kafka Topic              | Partitions | Partition Replicas
-----
confluent_rmqoff_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_5b881ab4e4b8802d69caad89\", \"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_5b99726fe4b0fd62aa541201\", \"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:** Data stored in mysql database

```
mysql> select * from news limit 2 offset 1;
+-----+-----+-----+-----+-----+
| TITLE                                     | SUMMARY                                     |
+-----+-----+-----+-----+-----+
| Ireland Votes To Repeal Abortion Amendment In Landslide Referendum | Irish women will no longer have to travel to the United Kingdom to end their pregnancies |
| politics | huffingtonpost | 10-09-2021 |
+-----+-----+-----+-----+-----+
| 8 Majestic Islands In Europe That Most Tourists Dont Know About | If you are dreaming about a romantic European getaway that doesnt involve a gazillion tourists, then consider these beautiful isles |
| travel | huffingtonpost | 10-10-2021 |
+-----+-----+-----+-----+-----+
2 rows in set (0.00 sec)
```

- **Tools/libraries used:** Docker, Pycharm, kafka, Mysql, zookeeper, Spark, kafka

Challenges encountered:

We faced below mentioned challenges. However, we have resolved them.

- Finding better legal data sources
- Api rate limiting makes the pipeline slow
- Character encoding
- Data Labeling
- Connecting kafka to data store

News Classifier Project Report

- Network configurations for kafka connectors

Future Scope:

- We will be adding more features to build end to end news browsing application like
 - o bookmarking the news
 - o subscribing to specific news
 - o news-recommendations
 - o notifications etc.
- 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

Github link:

<https://github.com/biren162/Capstone>