Fake News Detection

Mythbusters

Outline

1. Introduction
2. System Architecture
3. Data Source
4. Machine Learning Implementation
5. Result Evaluation
6. Application Design

Introduction

- Wide spreading fake news on social media
- Heavy social and national impact
- Solution: intelligent platform for detecting fake news

System Architecture

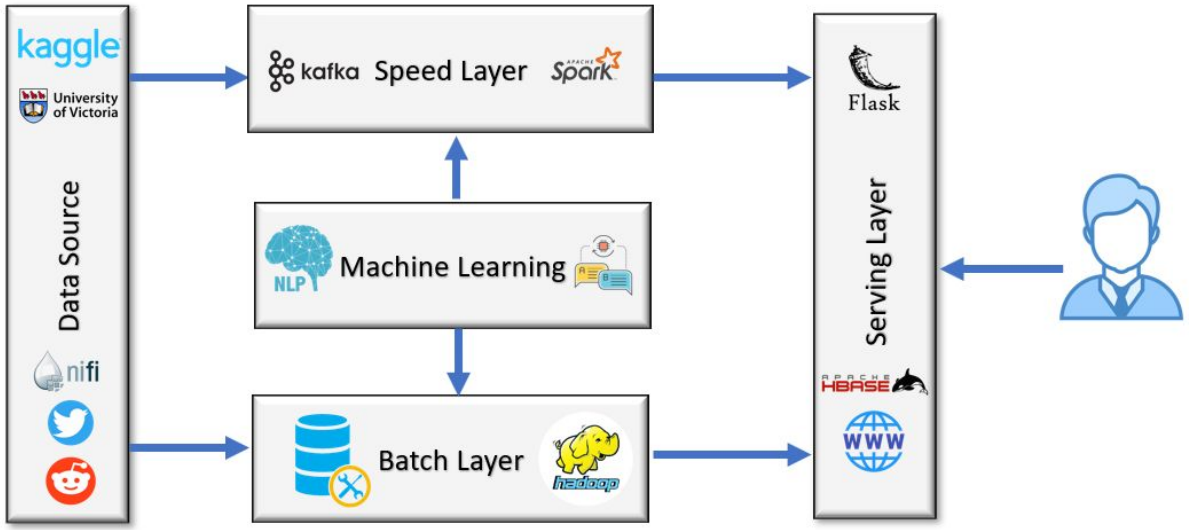


Figure 1: Lambda Architecture

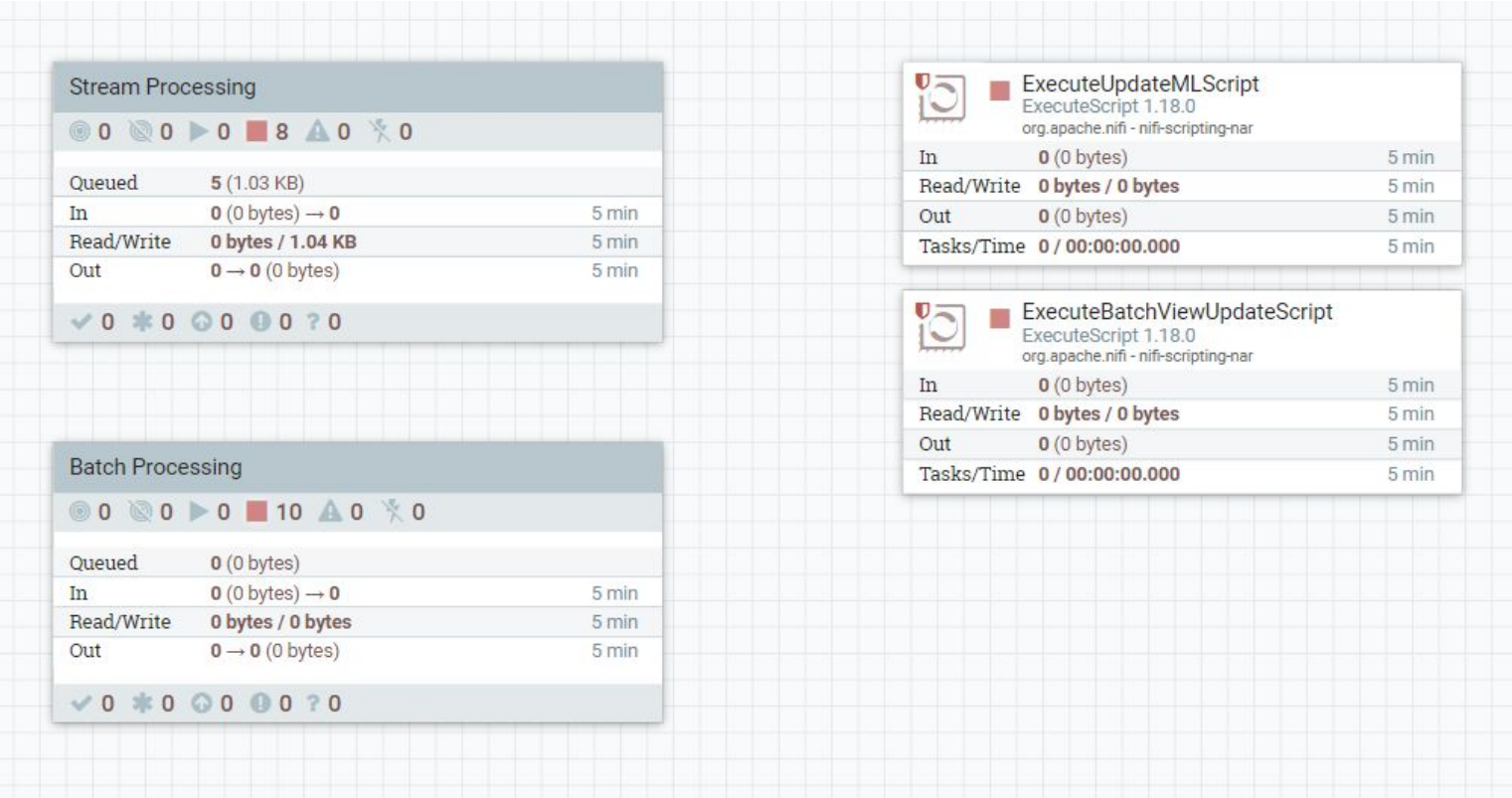
Data sources

- Labeled batch data
  - Fakeddit ➜ 1M reddit posts
  - LIAR ➜ 12,8K social media statements
  - ISOT ➜ 40K articles

artificially streamed to a daily frequency

- Streaming data
  - Twitter
  - Reddit

Data Ingestion

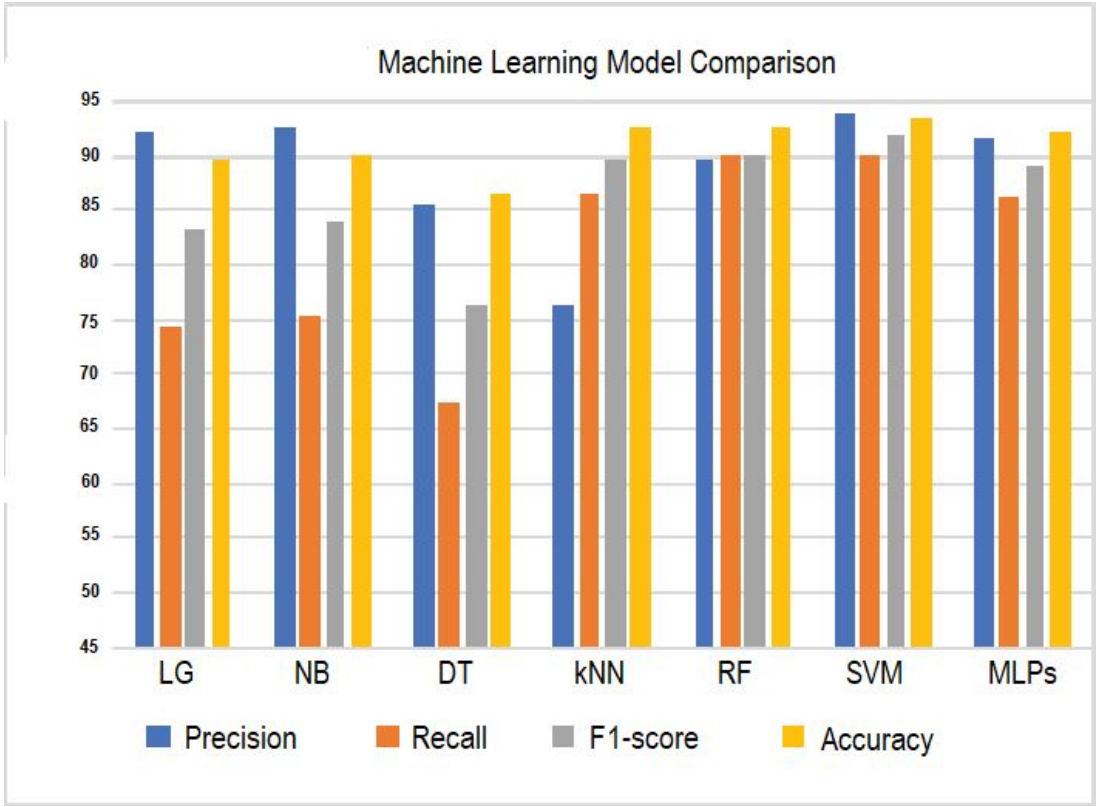


Machine Learning Implementation

Supervised Classification Techniques

1. Logistic Regression
2. Naive Bayes
3. Decision Tree
4. Random Forest
5. k-Nearest Neighbors
6. Support Vector Machine
7. Multilayer Perceptron

Result Evaluation



ML model updating

- Batch layer incremented on a daily basis
- Model re-trained at each update
- The versions of the model are stored on the server

Serving Layer

- Queries: individual statement or historical data
- Predict a statement directly on the model
- Aggregation of historical statistics stored in a Hive table

Application Design

Historical Data

```
SELECT 100*SUM(last_pred.prediction)/COUNT(*)
FROM (
SELECT prediction
FROM batch_view
WHERE (UNIX_TIMESTAMP(current_timestamp()) - UNIX_TIMESTAMP(datetimestamp) < 3600)
) AS last_pred;
```

```
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Job running in-process (local Hadoop)
2023-01-23 15:22:53,706 Stage-1 map = 0%,  reduce = 0%
2023-01-23 15:22:56,747 Stage-1 map = 100%,  reduce = 100%
Ended Job = job_local1314716597_0006
MapReduce Jobs Launched:
Stage-Stage-1:  HDFS Read: 0 HDFS Write: 0 SUCCESS
Total MapReduce CPU Time Spent: 0 msec
OK
49.63503649635037
Time taken: 4.334 seconds, Fetched: 1 row(s)
hive (mythdb)>
```

Ethical concerns

- Errors of classification:
    - Take part in propagating fake news
    - Impair the reputation or a journalist/celebrity/newspaper
- Inability to detect comedy or fiction
- Possibility of unequal treatment

Conclusion

- Implementation of a fake news detecting platform
- Lambda architecture to deal with big data
- Able to check veracity of a statement or analyse historical data
- Lots of ethical issues to deal with

Business perspectives

- Increase the volume of labeled data via web scraping
- Get access to more channels of streaming data
- Put the platform on a cluster
- Implement a more user-friendly and detailed front-end

Questions