EMT 678 Big Data Technologies

Objective : Classify news article based on its sentiments

Dataset : GDELT 2.0
SIZE : 2TB
Record Count: 24M

By **Sunderamurthy Yuvaraj**

Prediction Col - QuadClass 3M Dataset

1=Verbal Cooperation

2=Material Cooperation

3=Verbal Conflict

4=Material Conflict

QuadClass	count
1	1805710
3	361233
4	392632
2	353933

DATA ACQUISITION - BIG QUERY

```
SELECT C.GLOBALEVENTID,A.GKGRECORDID, C.FractionDate, C.Actor1Code,
C.Actor1CountryCode, C.Actor2Code, C.Actor2CountryCode, C.IsRootEvent, C.GoldsteinScale,
C.AvgTone, B.Actor1CharOffset, B.Actor2CharOffset, B.ActionCharOffset, B.Confidence,
B.MentionDocTone, A.V2Organizations, A.V2Themes, A.V2Tone,
A.DocumentIdentifier,C.QuadClass

FROM `gdelt-bq.gdeltv2.gkg` A, `gdelt-bq.gdeltv2.eventmentions` B, `gdelt-bq.gdeltv2.events` C

where B.Confidence=100 and A.DocumentIdentifier = B.MentionIdentifier and
B.GLOBALEVENTID = C.GLOBALEVENTID and C.FractionDate>=2023.0000 and
C.FractionDate<=2023.2000;
```

SCHEMA OF TABLE

```
root
 |-- GLOBALEVENTID: string (nullable = true)
 |-- GKGRECORDID: string (nullable = true)
 |-- Actor1Code: string (nullable = true)
 |-- Actor1CountryCode: string (nullable = true)
 |-- Actor2Code: string (nullable = true)
 |-- Actor2CountryCode: string (nullable = true)
 |-- GoldsteinScale: string (nullable = true)
 |-- AvgTone: string (nullable = true)
 |-- Actor1CharOffset: string (nullable = true)
 |-- Actor2CharOffset: string (nullable = true)
 |-- ActionCharOffset: string (nullable = true)
 |-- V2Organizations: string (nullable = true)
 |-- V2Themes: string (nullable = true)
 |-- V2Tone: string (nullable = true)
 |-- QuadClass: string (nullable = true)
```

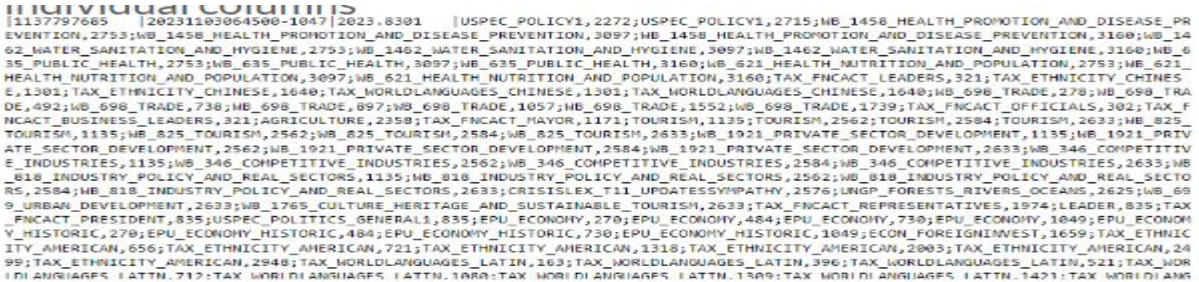
```
root
 |-- GLOBALEVENTID: string (nullable = true)
 |-- GKGRECORDID: string (nullable = true)
 |-- GoldsteinScale: integer (nullable = true)
 |-- AvgTone: integer (nullable = true)
 |-- QuadClass: integer (nullable = true)
 |-- Actor1CharOffset_I: integer (nullable = true)
 |-- Actor2CharOffset_I: integer (nullable = true)
 |-- ActionCharOffset_I: integer (nullable = true)
 |-- ArticleTone: integer (nullable = true)
 |-- PositiveScore: integer (nullable = true)
 |-- NegativeScore: integer (nullable = true)
 |-- Polarity: integer (nullable = true)
 |-- Actor1Code_SI: double (nullable = true)
 |-- Actor2Code_SI: double (nullable = true)
 |-- Actor1CountryCode_SI: double (nullable = true)
 |-- Actor2CountryCode_SI: double (nullable = true)
 |-- Theme_0_SI: double (nullable = true)
 |-- Theme_1_SI: double (nullable = true)
 |-- Theme_2_SI: double (nullable = true)
 |-- Theme_3_SI: double (nullable = true)
 |-- Theme_4_SI: double (nullable = true)
```

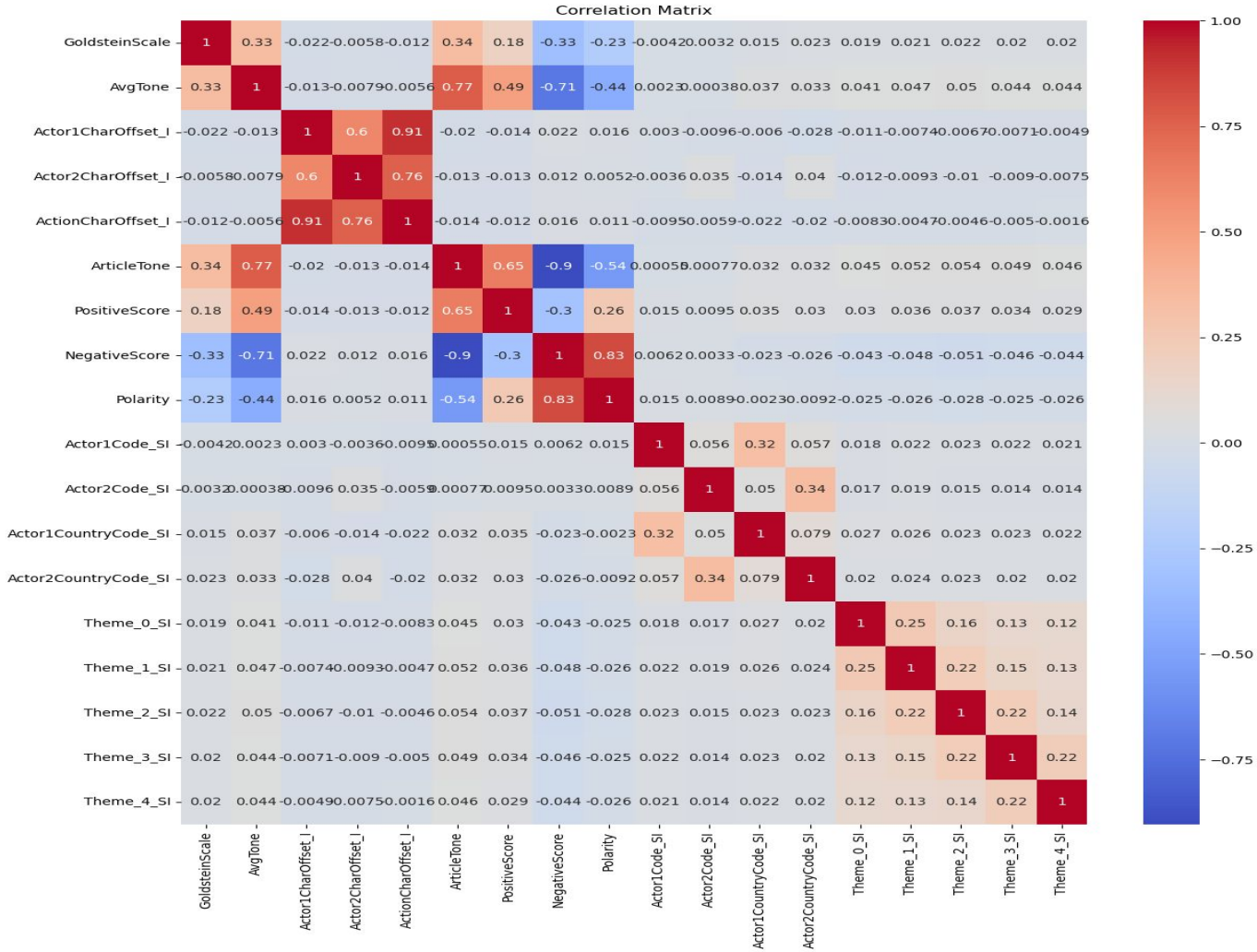
Row Size = 3M

PK= (GLOBAL,GKG)ID

DATA PREPROCESSING

- Removing Duplicate Articles(unique (GLOBALEVENTID,GKGRECORDID))
- Handling Null Values
- Conversion of String to Integers
- Theme Analysis by Splitting themes sorting top 5 themes for each article
  - UDF
- Assigning Integers to Categorical Data using StringIndexer
- Dropping values with inconsistent data




Correlation Matrix
GoldsteinScale
AvgTone
Actor1CharOffset_I
Actor2CharOffset_I
ActionCharOffset_I
ArticleTone
PositiveScore
NegativeScore
Polarity
Actor1Code_SI
Actor2Code_SI
Actor1CountryCode_SI
Actor2CountryCode_SI
Theme_0_SI
Theme_1_SI
Theme_2_SI
Theme_3_SI
Theme_4_SI
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75

Model - Logistic Regression

Model Flow

1. Features to Vectors
2. Normalise Continuous Columns
3. LR Model

Hyper Tuning Parameter using ElasticNetParam [0.2,0.5]

Strategy - Train Validation Split

LOGISTIC REGRESSION MODEL - PERFORMANCE

Precision	
label 1:	**0.8786058076**
label 2:	**0.7547188457**
label 3:	**0.6349439939**
label 4:	**0.8548670346**
Recall	
label 1:	**0.9901311331**
label 2:	**0.1071634042**
label 3:	**0.8303404544**
label 4:	**0.7791440658**
Accuracy	**0.8341150553**

ElasticNetParam
Values = [0.2,0.5]

Confusion Matrix

QuadClass	1.0	2.0	3.0	4.0
1	357973	2482	1085	1
3	1842	null	59973	10412
4	null	null	17386	61335
2	47618	7637	16010	null

Scaling Strategies

1 Master 5 Workers

1 Master 4 Workers

1 Master 3 Workers

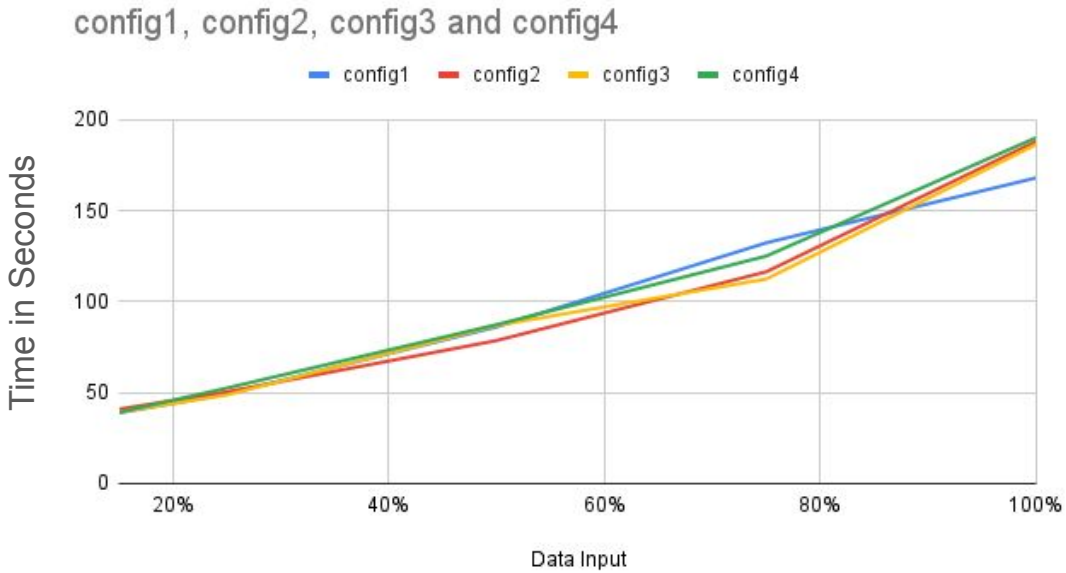
1 Master 2 Workers

Master
2vCPU
8GB

WORKER
2vCPU
5GB

Scaling Results -3M

Data Input	config1	config2	config3	config4
15%	38.82	40.84	39.19	39.12
25%	49.18	50.41	48.66	52.46
50%	86.1	78.61	86.89	87.36
75%	132.41	116.47	112.48	125.11
100%	168.12	188.23	186.48	190.24



Config1 1M 5N

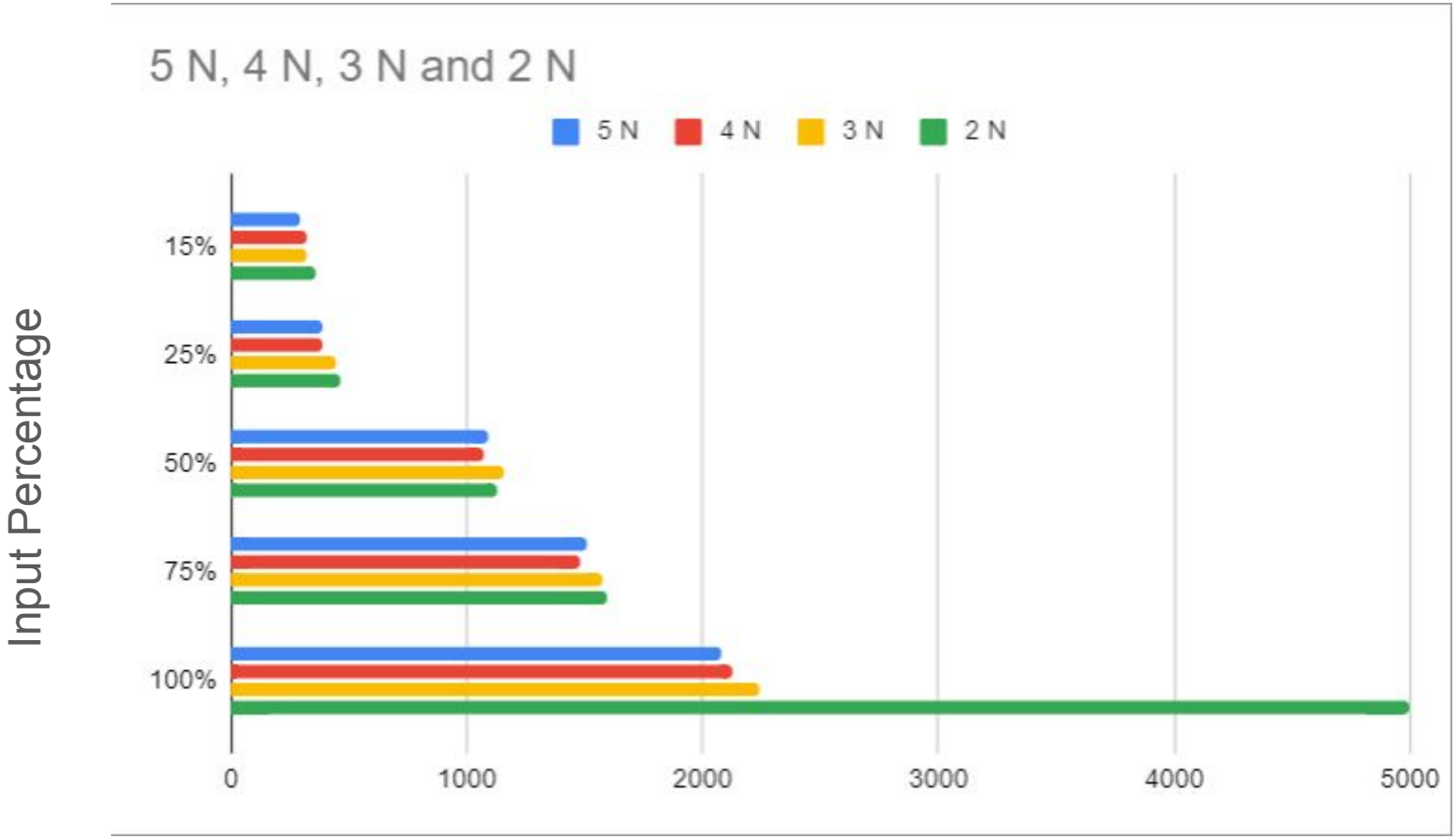
Config2 1M 4N

Config3 1M 3N

Config4 1M 2N

Scaling Results for 24M

	5 N	4 N	3 N	2 N
15%	296.69	317.61	324.45	359.16
25%	387.98	386.42	440.72	461.38
50%	1091.28	1075.7	1157.13	1130.16
75%	1503.8	1476.21	1578.38	1597.47
100%	2074.53	2126.29	2243.19	5000


5 N, 4 N, 3 N and 2 N
5 N
4 N
3 N
2 N
Input Percentage
15%
25%
50%
75%
100%
0
1000
2000
3000
4000
5000

Important Observations

- Spark was able to process and train 24M records
- 2 worker Node configuration was not able to handle 24M data
- Some configuration with lesser worker was able to perform better than the higher nodes due to factors like Resource Utilisation and Overhead to run data on multi nodes