

# 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

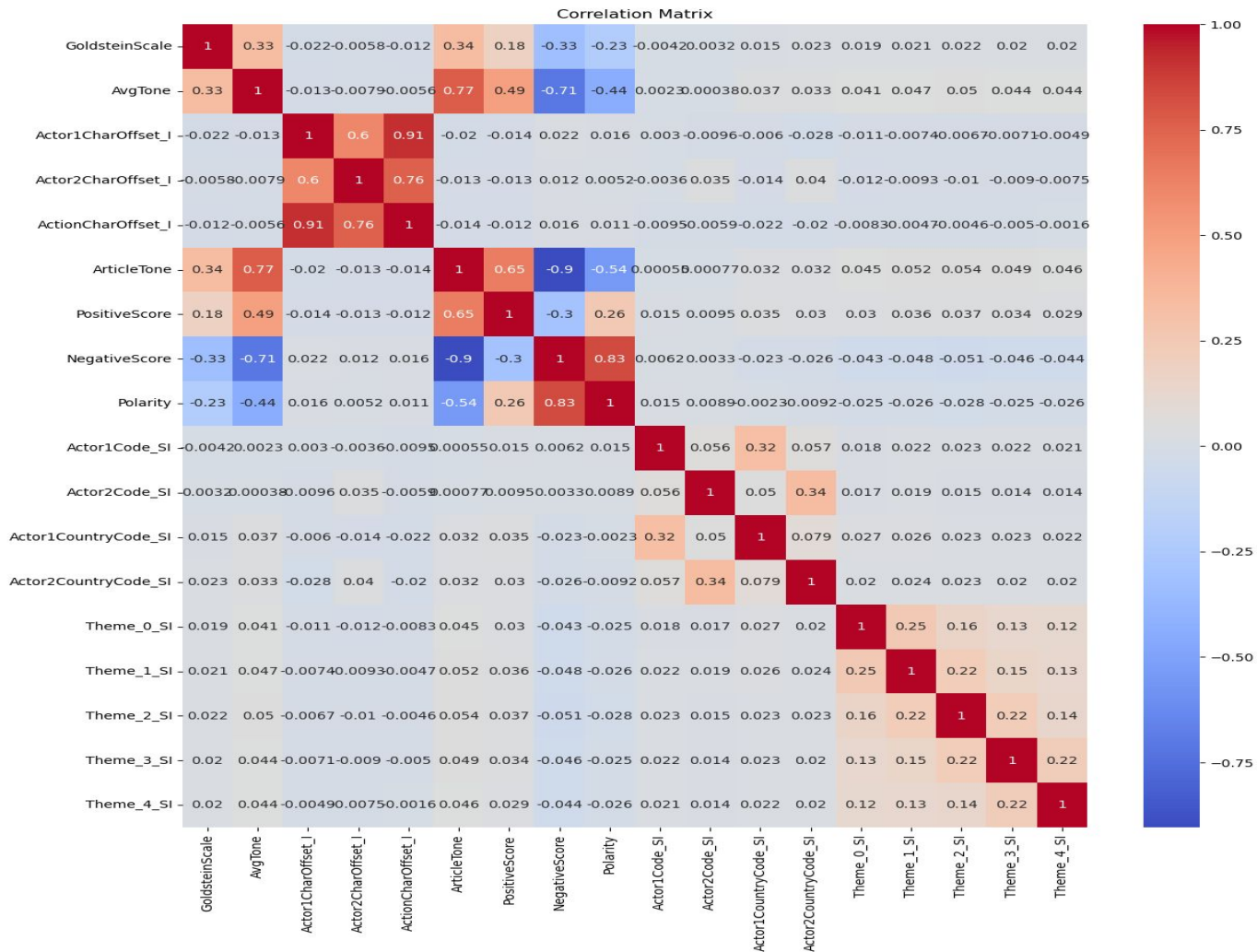
PK= (GLOBAL,GKG)ID

# 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

## Individual columns

```
[1137797685 |20231103064500-1047|2023.8301 |USPEC_POLICY1,2272;USPEC_POLICY1,2715;WB_1458_HEALTH_PROMOTION_AND_DISEASE_PR  
EVENTION,2753;WB_1458_HEALTH_PROMOTION_AND_DISEASE_PREVENTION,3097;WB_1458_HEALTH_PROMOTION_AND_DISEASE_PREVENTION,3160;WB_14  
62_WATER_SANITATION_AND_HYGIENE,2753;WB_1462_WATER_SANITATION_AND_HYGIENE,3097;WB_1462_WATER_SANITATION_AND_HYGIENE,3160;WB_6  
35_PUBLIC_HEALTH,2753;WB_635_PUBLIC_HEALTH,3097;WB_635_PUBLIC_HEALTH,3160;WB_621_HEALTH_NUTRITION_AND_POPULATION,2753;WB_621_  
HEALTH_NUTRITION_AND_POPULATION,3097;WB_621_HEALTH_NUTRITION_AND_POPULATION,3160;TAX_FNCFACT_LEADERS,321;TAX_ETHNICITY_CHINES  
E,1301;TAX_ETHNICITY_CHINESE,1640;TAX_WORLDLANGUAGES_CHINESE,1301;TAX_WORLDLANGUAGES_CHINESE,1640;WB_698_TRADE,278;WB_698_TRA  
DE,492;WB_698_TRADE,738;WB_698_TRADE,897;WB_698_TRADE,1057;WB_698_TRADE,1552;WB_698_TRADE,1739;TAX_FNCFACT_OFFICIALS,302;TAX_F  
NCFACT_BUSINESS_LEADERS,321;AGRICULTURE,2358;TAX_FNCFACT_MAYOR,1171;TOURISM,1135;TOURISM,2562;TOURISM,2584;TOURISM,2633;WB_825_  
TOURISM,1135;WB_825_TOURISM,2562;WB_825_TOURISM,2584;WB_825_TOURISM,2633;WB_1921_PRIVATE_SECTOR_DEVELOPMENT,1135;WB_1921_PRIV  
ATE_SECTOR_DEVELOPMENT,2562;WB_1921_PRIVATE_SECTOR_DEVELOPMENT,2584;WB_1921_PRIVATE_SECTOR_DEVELOPMENT,2633;WB_346_COMPETITIV  
E_INDUSTRIES,1135;WB_346_COMPETITIVE_INDUSTRIES,2562;WB_346_COMPETITIVE_INDUSTRIES,2584;WB_346_COMPETITIVE_INDUSTRIES,2633;WB  
_818_INDUSTRY_POLICY_AND_REAL_SECTORS,1135;WB_818_INDUSTRY_POLICY_AND_REAL_SECTORS,2562;WB_818_INDUSTRY_POLICY_AND_REAL_SECTO  
RS,2584;WB_818_INDUSTRY_POLICY_AND_REAL_SECTORS,2633;CRISISLEX_T11_UPDATESSYMPATHY,2576;UNGP_FORESTS_RIVERS_OCEANS,2625;WB_69  
9_URBAN_DEVELOPMENT,2633;WB_1765_CULTURE_HERITAGE_AND_SUSTAINABLE_TOURISM,2633;TAX_FNCFACT_REPRESENTATIVES,1974;LEADER,825;TAX  
_FNCFACT_PRESIDENT,835;USPEC_POLITICS_GENERAL1,835;EPU_ECONOMY,270;EPU_ECONOMY,484;EPU_ECONOMY,730;EPU_ECONOMY,1049;EPU_ECONOM  
Y_HISTORIC,270;EPU_ECONOMY_HISTORIC,484;EPU_ECONOMY_HISTORIC,730;EPU_ECONOMY_HISTORIC,1049;ECON_FOREIGNINVEST,1659;TAX_ETHNIC  
ITY_AMERICAN,656;TAX_ETHNICITY_AMERICAN,721;TAX_ETHNICITY_AMERICAN,1318;TAX_ETHNICITY_AMERICAN,2003;TAX_ETHNICITY_AMERICAN,24  
99;TAX_ETHNICITY_AMERICAN,2584;TAX_WORLDLANGUAGES_LATIN,2633;TAX_WORLDLANGUAGES_LATIN,396;TAX_WORLDLANGUAGES_LATIN,521;TAX_WOR  
LD_LANGUAGES_LATIN,712;TAX_WORLDLANGUAGES_LATIN,1000;TAX_WORLDLANGUAGES_LATIN,1309;TAX_WORLDLANGUAGES_LATIN,1421;TAX_WORLD
```



# 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

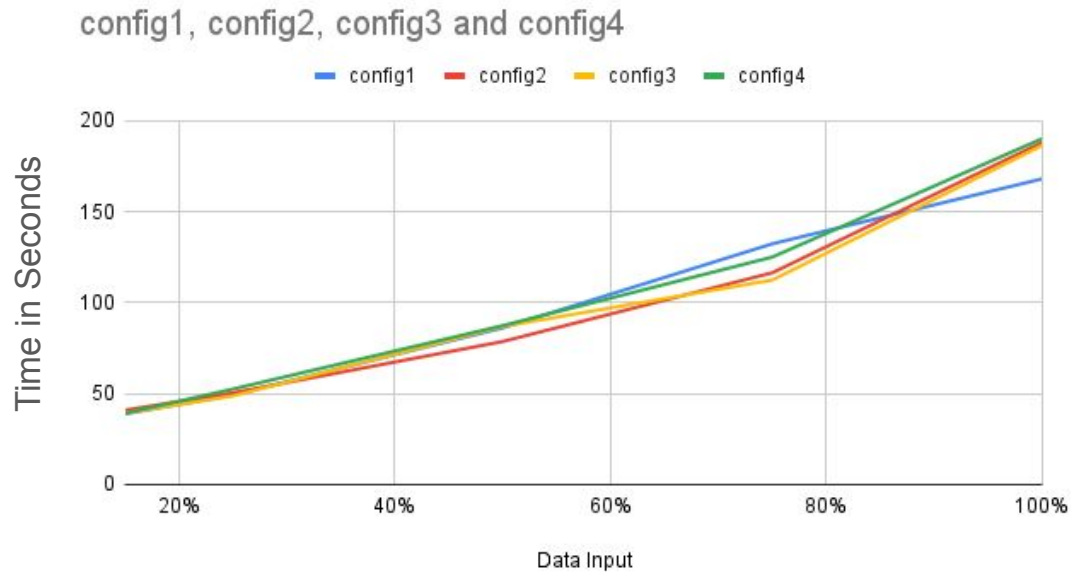
1 Master 2 Workers

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

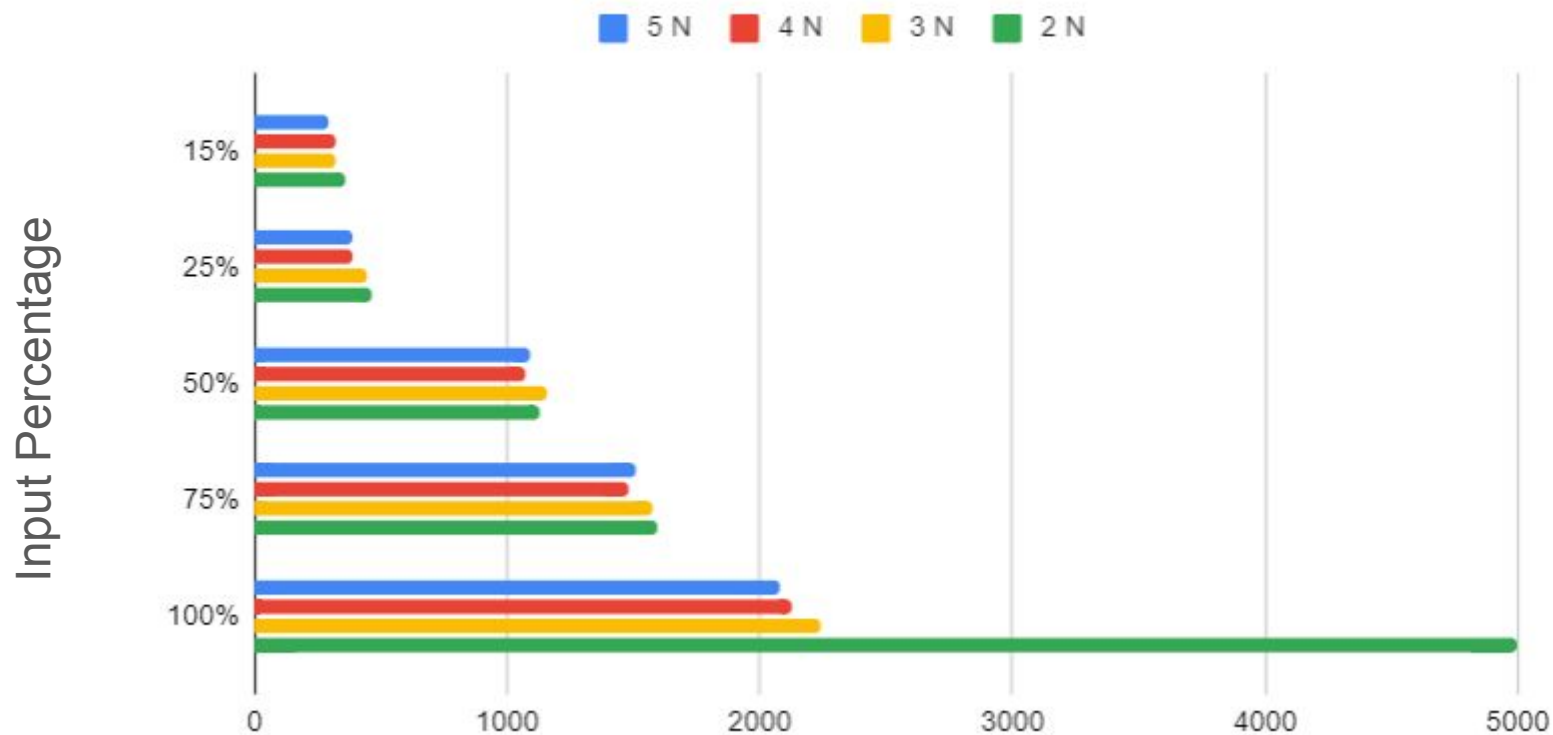
Config3 1M 3N

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



# 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