Transnational and National level conflict analysis using the GDELT Event database

Aditya Mehta (52106261)
Master of Science in Data Science
University of Aberdeen

Under the Guidance of
Dr. Marco Thiel
University of Aberdeen

ABSTRACT

World conflict and Civilization-threatening tensions are principal dimensions of human suffering. Its drastic surge over last few years on both global and local level is a detrimental strain challenging all nations fetching attention of researchers and policymakers. Consequently, with over a million news stories and articles published every day, there is a need to analyse implicit factors, summarize core rationales and predict the hidden trend for optimum endeavour to restrain future Geopolitical turmoil. In the current study, we exploit information extracted from *Global Data on Events, Location, and Tone* (GDELT) digital news database which is the largest real-time, most comprehensive, and modernised data source of useful attributes of every recorded event within its timeline with over a quarter-billion events on record from 1979 to present, to study significance and severity of particular events across worldwide regions and distinct time periods for an attempt to build a foresighted pre-warning mechanism to announce on anticipated Transnational strife and upheaval.

Keywords: GDELT, Conflict, Data analysis.

1.INTRODUCTION

1.1 GDELT

The open-source *Global Database of Events, Language, and Tone* (GDELT) is the biggest most thorough and modernized data source of valuable attributes of every single recorded event that takes place within its respective timeline [1]. GDELT extracts more than 300 categories of events, millions of themes, thousands of emotions, and the networks that connect them using some of the most advanced natural language and data mining algorithms in

the world, including the most potent deep learning algorithms [2]. Over a quarter-billion events on record from 1979 till present date, GDELT Project is a free open platform for computing on the entire planet [2]. It tracks news coverage from nearly every country and watches broadcast, print, and online sources in more than 100 languages to identify the people, places, organizations, themes, sources, feelings, counts, quotes, images, and events that shape our global society every second of every day. [2]. GDELT is CAMEO-coded which stands for *(Conflict and Mediation Event Observations)*, utilizes the TABARI system for events and Geological Names, Latitudes-Longitudes for geocoding, and frequently adds new data improvements [3]–[5].

1.2 CONFLICT

Peace and conflict researchers have studied conflict causes, offered conflict interventions, and predicted conflict spread [6], [7]. GDELT open-source database serves best for analysing conflict trends across different timeline, offering not only international but city level precision [8]–[11]. This report focuses on two main objectives.

Local level Evaluation: Analysis of positive-negative event distribution highlighted based on crucial factors like

1. *GoldsteinScale*: theoretical potential effect of an incident on stability of a nation [9].
2. *AvgTone* of the article: sentiment score of the published event.
3. *ActorType*: role or type of participating actors. For example, *Police, Judiciary, Legal, business*, etc.
4. *ActionGeoFullName*: human-legible name of the event conflict location that was matched.
5. *NumMentions*/*NumSources*: This is the overall number of times this event has been mentioned in all other sources/articles.

Global level Evaluation: Analysis of global conflicts contributing to a war like situation and highlighting rise of incidents within participating nations against global peace filtered by means of:

1. *EventCode*: The act that Actor1 committed against Actor2 is described in this raw CAMEO action code.[3]
2. *Actor1CountryCode, Actor2CountryCode*: Participation of pair of countries involved in specific event category.

2. TOOLS

Wolfram Mathematica (Version 13.0.0.0)

Wolfram Mathematica is used for the data analysis in this report. It is a piece of computer software used for data analysis, symbolic computation, and technical computation [12].

MySQL (8.0.28)

For data storage and access, MySQL is a flexible and dependable open-source database. Only for one calendar year, our data has over 2.5 million rows. As a result, 10-year conflict trend analysis of any country requires a versatile and lightweight database [13].

Tableau (2022.4.0)

Data visualization software used to create interactive and visually appealing dashboards, reports, and charts[14].

3.DATA

Apart from *GDELT Analysis Service* and *Google Big Query*, GDELT offers raw data files in the form of archived format over 20 sub-projects, for this report we have used *GDELT 1.0 Event Database* as main and only source of data for conflict analysis [15]. The daily update is published by 6AM EST every morning on GDELT website. Name of the file, which is in the format *"YYYYMMDD.export.CSV.zip,"* contains the date from the previous day [16]. (For instance, a new file with the name *"20130523.export.CSV.zip"* is added the morning of May 24, 2013). Length of data for each distinct day spans over 50,000 - 150,000 rows and 58 variables (columns). Each row is an event recorded by GDELT service with source URL of the article, timestamp, and numerous valuable aspects such as impact threshold, participating actors with their affiliated groups, precise locations, number of mention/article/sources etc.

Group No.	Variable	Information
1	*GlobalEventID*	Globally unique identifier assigned to each event record
2	*SQLDATE, MonthYear, Year, FractionDate, DATEADDED*	Date the event took place in different formats
3	*Actor1Code, Actor2Code*	The complete raw CAMEO code for Actor
4	*Actor1Name, Actor2Name*	The actual name of the Actor
5	*Actor1CountryCode, Actor2CountryCode*	The 3-character CAMEO code for the country affiliation of Actor
6	*Actor1KnownGroupCode, Actor2KnownGroupCode*	Code for actor affiliated group or organization
7	*Actor1EthnicCode, Actor2EthnicCode, Actor1Religion1Code, Actor2Religion1Code, Actor1Religion2Code, Actor2Religion2Code*	Ethnic and religious affiliation of Actor
8	*Actor1Type1Code, Actor2Type1Code, Actor1Type2Code, Actor2Type2Code, Actor1Type3Code, Actor2Type3Code*	The 3-character CAMEO code of the CAMEO "type" or "role" of Actor
9	*IsRootEvent*	Factor which classifies event as an important event
10	*EventCode, EventBaseCode, EventRootCode*	This is the raw CAMEO action code describing the action that Actor1 performed upon Actor2
11	*QuadClass*	This field specifies this primary classification for the event type, allowing analysis at the highest level of aggregation.
12	*GoldsteinScale*	Each CAMEO event code is assigned a numeric score from-10 to+10
13	*NumMentions, NumSources, NumArticles*	This is the total number of mentions, sources, articles of this event across all source documents.
14	*AvgTone*	This is the average "tone" of all documents containing one or more mentions of this event.
15	*Actor1Geo_FullName, Actor2Geo_FullName, ActionGeo_FullName*	This is the full human-readable name of the matched location.
16	*Actor1Geo_Type, Actor2Geo_Type, ActionGeo_Type*	This field specifies the geographic resolution of the match type and can be used to filter events by geographic specificity
17	*Actor1Geo_CountryCode, Actor2Geo_CountryCode, ActionGeo_CountryCode*	This is the 2-character FIPS10-4 country code for the location.
18	*Actor1Geo_ADM1Code, Actor2Geo_ADM1Code, ActionGeo_ADM1Code*	This is the 2-character FIPS10-4 country code followed by the 2-character FIPS10-4 administrative division 1 (ADM1) code for the administrative division housing the landmark.
19	*Actor1Geo_Lat, Actor1Geo_Long, Actor2Geo_Lat, Actor2Geo_Long, ActionGeo_Lat, ActionGeo_Long*	Centroid latitude and longitude of the landmark for mapping
20	*Actor1Geo_FeatureID, Actor2Geo_FeatureID, ActionGeo_FeatureID*	GNS or GNIS FeatureID for this location
21	*SOURCEURL*	URL of the news article the event was found in

Figure i Summary of all variables of an event offered by GDELT Event database 1.0 [1].

4. METHODOLOGY

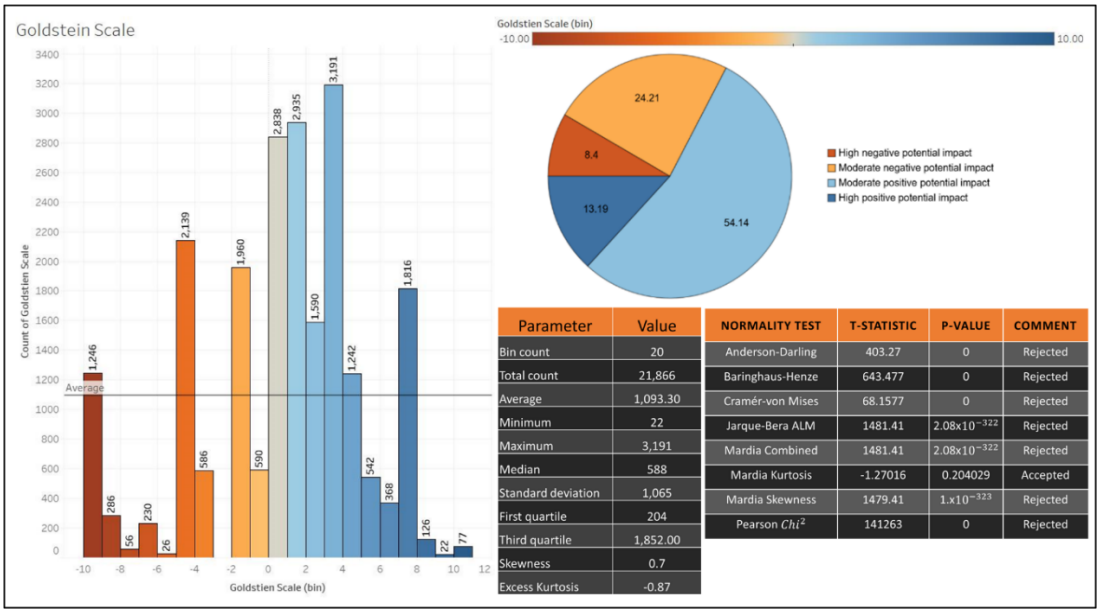
4.1 LOCAL

United States of America (USA) is selected for the scope of Local event data analysis in this research [17], [18]. A total of 21866 event records under the category *ActionGeo_Type* with value '2' = USSTATE, '3' = USCITY are available for the day of March 18,2023 from *GDELT 1.0 Event Database*.

4.1.1 IMPACT FACTORS

Goldstein Scale

"Goldstein scale" by Joshua Goldstein assigned to each CAMEO event code measures worldwide conflict and cooperation [9]. International relations trends are assessed using this scale. -10 is the most strife and +10 the most cooperation on the Goldstein scale. A -10 event may involve armed conflict, violence, or threats of violence, while a +10 event may involve diplomatic negotiations, economic cooperation, or cultural exchange.



Parameter	Value
Bin count	20
Total count	21,866
Average	1,093.30
Minimum	22
Maximum	3,191
Median	588
Standard deviation	1,065
First quartile	204
Third quartile	1,852.00
Skewness	0.7
Excess Kurtosis	-0.87

NORMALITY TEST	T-STATISTIC	P-VALUE	COMMENT
Anderson-Darling	403.27	0	Rejected
Baringhaus-Henze	643.477	0	Rejected
Cramér-von Mises	68.1577	0	Rejected
Jarque-Bera ALM	1481.41	2.08×10^{-322}	Rejected
Mardia Combined	1481.41	2.08×10^{-322}	Rejected
Mardia Kurtosis	-1.27016	0.204029	Accepted
Mardia Skewness	1479.41	$1.x10^{-323}$	Rejected
Pearson Chi^2	141263	0	Rejected

Figure 2 – *GoldsteinScale* (Histogram, coverage, statistical parameters, normality test results)

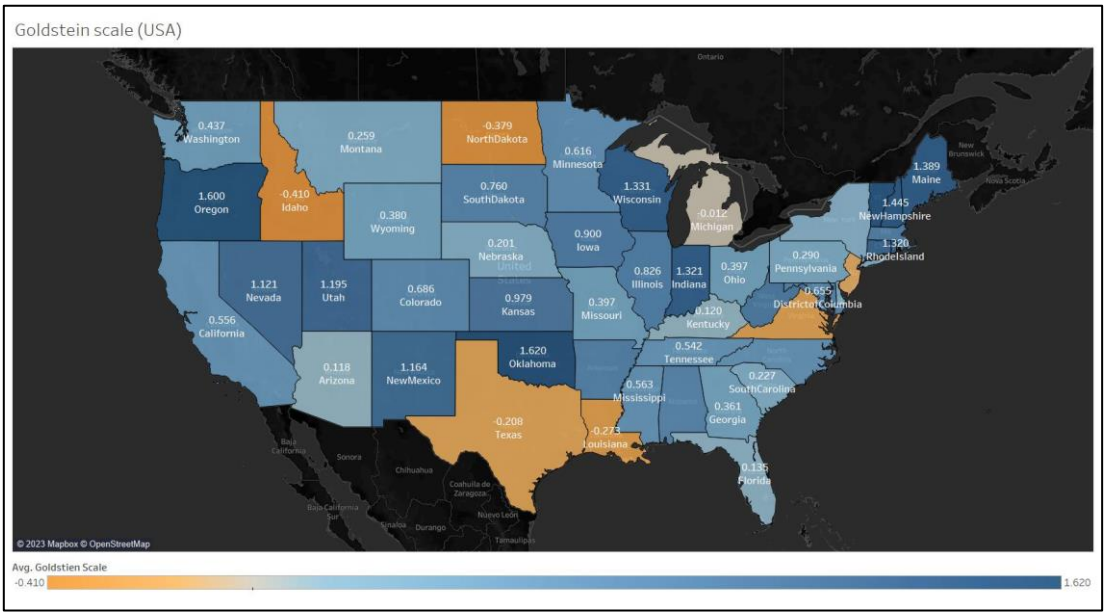
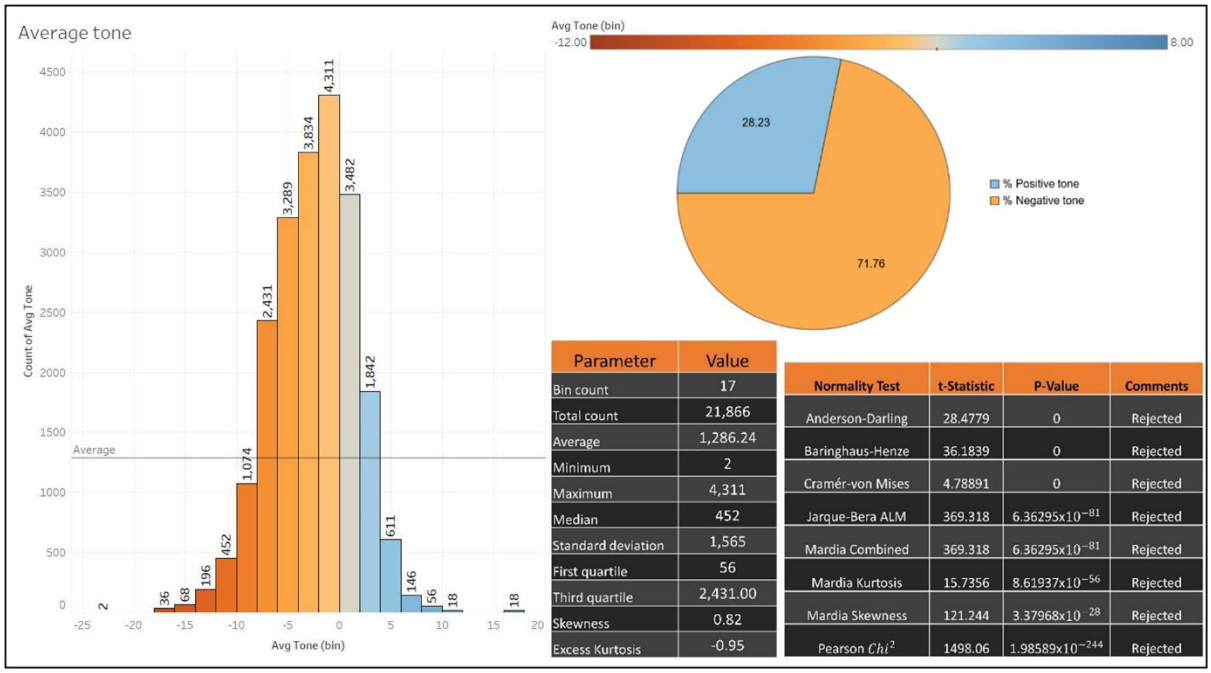


Figure 3 – Average *GoldsteinScale* distribution in United States of America

Histogram plot and Normality test results from *Figure 2* indicate that the scale is not normally distributed. Except for the *Mardia-Kurtosis* test, all other tests provide strong evidence to reject normality. In addition, it does not match any other statistical distribution. The *Figure 3* depicts the average distribution of impact by occurring events, with *Idaho* and *Oklahoma* having the greatest negative and positive impacts, respectively.

Average Tone

This is the average "tone" of all documents mentioning this event at least once. The score ranges from -100 (extremely negative) to +100 (extremely positive). Common values range from -10 to +10, with 0 representing neutrality. This can be used to filter the "context" of events as a subtle indicator of significance of an event. For instance, a riot event with a mildly negative average tone is likely to have been a minor occurrence, whereas a significantly negative average tone indicates a significantly more serious occurrence.



Parameter	Value
Bin count	17
Total count	21,866
Average	1,286.24
Minimum	2
Maximum	4,311
Median	452
Standard deviation	1,565
First quartile	56
Third quartile	2,431.00
Skewness	0.82
Excess Kurtosis	-0.95

Normality Test	t-Statistic	P-Value	Comments
Anderson-Darling	28.4779	0	Rejected
Baringhaus-Henze	36.1839	0	Rejected
Cramér-von Mises	4.78891	0	Rejected
Jarque-Bera ALM	369.318	6.36295×10^{-81}	Rejected
Mardia Combined	369.318	6.36295×10^{-81}	Rejected
Mardia Kurtosis	15.7356	8.61937×10^{-56}	Rejected
Mardia Skewness	121.244	3.37968×10^{-28}	Rejected
Pearson Chi^2	1498.06	1.98589×10^{-244}	Rejected

Figure 4 - *AvgTone* (Histogram, coverage, statistical parameters, normality test results)

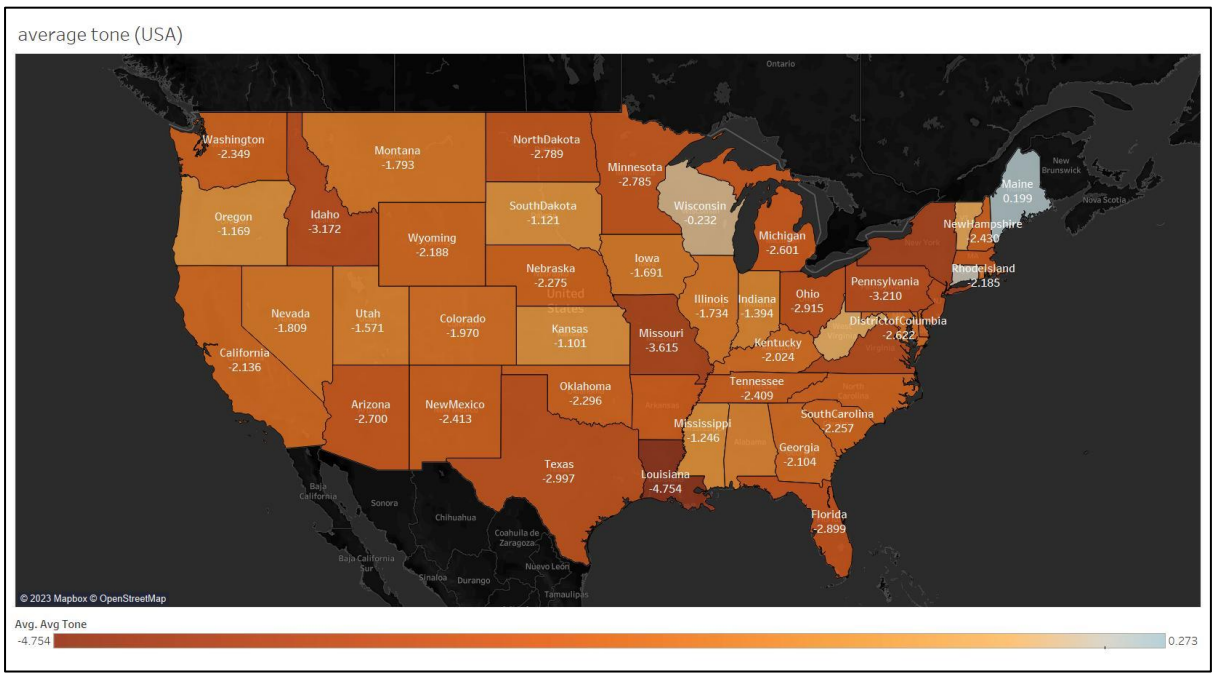


Figure 5 – Average *AvgTone* distribution in United States of America

Figure 4 depicts the average tone, which visually appears to be normally distributed; however, like the *Goldstein scale*, the *AvgTone* fails all normality tests. Furthermore, it demonstrates that majority of articles have a negative tone. *Figure 5* indicates Louisiana on average has most negative tone of all the articles related to that state on day (March 18,2023).

Group type, Numeric Mentions and Quad Class

Linking the available affiliated group type of the participating actor, as provided by GDELT, across the four primary classifications (*Verbal Cooperation, Material Cooperation, Verbal Conflict, and Material Conflict*) under which the entire CAMEO event taxonomy is ultimately organized provides an overview of event impact dispersion [3]. Along with that, *NumMentions* represents the total number of occurrences of an event in all source documents. Multiple references to an occurrence within a single document contribute to this total. This is a method for determining the "importance" of an event. Simply put, the more that event is discussed, the more likely it is to be significant.

Group Type	Participation count	Numeric Mentions	Quad Class	Percentage by Class
Agriculture	83	670	Verbal Cooperation	60.24
			Material Cooperation	22.89
			Verbal Conflict	9.639
			Material Conflict	7.229
Business	1331	13 882	Verbal Cooperation	66.34
			Material Cooperation	14.27
			Verbal Conflict	11.27
			Material Conflict	8.114
Police forces	1724	15 882	Verbal Cooperation	41.36
			Material Cooperation	13.46
			Verbal Conflict	8.469
			Material Conflict	36.72
Criminal	362	3234	Verbal Cooperation	41.16
			Material Cooperation	12.43
			Verbal Conflict	17.68
			Material Conflict	28.73
Civilian	1559	13 547	Verbal Cooperation	64.91
			Material Cooperation	12.64
			Verbal Conflict	12.83
			Material Conflict	9.622
Education	1973	16 388	Verbal Cooperation	69.94
			Material Cooperation	11.25
			Verbal Conflict	9.478
			Material Conflict	9.326
Elites	432	4471	Verbal Cooperation	59.72
			Material Cooperation	12.04
			Verbal Conflict	9.259
			Material Conflict	18.98
Environmental	7	36	Verbal Cooperation	71.43
			Material Cooperation	14.29
			Verbal Conflict	14.29
Government	3114	35 810	Verbal Cooperation	65.35
			Material Cooperation	10.63
			Verbal Conflict	13.29
			Material Conflict	10.73

Figure 6 Classification of events based on *QuadClass* and *ActorType*

For example, from the *Figure 6*, *Police Forces* account for 36% of *Material Conflict* and only 13% of *Material cooperation* with overall more than 15,000 mentions of that day (March 18,2023).

4.1.1 STATISTICAL TESTS

GoldsteinScale and AvgTone

ρ_{xy} is the population correlation coefficient for *GoldsteinScale* and *AvgTone*.

$H_0 : \rho_{xy} == \rho_0 : There\ is\ no\ correlation$
$H_a : \rho_{xy} \neq \rho_0 \quad : Data\ is\ correlated$

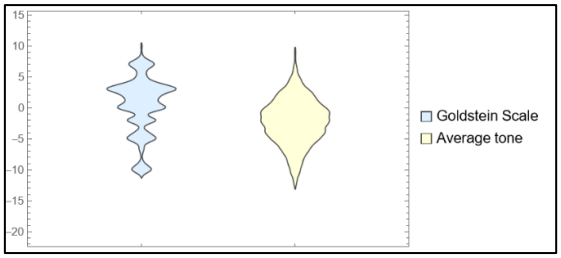


Figure 7 Distribution plot *Goldstein scale vs AvgTone*

Table 1. Correlation test results between *Goldstein scale and AvgTone*	
Test	Spearman Rank Correlation Test
p-value	0.
Test Statistic	0.339
Significance Level	95%

Conclusion : The null hypothesis that the population rank correlation coefficient is equal to 0. is rejected at the 5 percent level based on the Spearman Rank test [19].

NumMentions and NumSources

ρ_{xy} is the population correlation coefficient for *NumMentions* and *NumSources*

.

$H_0 : \rho_{xy} == \rho_0 : There\ is\ no\ correlation$
$H_a : \rho_{xy} \neq \rho_0 \quad : Data\ is\ correlated$

ii

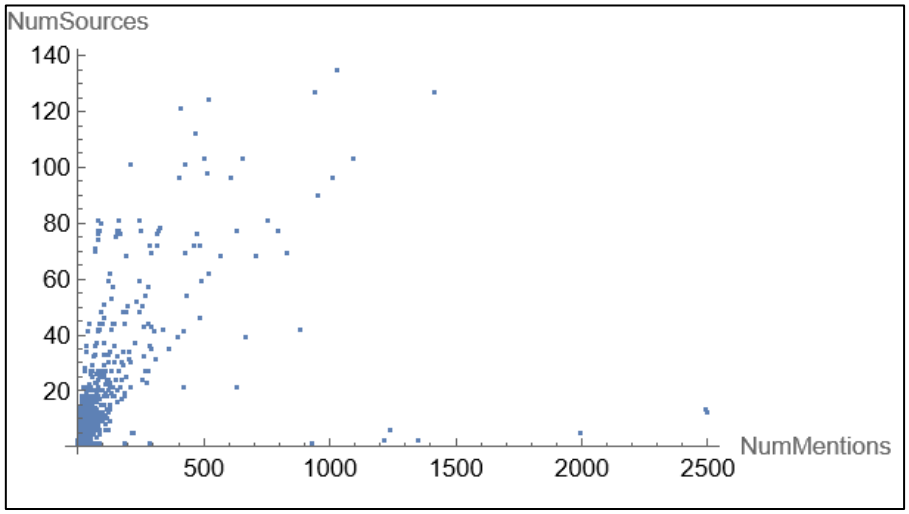


Figure 8 List plot *NumMentions* vs *NumSources*

Table 2. Correlation test between *NumMentions* and *NumSources*	
Test	Spearman Rank Correlation Test
p-value	0.
Test Statistic	0.59
Significance Level	95%

Conclusion : The null hypothesis that the population rank correlation coefficient is equal to 0. is rejected at the 5 percent level based on the Spearman Rank test [19].

Final Interpretation : The values of 0.339 and 0.565 for the Test Statistic indicates a moderately positive linear relationship between the tested variables. Moreover, it is possible that each variable has a significant effect on the outcome of interest, and it may therefore be beneficial to consider each variable separately.

Correlation Matrix

The *GoldsteinScale* is divided into positive and negative impact, therefore there is a possibility that number of mentions can be biased towards one or the other. To check impact of *AvgTone*, *NumMentions*, *NumSources* on positive and negative *GoldSteinScale* separately, we plot correlation matrix as follows

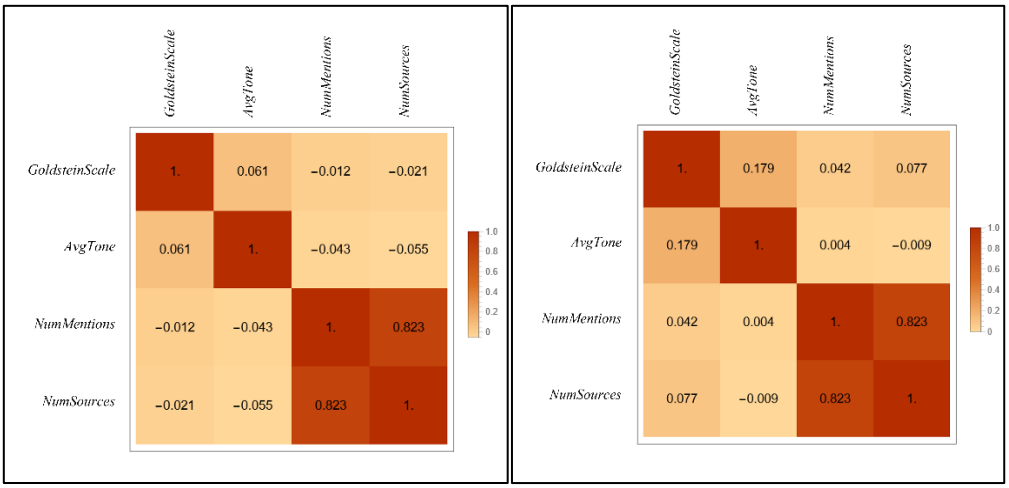


Figure 9 Correlation Matrix of variables for separate positive & negative *GoldsteinScale*

Figure 9 visually shows both correlation matrix has very high similarity where most of the values except *NumMentions* vs *NumSources* lie very close to 0, meaning there is no considerable correlation. Moreover, value of -0.012 and 0.042 for positive and negative *GoldsteinScale* vs *NumMentions* respectively, concludes that there is favouritism of magnitude of event mentions on the type of impact.

Linear Relations

To know combined influence of all factors on each other, a linear model is fitted with each variable as dependent variable separately.

Table 3 Linear function fit for each variables

	Dependent Variable	Independent variables	Estimate	Standard Error	t-Statistic	P-Value
1	GoldsteinScale					
		1	1.3065	0.0449846	29.0434	1.91064*10^-179
		Avgtone	0.411059	0.00897134	45.8191	0
		NumMentions	0.00015228	0.0013038	0.116797	0.907023
		NumSources	0.00815928	0.00703907	1.15914	0.246421
	Equation	**1.3065 + 0.411059 Avgtone + 0.00015228 NumMentions + 0.00815928 NumSources**				
2	Avgtone					
		1	-2.44161	0.0371319	-65.7549	0
		GoldsteinScale	0.354097	0.00772815	45.8191	0
		NumMentions	0.00166886	0.00121	1.37922	0.167852
		NumSources	-0.0253922	0.00652953	-3.88883	0.000101256
	Equation	**-2.44161 + 0.354097 GoldsteinScale + 0.00166886 NumMentions - 0.0253922 NumSources**				
3	NumMentions					
		1	1.28831	0.321047	4.01285	6.03451E-05

		GoldsteinScale	0.00726889	0.0622353	0.116797	0.907023
		Avgtone	0.0924758	0.0670493	1.37922	0.167852
		NumSources	4.43431	0.0277468	159.814	0
	Equation	**1.28831 + 0.0924758 Avgtone + 0.00726889 GoldsteinScale + 4.43431 NumSources**				
4	NumSources					
		1	0.649711	0.0592125	10.9725	6.97263*10^-28
		GoldsteinScale	0.0133605	0.0115262	1.15914	0.246421
		NumMentions	0.152115	0.000951825	159.814	0
		Avgtone	-0.0482673	0.0124118	-3.88883	0.000101256
	Equation	**0.649711 - 0.0482673 Avgtone + 0.0133605 GoldsteinScale + 0.152115 NumMentions**				

Table 3 shows

1. *Avgtone* and *GoldsteinScale* are positively correlated (p-value = 0) with an estimate of 0.411059, where *GoldsteinScale* is increased by 0.4110592 units due to *Avgtone*. *NumMentions* and *NumSources* have respective p-values of 0.907023 and 0.246421, therefore no relationship with *GoldsteinScale*. At 1.3065, the intercept term is statistically significant with a standard error of 0.0449846 and a t-statistic of 29.0434, which suggests *GoldsteinScale* is 1.3065 even if all independent factors are 0.
2. *NumSources* is negatively correlated with *Avgtone* (p=0.000101256), with an estimate of -0.0253922 (p-value = 0.000101256), means while *NumSources* grows, *AvgTone* decreases by 0.0253922 units. With a p-value of 0.167852, *AvgTone* is not significantly associated to *NumMentions*. The statistical significance of the intercept term is -2,44161, suggest that even without any independent factors, the expected value of *AvgTone* is -2,44161.
3. *NumSources* and *NumMentions* have a positive correlation (p-value = 0) with an estimate of 4.43431. As *NumSources* increase, *NumMentions* rise by 4.43431. *GoldsteinScale* and *AvgTone* have insignificant p-values suggesting no relation with *NumMentions*.
4. *NumMentions* and *NumSources* have a 0.152115 positive association, *NumMentions* increases, *NumSources* increases by 0.152115 units. *AvgTone* is negatively correlated with *NumSources* (p-value=0.000101256), with an estimate of -0.0482673. Means *AvgTone* increases, *NumSources* decreases by 0.0482673 units. With a p-value of 0.246421, there is no proof that *GoldsteinScale* and *NumSources* are related. The intercept term has a statistically significant positive value of 0.649711.Even with zero independent factors, *NumSources'* expected value is 0.649711.

4.1.2 Anticipated Event Impact (AEI)

Following Quote raises a requirement for a more revised factor to measure the impact of an event more fairly.

"This score is based on the type of event, not the specifics of the actual event record being recorded – thus two riots, one with 10 people and one with 10,000, will both receive the same Goldstein score"[1].

"The core engine for the event coding was the open-source Tabari program"[1].

TABARI does make mistakes with complex sentences and sentences containing atypical grammatical constructions[20]. Notably, the accuracy of TABARI is highly dependent on the source text, the event coding scheme, and the nature of event being coded [11]. For instance, if the algorithm fails to recognise the actor or actor group of an event, this can result in an inaccurate categorization and an erroneous Goldstein scale. The objective of *AEI* is to combine event categorization with factors that contribute to designating an event as "Important" in order to disclose the actual influence of articles on locations and actors. The marker (variable) called *IsRootEvent* can serve as an indicator of the approximate significance of an event to generate subsets of the event stream, therefore for the estimation of *AEI*, only the events with *IsRootEvent* = TRUE are taken into account[1]. To calculate Anticipated Event Impact (*AEI*), the factors used are:

$$p = Goldsteinscale$$
$$q = AvgTone$$

$$r = NumMentions$$

iii

$$n = \begin{cases} negative, & p_n < 0 \\ positive, & p_n \geq 0 \end{cases}$$

iv

$$p'_{n_i} = \frac{p_{n_i} - p_{n_{min}}}{p_{n_{max}} - p_{n_{min}}}$$

v

$$q'_i = \frac{q_i - q_{min}}{q_{max} - q_{min}}$$

vi

$$r'_i = \frac{r_i - r_{min}}{r_{max} - r_{min}}$$

vii

$$a = AvgTone\ weighing\ factor\ (experimental)$$

viii

$$AEI_{n_i} = \frac{(p'_{n_i} + a(q'_i))}{2} + r'_i$$

ix

$$AEI'_i = S\left(\frac{AEI_{n_i} - AEI_{n_{min}}}{AEI_{n_{max}} - AEI_{n_{min}}}\right), S = \begin{cases} 10, & n = positive \\ -10, & n = negative \end{cases}$$

x

$$\overline{AEI'_{state}} = \frac{\sum_{i=1}^{n} AEI'_{i(state)}}{n}$$

xi

Equations *v, vi, vii* converts variables from Equation *iii* into normalized form. Equation *viii* shows weighing factor for the two variables which will be different for each country. For calculation of USA based events on March 18,2023 the value of a and b is approximately assumed 0.4 respectively based on correlation test results. *AEI* value is normalized value ranging -10 (most negative impact) to +10(most positive impact).

State name	AEI_{state}	Actor Group 1	Actor Group 2	Actor Group 3
New Mexico	1.149049719	Government	Legislature	Education
Oregon	1.102438343	Education	Civilian	Government
Oklahoma	1.101171161	Business	Judiciary	Government
Vermont	1.079437287	Government	Judiciary	Education
Connecticut	1.038678357	Business	Government	Education
North Carolina	**0.108361777**	Judiciary	Education	Legislature
Ohio	**0.064020929**	Government	Civilian	Police forces
New York	**-0.015117651**	Judiciary	Government	Media
New Jersey	**-0.069708**	Education	Government	Judiciary
Mississippi	**-0.150268963**	Education	Judiciary	Government

Figure 10 Average *AEI* of some states of USA

Figure 10 displays average *AEI* of few states of USA and major actor group responsible in order of columns shown. For example, Mississippi is below the neutral level where *Education* actor group is responsible.

4.2 Global

On a global level, international conflicts occurring worldwide from 2014/01/01 till 2023/03/09 are chosen as centre of significance for this study[21], [22].

$$Pair_{(x,y)} = Participating\ actor\ pair\ (x,y)\ on\ certian\ calendar\ day$$
$$Score_{(x,y)} = Total\ Occurance\ count\ of\ actor\ pair\ (x,y) on\ certain\ calendar\ day$$

xii

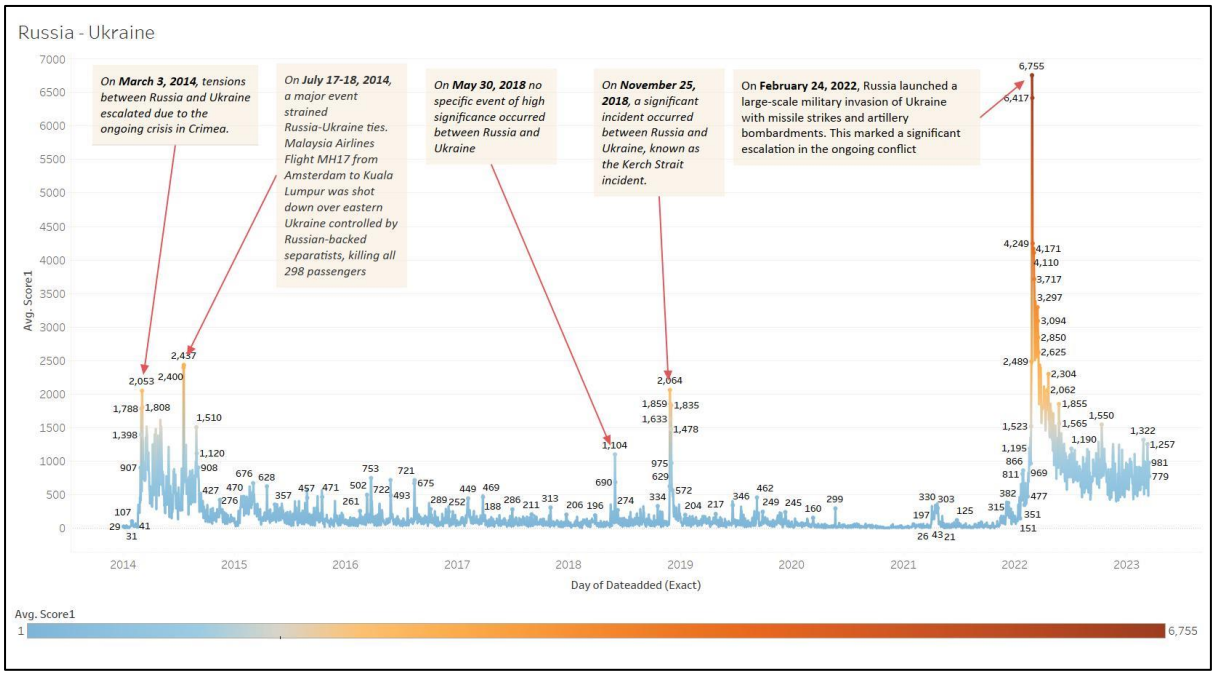


Figure 11 Russia-Ukraine Conflict score plot with major conflict event description and timestamp

Goldstein Scale	Event Code	Event Description
-10	190	Use conventional military force, not specified below
-10	202	Engage in mass killings
-9.5	192	Occupy territory
-9.5	201	Engage in mass expulsion
-8	163	Impose embargo, boycott, or sanctions
-7.5	145	Protest violently, riot
-7.5	1451	Engage in violent protest for leadership change
-6.5	1422	Conduct hunger strike for policy change
-5.8	134	Threaten to halt negotiations
-5	1724	Impose state of emergency or martial law

Figure 12 Some of the total *EventCode* selected for global study

Figure 12 exhibits opted *GoldsteinScale* for analysis from range -5 to -10 assigned to distinct *EventCode* will aid in classification of high to extreme uncooperative and detrimental exchange between nations respectively.

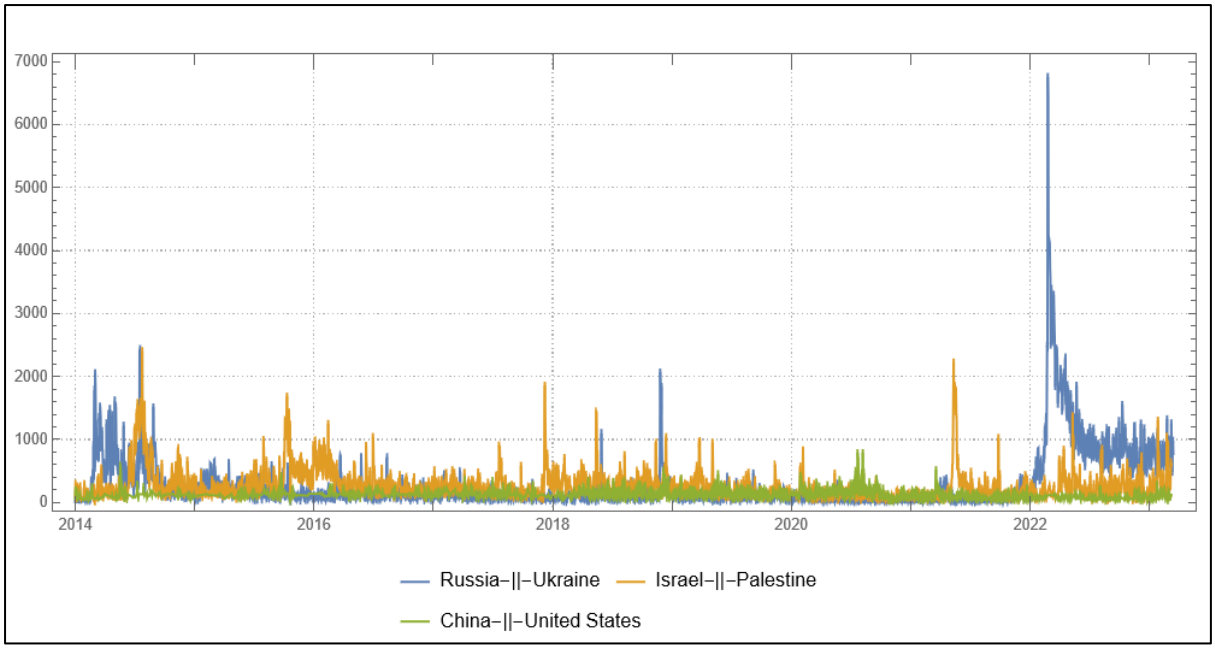


Figure 13 Conflict score time series plot for 3 actor pairs

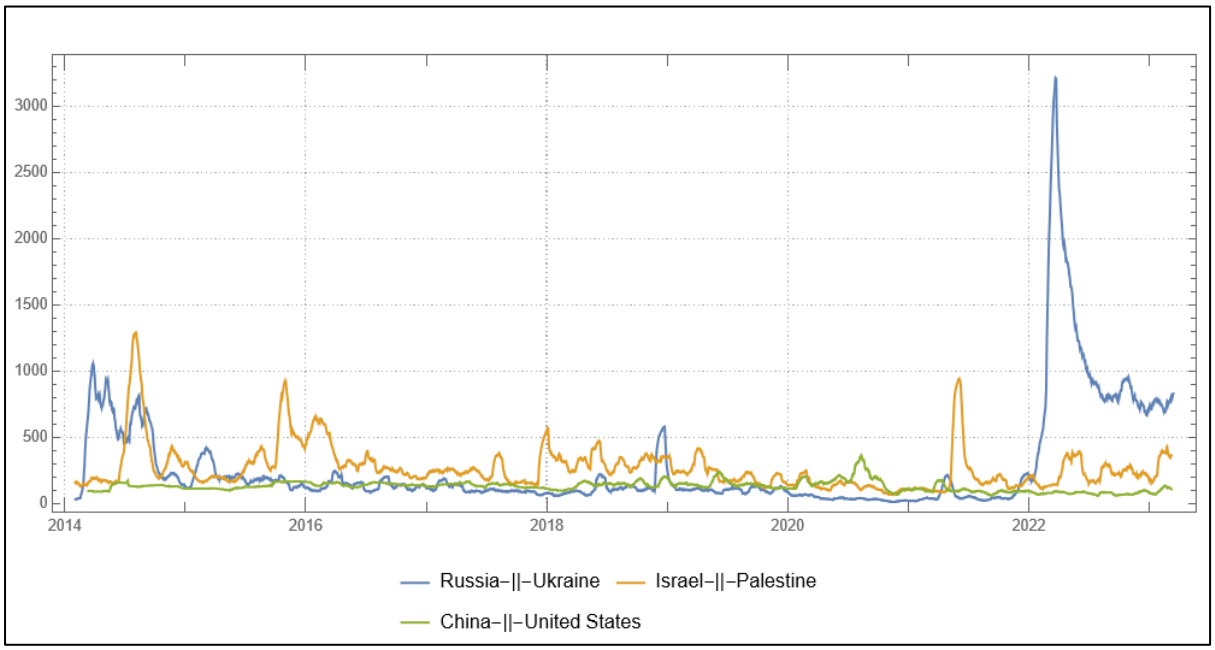


Figure 14 Conflict score time series plot for 3 actor pair filtered with moving average at lag 30 days.

Figure 14 displays Placement of $Pair_{(x,y)}$ with their respective $Score_{(x,y)}$ spanned over 9 years hints about conflict trend. visual comparison between three actor pairs with peaks being the point of interest reflects rise of Russia-Ukraine to war like circumstances, exact day of war announcement, and decline. Accordingly, Russia-Ukraine conflict timeline is selected for further detailed timeseries analysis.

4.2.2 Statistical Tests

Auto-correlation test

ρ_k is the auto-correlation coefficient at lag k.

$H_0 : \rho_k == 0 : Data\ are\ uncorrelated$
$H_a : \rho_i \neq 0 \quad : Presence\ of\ Auto - correlation\ at\ particular\ lag$

xiii

Table 4 Auto-correlation test results	
Test	Ljung-Box
p-value	0.
Test Statistic	21502.3
Significance Level	95%

Conclusion : The null hypothesis that the data are uncorrelated to lag 9 is rejected at the 5 percent level based on the Ljung-Box test [23].

Stationarity Test

$H_0 : The\ time\ series\ satisfying\ an\ AR\ model\ has\ a\ unit\ root$
$H_a : Timeseries\ has\ no\ unit\ root$

xiv

Table 5 Unit Root test results	
Test	Dickey-Fuller F
p-value	1.54522*10^-10
Test Statistic	-152.477
Significance Level	95%

Conclusion : The null hypothesis that the Timeseries contains a unit root is rejected at the 5 percent level based on the Dickey-Fuller F test [24].

Final Interpretation *: Ljung-Box* test statistic is 21502.3, which is a very large value and indicates very strong auto-correlation at multiple latency levels, which violates one of the assumptions of linear regression, namely that residuals are independent and uniformly distributed. This can result in erroneous parameter estimates and forecasts. Moreover, whilst visual graph exhibits clear multiple peaks, the stationarity test is passed, nevertheless. This might result in unfit model such as ARIMA and SARIMA which assumes the series to be stationary.

4.2.3 Time series analysis

Before fitting model to the timeseries data, multiplicative decomposition of series is performed[25]. The decomposition of a time series can be beneficial for identifying underlying data patterns and predicting future values[26]. Once the components are separated, statistical models can be fitted separately to each component and then combined to create predictions for the original time series. trend component which represents long-term pattern of data, which may be increasing, decreasing, or stable. The seasonality component represents periodic data fluctuations, such as daily, weekly, or annual patterns. The noise component represents random fluctuations or data defects not accounted for by the trend and seasonality components.

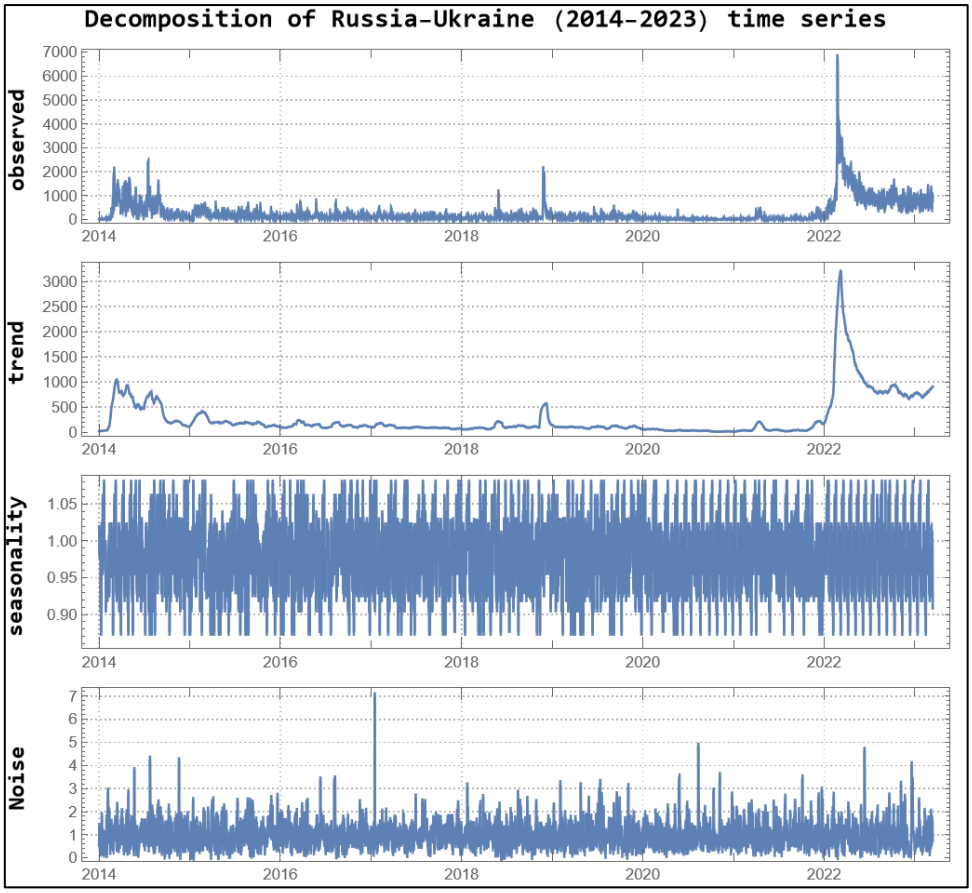


Figure 15 Decomposition of Russia-Ukraine time series data

After detrending the timeseries data by dividing it with moving average of 30-day period, *Figure 15* shows a flat linear trend that can be assumed if we exclude the peak after 2022. No specific pattern in seasonality of data is observed on a 30 day lag scale and due to inherent variability, the random noise component dominates the time series, making it difficult to identify any underlying patterns or trends. Since there is no discernible trend or seasonality in the data, it is enticing to conclude that the time series is stationary. The large random noise component, however, can make it challenging to derive meaningful conclusions from summary statistics or visual inspection of the data. Therefore, based on decomposition and statistical test performed, AR (Auto-Regressive) and MA (Moving Average) statistical model are best suit for fitting the time series data.

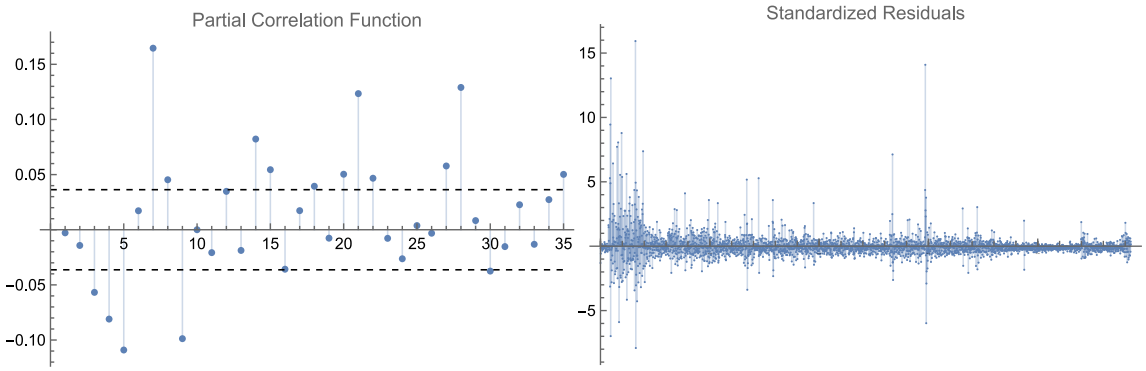


Figure 16 PACF and Standardized residuals plot for AR model order 3

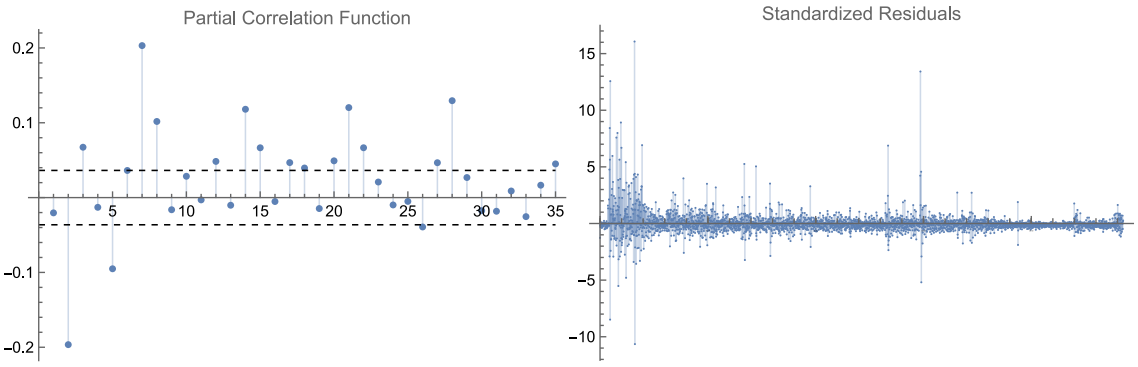


Figure 17 PACF and Standardized residuals plot for AR model order 2

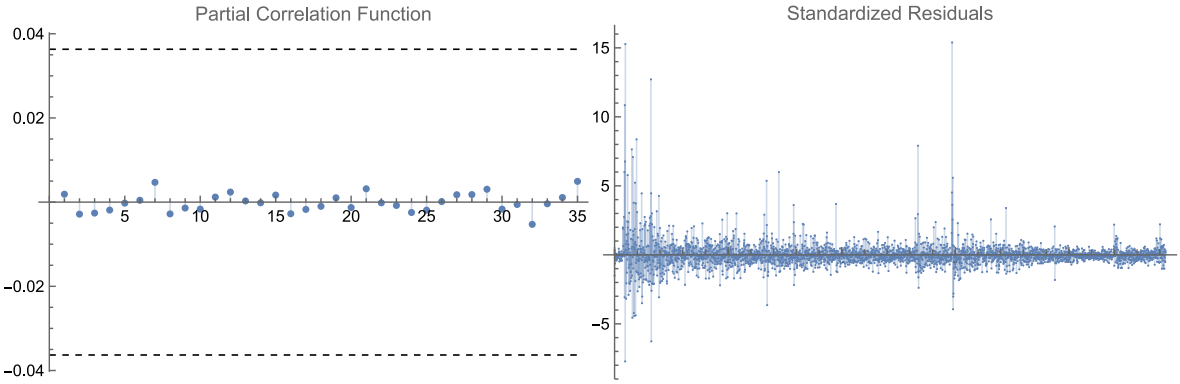


Figure 18 PACF and Standardized residuals plot for MA model order 421

Partial correlation function plot in both *Figure 16* and *Figure 17* shows significant non-zero correlations at multiple lag levels which also states that the timeseries is non-random. 95% threshold levels above and below x-axis indicated by dashed line is crossed multiple times at lag 3,5,7,8,14,15,12 and more. There is no sign of geometric decay in both *Figure 16 and Figure 17* . ACF and PACF plots of AR and MA models are very straight forward. However, some ACF and PACF graphs for real-world time series data sets are ambiguous. Considerable number of outliers or extreme values in the standardized residual plot in *Figure 16, Figure 17* and *Figure 18* indicates influential points that may have an undue impact on the Timeseries model. In addition, the majority of residual values are distant from x-axis, indicating that the relationship between the predictor variables and the response variable may be nonlinear, which can lead to large prediction errors.

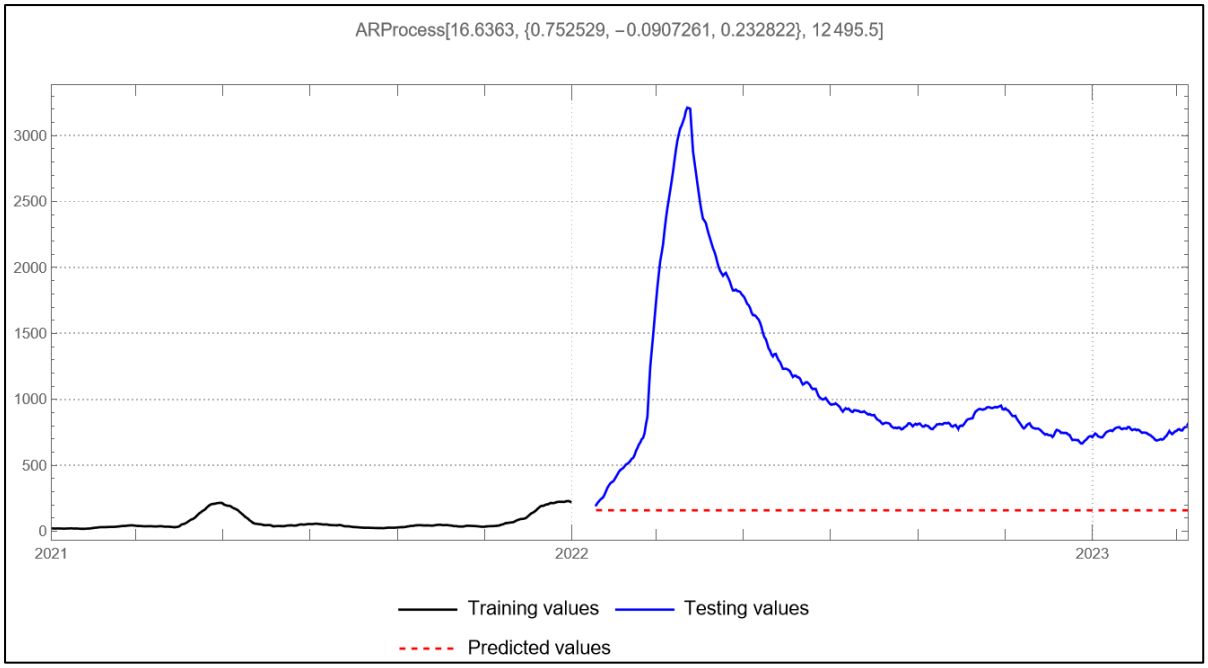


Figure 19 Russia-Ukraine conflict peak prediction plot for model AR with order 3

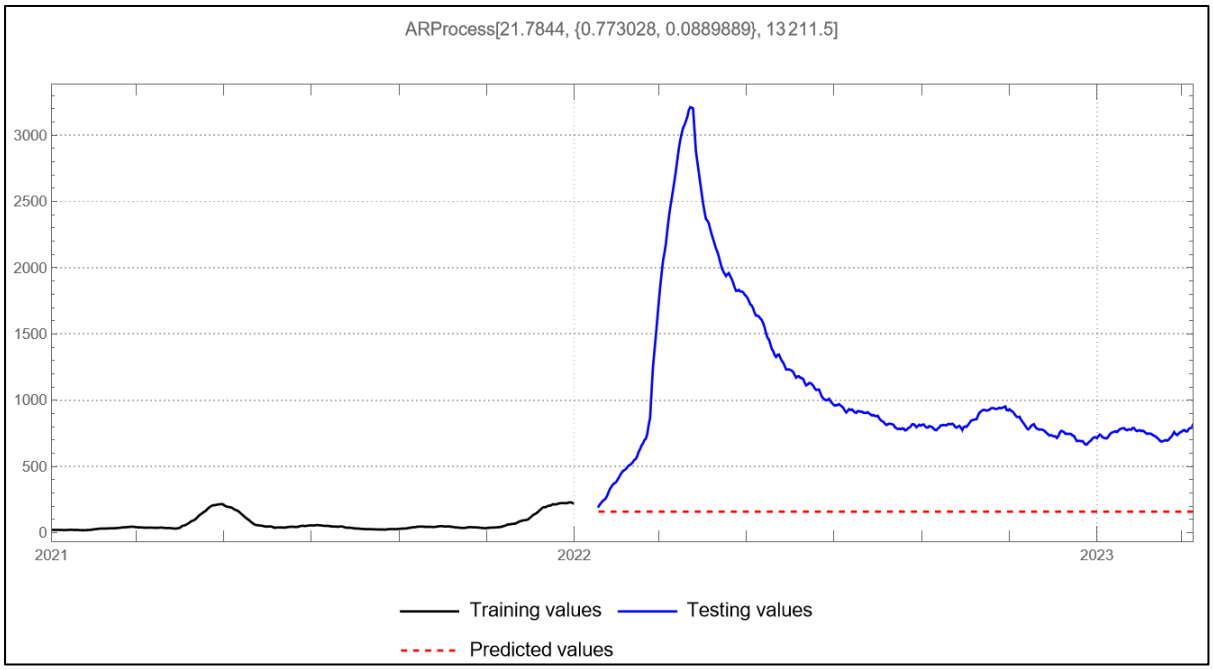


Figure 20 Russia-Ukraine conflict peak prediction plot for model AR with order 2

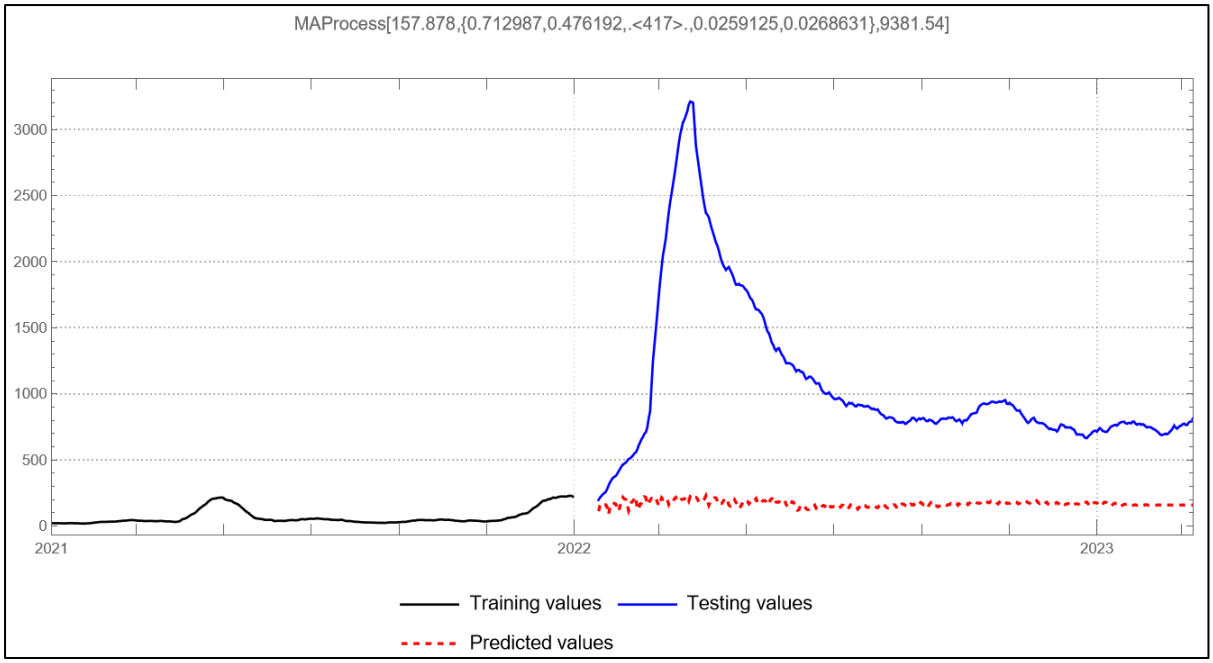


Figure 21 Russia-Ukraine conflict peak prediction plot for model MA with order 421

Parameters	**Values**		
Model Family	Auto Regressive	Auto Regressive	Moving Average
Order	3	2	421
AIC (Akaike Information Criterion)	27479.3	27639.5	27480.6
BIC (Bayesian Information Criterion)	27511.3	27665.7	28437.3
AICc (corrected Akaike Information Criterion)	27481.3	27641.5	27627.5
SBC (Schwarz Bayesian Criterion)	27509.1	27663.4	30008.7
Best fit parameters	(0.752529, -0.0907261, 0.232822)	(0.773028, 0.0889889)	*
Error Variance	12495.5	13211.5	9381.54

Figure 22 Parameters and results of all 3 timeseries model applied

Figure 22 displays MA model has comparatively less AIC value and error variance which indicates a better goodness-of-fit [27]. The contrast between testing values and predicted values in *Figure 19*, *Figure 20* and *Figure 21* displays highly inadequate attempt to forecast the trending peak of the conflict score. The difference between the lines is considerable which states a substandard model fit. Based on *Figure 22* and final predictions, it can be concluded that none of the models are able to capture a significant amount of the variability in the data, the residuals are relatively large. and therefore, cannot be used for conflict prediction at any interval [28].

5. RESULTS

5.1 LOCAL

Trending Events

With the help of revised Impact scale *AEI* (Anticipated Event Impact), most positive and negative resulting events over any location can be filtered.

Event	Location	AEI
https://santamariatimes.com/news/national/govt-and-politics/judge-orders-more-trump-lawyer-testimony-in-mar-a-lago-probe/article_2eadb810-3712-5e2d-ae66-5cc3cdd9836a.html	Washington, District of Columbia, United States	10
https://mynorthwest.com/3859859/tejano-musician-fito-olivares-dies-in-houston-at-75/	Houston, Texas, United States	7.497940612
https://menafn.com/1105805639/Manhattan-Neighborhood-Network-To-Honor-Trailblazer-Ralph-Mcdaniels-During-Celebration-Of-New-Multimedia-Facility	New York, United States	7.317738309
https://ktvz.com/politics/cnn-us-politics/2023/03/17/trump-attorney-ordered-to-testify-before-grand-jury-investigating-former-president/	Florida, United States	6.589896588
https://vancouversun.com:443/entertainment/celebrity/dark-moments-sam-neill-receives-treatment-for-blood-cancer/wcm/a302ffc5-fc04-4029-b2f6-15a9ec2f1f9c	Hollywood, California, United States	6.525095562
https://www.kob.com/news/us-and-world-news/trump-expects-to-be-arrested-tuesday-as-da-eyes-charges/	Manhattan, New York, United States	-6.769665233
https://news.wgcu.org/government-politics/2023-03-17/local-students-travel-to-tallahassee-to-testify-against-abortion-ban	Florida, United States	-6.909290005
https://torontosun.com/news/world/accused-wife-killer-made-prophetic-joke-as-family-feud-contestant	New York, United States	-7.469053714
https://santamariatimes.com/news/national/govt-and-politics/judge-orders-more-trump-lawyer-testimony-in-mar-a-lago-probe/article_2eadb810-3712-5e2d-ae66-5cc3cdd9836a.html	Washington, District of Columbia, United States	-8.769432816
https://www.durangoherald.com/articles/wyoming-governor-signs-measure-prohibiting-abortion-pills/	Wyoming, United States	-10

Figure 23 Top 10 trending events based on most negative and positive AEI scale.

Figure 23 provides us with top 10 trending events that occurred on March 18 2023, in United States of America (USA) over mentioned states which was our sample data for the research.

5.2 GLOBAL

Alert System

Based on performance and results in time series analysis in methodology, it gives a reason that resulting statistical model used are not reliable and therefore a need arises for a manual alert mechanism which can warn about probable war like situation. Slope pattern for multiple highest conflict peaks provides us with this approach.

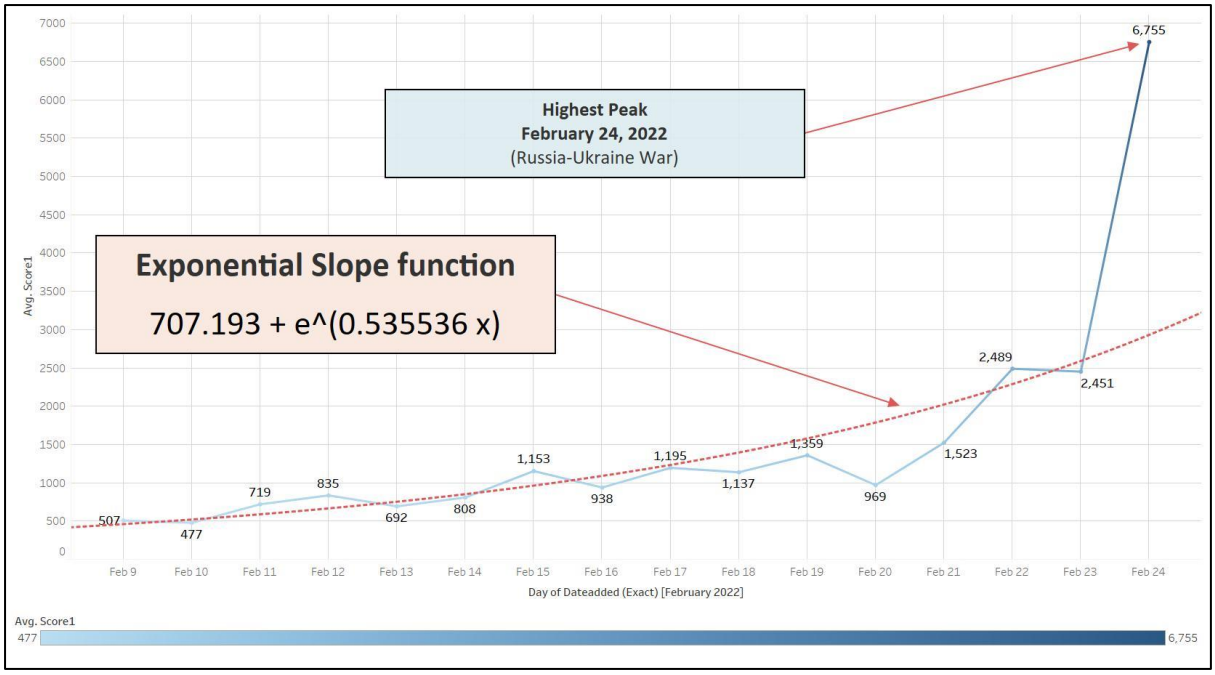


Figure 24 Exponential slope function fit over 15 days before day of war (February 24, 2022)

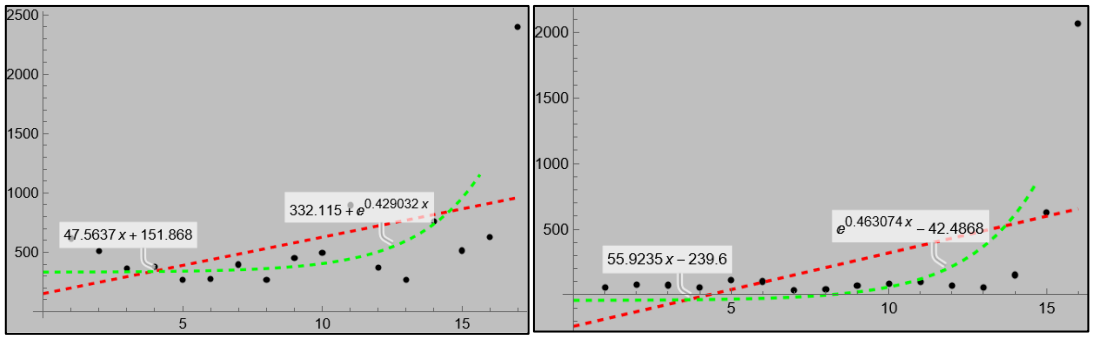


Figure 25 2nd highest and 3rd highest peaks with slope funtions

Peaks	Slope function	R-Squared value
1st highest (2022/02/24)	-373.9 + 220.51 x	0.476749
	707.193 + $e^{(0.535536\ x)}$	0.95536
2nd highest (2014/7/17)	151.868 + 47.5637 x	0.229011
	332.115 + $e^{(0.429032\ x)}$	0.849437
3rd highest (2018/11/26)	-239.6 + 55.9235 x	0.275344
	-42.4868 + $e^{(0.463074\ x)}$	0.839037

Figure 26 Linear and Exponential slope functions with fit comparision

Figure 26 suggests that after fitting linear and exponential fit, the latter one gives better results. Moreover, *Figure 24* shows Exponential function slope with high R-Squared value which indicates a good fit. That provides us with enough evidence to assume exponential slope function of highest peaks as base of our alert system.

$$a = Slope\ coefficient$$
$$b = Intercept$$
$$c_n = day\ of\ n^{th}\ highest\ conflict\ peak$$
$$s_{c_n} = conflict\ score\ of\ n^{th}\ highest\ conflict\ peak$$

$$p = day\ of\ Interest\ (prediction\ day)$$

xv

$$P = (p, p-1, p-2, \ldots, p-10)$$

xvi

$$x_n = (s_{c_n}, s_{c_n-1}, s_{c_n-2} \ \ldots \ s_{c_n-14}, s_{c_n-15})$$

xvii

$$S_n = f(x_n) = b + e^{ax_n}$$

xviii

$$S_{avg} = \frac{S_1 + S_2 + \ldots + S_n}{n}$$

xix

$$S_P = f(P) = b + e^{ap_n}$$

xx

$H_0 : S_{avg} > S_P : Expectation\ of\ war\ like\ situation\ is\ low$
$H_a : S_{avg} =< S_P : Expectation\ of\ war\ like\ situation\ is\ very\ high$

xxi

As shown in *Equation xix*, the intercept plays a major role as it sets a threshold value based on normal level ongoing events which is different for different countries. For example, slope equation for countries with high peace index like Canada or Bhutan would not work for countries drowned in conflicts like Syria or South Sudan, and will result in setting off false alarms or worse other way around resulting in no alert warnings. *Equation xviii* and *Equation xix* exhibits 5 day lag between period for used for calculation of conflict slope and observation slope respectively to get an early warning for a conflict peak.

DISCUSSION

Our study based on Local level gives us an idea about the trend on a state level and a necessity for a more reworked impact scale contemplating other significant variable from connecting datasources. Conflict analysis on the global level raises a need for a alert mechanism about high rising material conflicts. Moreover, Future project includes

- Working with GDELT 2.0 on city level to accomplish more defined trend and pattern by cross comparing GDELT with attributes like household income and size data, occupation data, race & ethinicity data.

- Text recognisation and sentiment analysis using nth dimentional word vectors of online conflict articles available through *SOURCEURL* via GDELT to summarise impact level on deeper level.

- Automated monitoring, reporting and alerting system by conflit peak pattern analysis hosted on cloud, easily accessible to people.

References

[1] K. Leetaru and P. A. Schrodt, “GDELT: Global data on events, location, and tone,” *ISA Annual Convention*, 2013, [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.686.6605

[2] “The GDELT Project.” https://www.gdeltproject.org/ (accessed Jan. 16, 2023).

[3] P. A. Schrodt, “CAMEO Conflict and Mediation Event Observations Event and Actor Codebook.” [Online]. Available: http://eventdata.psu.edu/

[4] P. A. Schrodt, “TABARI Textual Analysis by Augmented Replacement Instructions Version 0.8.4,” 2014, Accessed: Jan. 17, 2023. [Online]. Available: http://eventdata.parusanalytics.com/

[5] K. H. Leetaru, “Fulltext geocoding versus spatial metadata for large text archives: Towards a geographically enriched wikipedia,” *D-Lib Magazine*, vol. 18, no. 9–10, Sep. 2012, doi: 10.1045/SEPTEMBER2012-LEETARU.

[6] N. B. Weidmann and M. D. Ward, “Predicting conflict in space and time,” *Journal of Conflict Resolution*, vol. 54, no. 6, pp. 883–901, Jul. 2010, doi: 10.1177/0022002710371669/SUPPL_FILE/DS_10.1177_0022002710371669.ZIP.

[7] N. B. Weidmann and M. D. Ward, “Erratum to: Predicting Conflict in Space and Time(Journal of Conflict Resolution, (2010), 54, 6 (883-901), 10.1177/0022002710371669),” *Journal of Conflict Resolution*, vol. 55, no. 2, p. 321, Apr. 2011, doi: 10.1177/0022002711405502.

[8] J. E. Yonamine, “A NUANCED STUDY OF POLITICAL CONFLICT USING THE GLOBAL DATASETS OF EVENTS LOCATION AND TONE (GDELT) DATASET,” 2013.

[9] J. S. Goldstein, Goldstein, and J. S., “A Conflict-Cooperation Scale for WEIS Events Data,” *Journal of Conflict Resolution*, vol. 36, no. 2, pp. 369–385, 1992, doi: 10.1177/0022002792036002007.

[10] D. Csala, “Insurgent Dynamics: A systematic analysis of social unrest using the GDELT Event database Sustainable Energy Transitions View Project Sustainable Aviation View project,” 2015, doi: 10.13140/RG.2.1.1095.9526.

[11] S. Keertipati, B. T. R. Savarimuthu, M. Purvis, and M. Purvis, “Multi-level analysis of peace and conflict data in GDELT,” in *ACM International Conference Proceeding Series*, Association for Computing Machinery, Dec. 2014, pp. 33–40. doi: 10.1145/2689746.2689750.

[12] J. Villalobos Alva, “Beginning Mathematica and Wolfram for Data Science,” *Beginning Mathematica and Wolfram for Data Science*, 2021, doi: 10.1007/978-1-4842-6594-9.

[13] S. Challawala, “MySQL 8 for Big Data,” p. 266, 2017.

[14] A. Khan, “Jumpstart Tableau : A Step By Step Guide to Better Data Visualization,” *Apress*, p. 388, 2016.

[15] “Data: Querying, Analyzing and Downloading: The GDELT Project.” https://www.gdeltproject.org/data.html (accessed Jan. 31, 2023).

[16] “All GDELT Event Files.” http://data.gdeltproject.org/events/index.html (accessed Feb. 10, 2023).

[17] “Crime in the U.S.: Key questions answered | Pew Research Center.” https://www.pewresearch.org/fact-tank/2020/11/20/facts-about-crime-in-the-u-s/ (accessed Feb. 15, 2023).

[18] “90% of Americans expect 2023 will be year of political conflict in US: Poll.” https://www.aa.com.tr/en/americas/90-of-americans-expect-2023-will-be-year-of-political-conflict-in-us-poll/2779002 (accessed Feb. 26, 2023).

[19] “CorrelationTest—Wolfram Language Documentation.” https://reference.wolfram.com/language/ref/CorrelationTest.html (accessed Mar. 01, 2023).

[20] R. H. Best, C. Carpino, and M. J. C. Crescenzi, “An analysis of the TABARI coding system,” *http://dx.doi.org/10.1177/0738894213491176*, vol. 30, no. 4, pp. 335–348, Jul. 2013, doi: 10.1177/0738894213491176.

[21] “Alert 2021! Report on conflicts, human rights and peacebuilding.”

[22] U. Nations, “United Nations | Peace, dignity and equality
on a healthy planet”, Accessed: Mar. 01, 2023. [Online]. Available: https://www.un.org/en/

[23] “AutocorrelationTest—Wolfram Language Documentation.” https://reference.wolfram.com/language/ref/AutocorrelationTest.html (accessed Mar. 07, 2023).

[24] “UnitRootTest—Wolfram Language Documentation.” https://reference.wolfram.com/language/ref/UnitRootTest.html (accessed Mar. 07, 2023).

[25] “plotting - Time-series decomposition in Mathematica - Mathematica Stack Exchange.” https://mathematica.stackexchange.com/questions/16723/time-series-decomposition-in-mathematica (accessed Mar. 07, 2023).

[26] I. Rojas, H. Pomares, O. Valenzuela, and S. International Work-Conference on Time Series (2017 : Granada, “Time series analysis and forecasting : selected contributions from ITISE 2017,” p. 340.

[27] “TimeSeriesModelFit—Wolfram Language Documentation.” https://reference.wolfram.com/language/ref/TimeSeriesModelFit.html (accessed Mar. 08, 2023).

[28] “TimeSeriesForecast—Wolfram Language Documentation.” https://reference.wolfram.com/language/ref/TimeSeriesForecast.html (accessed Mar. 08, 2023).