Predicting Socio Economic Indicators using NEWS Event

Hemang Akbari - 20162112
Mohsin Mulani - 20162053
Rohan Adesara - 20162127
Naveen Sharma - 20162107

Guided by : Prof Naresh Manwani
Mentor : Vaibhav Kumar

Introduction:

ARIMA + EVENT_BASED_PREDICTION_FUNCTION

This experiments demonstrate that incorporating event information in the prediction tasks reduces the root mean square error (RMSE) of prediction by 22% compared to the Standard ARIMA model.

Equation To Implement:

(1) Predict price (y):

$$y_t = \epsilon_t + \alpha_1 y_{t-1} + \ldots + \alpha_p y_{t-p} - \beta_1 e_{t-1} \ldots - \beta_q e_{t-q} \quad (1)$$

$$+ \sum_{k=1}^{K} \omega_t^k \phi_{tk} + \sum_{k=1}^{K} \omega_{t-1}^k \phi_{(t-1)k} + \ldots + \sum_{k=1}^{K} \omega_{t-\delta}^k \phi_{(t-\delta)k}$$

Consider a corpus D of news articles indexed by time t, so that Dt is the collection of news articles published at time t.

The news articles report real-world events and we suppose that the total number of events reported in the corpus is some fixed but unknown K.

there is some function $\varphi t: Dt \rightarrow pow([0,1], K)$ that maps a collection of news articles published at certain time t, to a vector $\varphi t(Dt) = (\varphi t1, \varphi t2, ..., \varphi tK)$ that specifies the “intensity” of each of the K events at time instant t. In other words, larger the value of φtk, more is the proportion of event $k \in [K] := \{1,2,...,K\}$ in corpus D.

(2) Spike prediction:

A spike is defined as a sudden change in the value of y_t from its previous value y_{t-1}.

For that here SVM based binary classifier is used.

EVENT CLASS MODEL

For any generic document - topics are important

For news articles - events are important

We focus on the “action words” that are representative of incidents reported in the article.

Event triggers are a set of words or phrases that describe an action between entities or some incident within text.e.g. “protesting”, “flooded” etc.

EVENT CLASS MODEL(Contd.)

Event class represented using a collection of related event triggers summarizing that category of events.

In essence, event classes encapsulate synonymous words to represent similarly themed events. We use these definitions of event class and event triggers to model events reported in a large collection of news articles.

Based on the typical structure of a news article, the information to be conveyed to the readers is usually mentioned in the title and the lead (first) paragraph of the article.

EVENT CLASS MODEL(Contd.)

Thus, we consider the triggers found in the title or the lead paragraph to be an indicator of the underlying event class, the central event of the article is drawn from.

A news article sampled from an event class is an instance of that class this instance is called an event.

For example, “accident” is an event class whereas a specific occurrence of an accident reported in an article is an event.

In this example, the trigger is “accident” but other words or phrases, e.g. crash, collision, rammed etc., can also replace this trigger without losing the essence of the event class.

EVENT CLASS MODEL(Contd.)

Subsidiary events are events mentioned in an article in addition to the main event of the article.It represents the additional events likely to happen along with the main event.

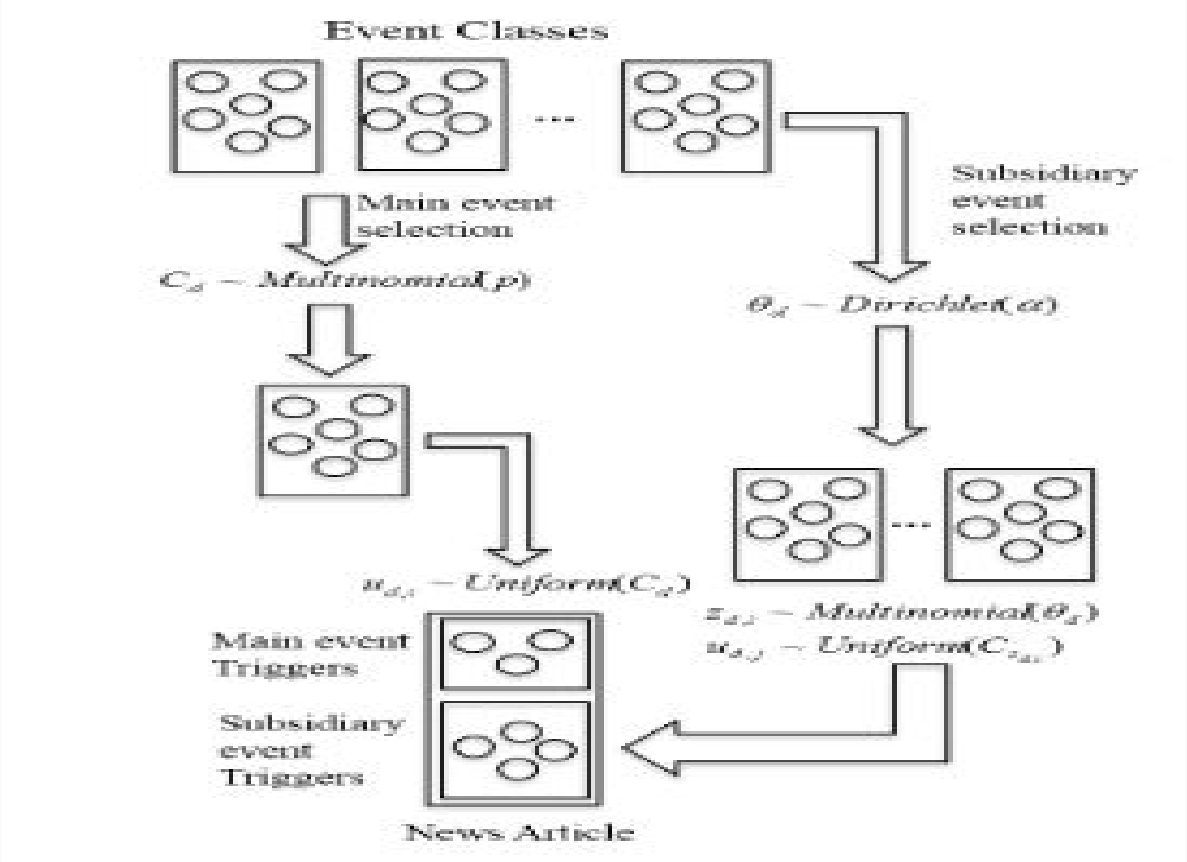
Consider the title “Blasts at Boston Marathon Kill 3 and Injure 100".

The event triggers in this title are “blast”,“kill” and “injure”. Clearly, “blast” is the central event in this article (which represents the event class related to blasts, explosion,bombing) but additional triggers (kill and injure) are two events that are closely associated with the central event.

These additional events as subsidiary events.

Generative Model of News Articles

Variables/ Parameters	Description
D	Number of news articles
K	Number of event classes
C_k	Event class k represented as a set of event triggers describing the class
p_k	Prior probability of event class C_k, specifies which event classes are more frequent
U_d	Event triggers present in article d
C_d	Event class from which main event in article d is generated
$\boldsymbol{\alpha}$	Dirichlet prior to generate subsidiary events
$\boldsymbol{\theta}_d$	The proportion of event classes as subsidiary events in article d
z_{dj}	Which event class produced the j^{th} subsidiary trigger in article d



Event class model and generation of news articles.

Event Trigger Extraction

Automatic Content Extraction (ACE)

8 class type / 33 class sub-type available, not much useful for our application

Conditional Random Field (CRF)

$$F_i(\boldsymbol{o}, \boldsymbol{x}) = \exp(\sum_j \lambda_j t_j(o_{i-1}, o_i, \boldsymbol{x}, i) + \sum_k \mu_k s_k(o_i, \boldsymbol{x}, i)) \quad (2)$$

$$\Pr(\boldsymbol{o} \mid \boldsymbol{x}; \boldsymbol{\lambda}, \boldsymbol{\mu}) = \frac{1}{Z(x)} \exp(\sum_i \log F_i(\boldsymbol{o}, \boldsymbol{x}))$$

Constructing Event Class

<explosions, bombing, blasting> are similar kind of event triggers.

NN based language model - to embed each word into vector space, with similar words nearby.

K-means to find optimal K.

E.g. K = 250 observed for total no. of event classes.

Event Driven Prediction

How is Phi_{tk} and y_t are found?

$$\phi_{tk} = \frac{\sum_{d \in \mathcal{D}_t} \mathbb{1}[C_d = k]}{|\mathcal{D}_t|}$$

$$y_t = \omega_t^0 + \sum_{k=1}^{K} \omega_t^k \phi_{tk} + \epsilon_t$$

Where y_t is the socio-economic indicator,

and 1l is the Iversion function.

Two Models

Historical

$$y_t = \omega_t^0 + \sum_{j=0}^{\delta} \sum_{k=1}^{K} \omega_{t-j}^k \phi_{(t-j)k} + \epsilon_t$$

Linear model that gives the prediction based on the time window *delta*.

Topic Driven

$$\phi_t(\mathcal{D}_t) = (\theta_t^1, \theta_t^2, \ldots, \theta_t^{K'})$$

$$\theta_t^i := \max_{d \in \mathcal{D}_t} \theta_{di}$$

Gives the most relevant topic in each article *d* from the corpus *D*.

Use case:Food Price Prediction

- We evaluated our model by demonstrating its ability in predicting the value of socio economic indicators.
- Events extracted from news article based on event model are used as features
- Food price fluctuation as an example scenario to demonstrate how event based predictive model can be built.
- Event driven prediction model built on top of ARIMA.
- For each event class the event class triggers are extracted from the news article.

Cont..

Event class triggers	Subsidiary event triggers
molest,kill,eliminate manhandle,kidnap abduct	including,denied,eliminated,killed left,set,chopped,elected, escalating, estimated, expressed
accused,suspects,killers,kingpin, conspirators,masterminded	arrested,found, told, raped, filed , registered, alleged, claimed, including
supporting,allies,backing, marxists,criticising,tacit	added, activists,advised,armed, arrested, attended, concerned, engaged, extending, found
drought,flood,worst, tsunami,situation,cyclone	provide, pump, added, adding,aired, allocated, announced, apathy, arrive, aila, assumed, beating, changing
campaigning,canvassing, mayoral,pitching,lobbying campaigned	ensure, campaigning, premises, canvassing, closed, conducted, including, leaving, prohibited, taking
capture,decode,recreate, propagate,arouse,ignite	managed, make, project, ruled, alleged, appealed, attacked, based, bored, capture, caste, change
gained,emerged,lost, boosted,transformed,demonstrated	purchased, exported, lift, ranging, reap, added, attached, availed, districts, enabled, fallen, growing
blast,bomb,malegaon, explosions,bakery,defusing	arrested, injured, sought, accused, made, picked, added, demanded, file, killed, occurred, found, involved,
protest,demonstration,protests, agitation,dharna,strike	held, staged, demanded, added, pay, protest, decided, told, alleged, died, proposed, protesting, submitted,

Example of event class extracted from news article

Printed from
THE TIMES OF INDIA

Heavy rain, hailstorms destroy crops in north India

TNN | Mar 17, 2015, 06.38 AM IST

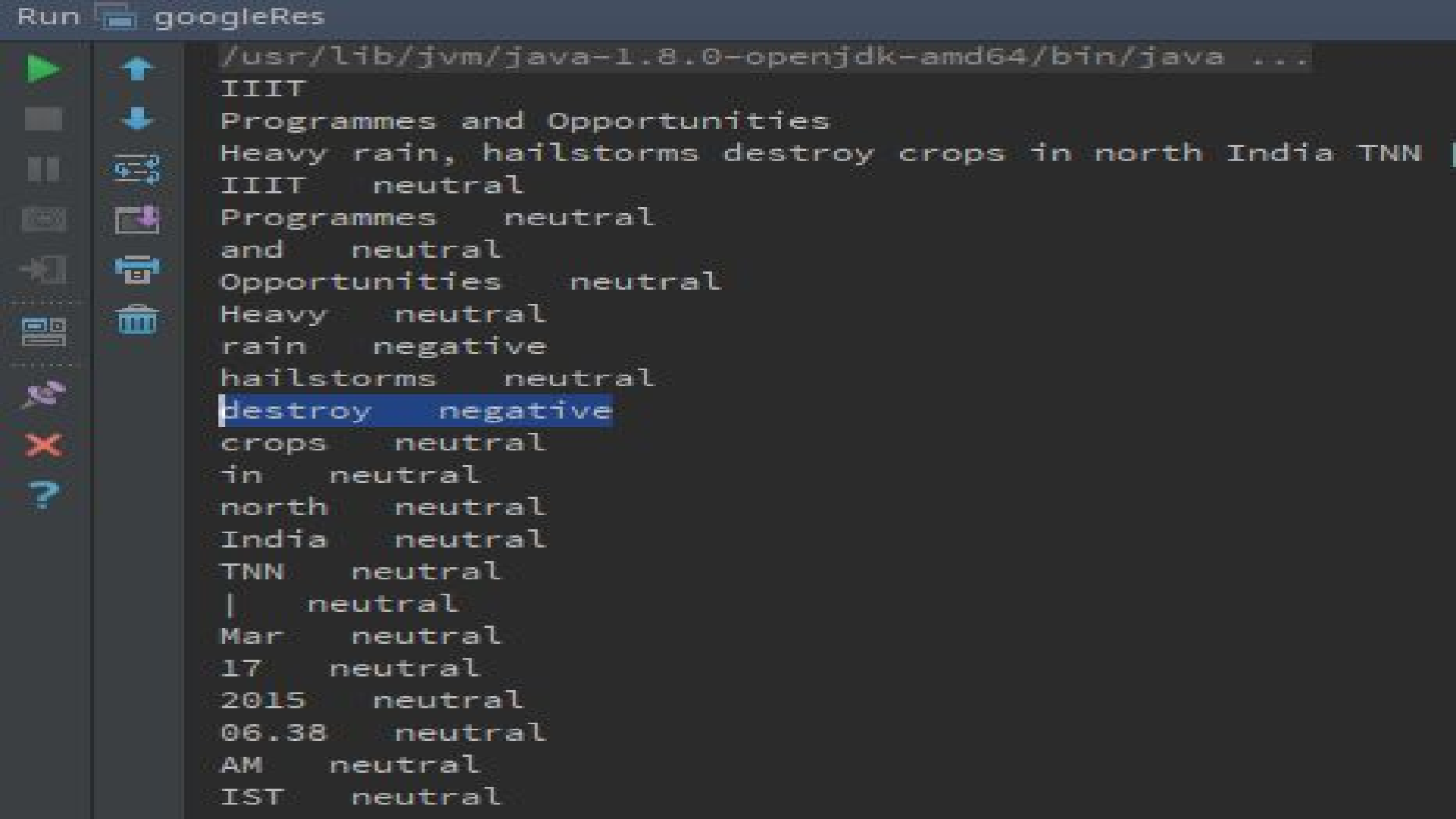
Unseasonal thundershowers and hailstorms left behind a trail of destruction, leveling standing crops across swathes of north India on Sunday, with the region still reeling under its effect on Monday even as Central authorities tried to assess the full extent of the cumulative losses.

Wheat, pulses, mustard, and gram took the brunt of sudden precipitation in east UP; Punjab, Haryana, Rajasthan, Madhya Pradesh, Uttar Pradesh and Maharashtra witnessed similar devastation. Landslides and snowfall led to the closure of the Jammu & Kashmir highway leaving thousands of people stranded.

UP farmers said they suffered crop losses of over 50% prompting chief minister Akhilesh Yadav to release Rs 200 crore from the state's emergency funds.

"Rains destroyed over 50% crops of wheat, mustard, pea and gram," lamented 75-year-old farmer Lalchand Patel of Jayapur village that was adopted by the Prime Minister Narendra Modi in May 2014.

A sample news article


Run
googleRes
/usr/lib/jvm/java-1.8.0-openjdk-amd64/bin/java ...
IIIT
Programmes and Opportunities
Heavy rain, hailstorms destroy crops in north India TNN |
IIIT neutral
Programmes neutral
and neutral
Opportunities neutral
Heavy neutral
rain negative
hailstorms neutral
destroy negative
crops neutral
in neutral
north neutral
India neutral
TNN neutral
| neutral
Mar neutral
17 neutral
2015 neutral
06.38 neutral
AM neutral
IST neutral

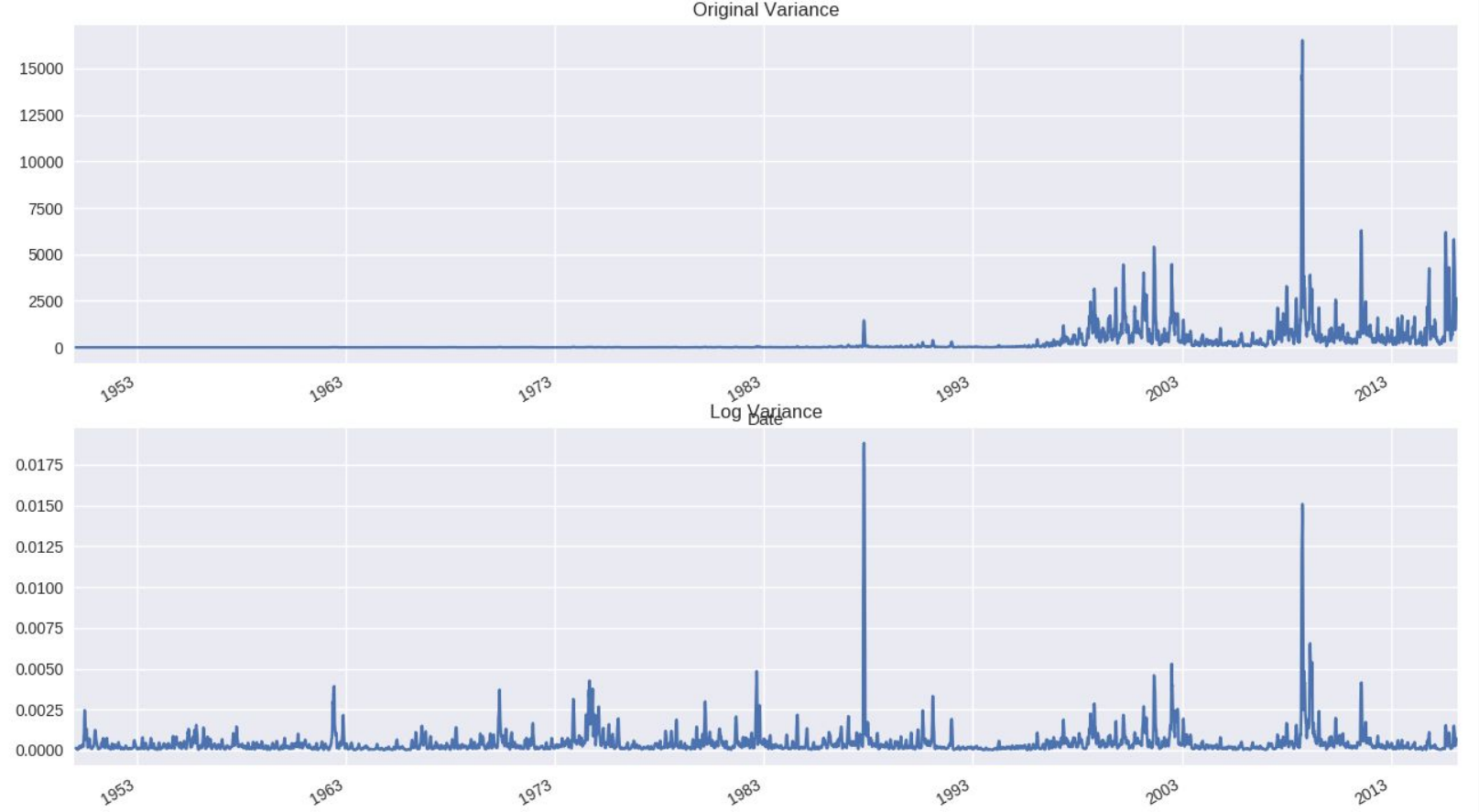
Problem with this approach

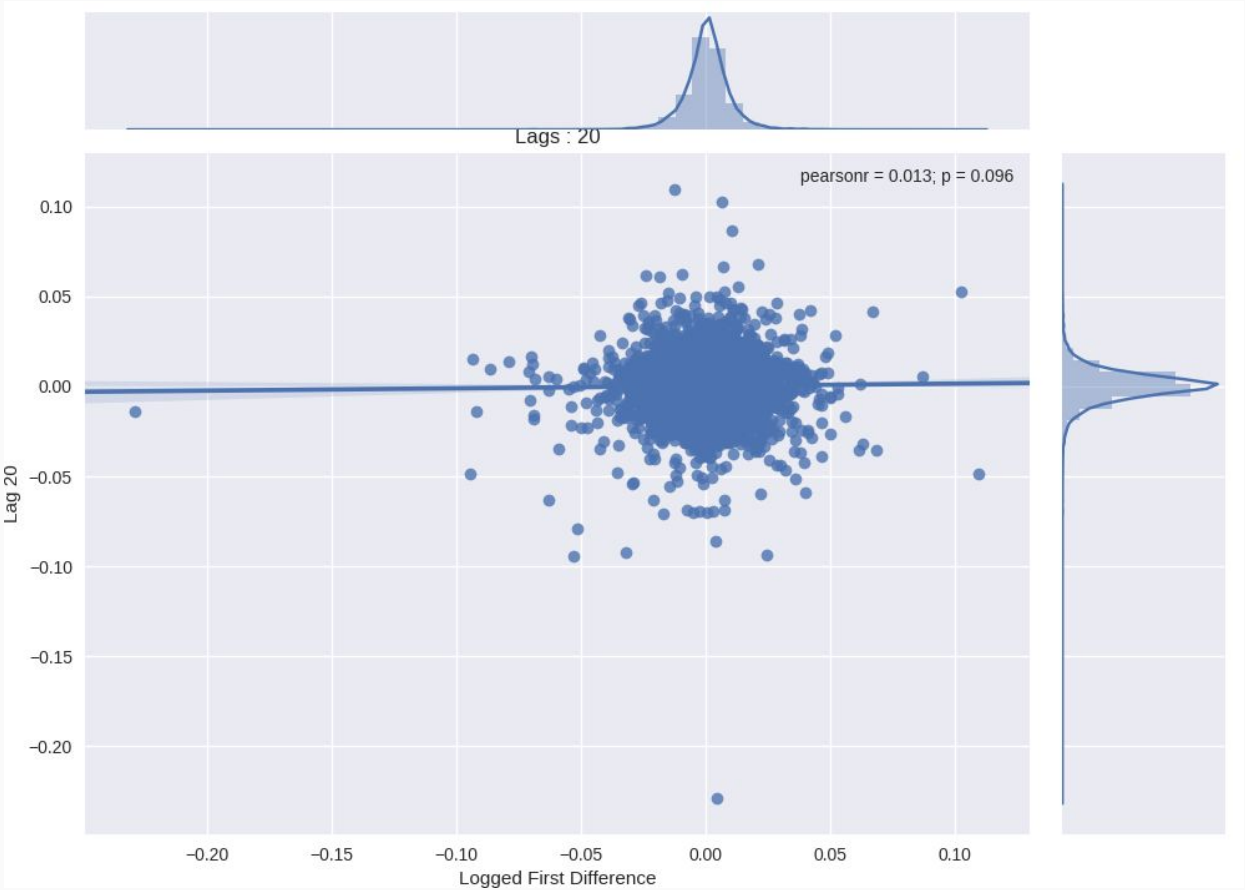
- It is ambiguous to decide if a event trigger will contribute in rise of price or not

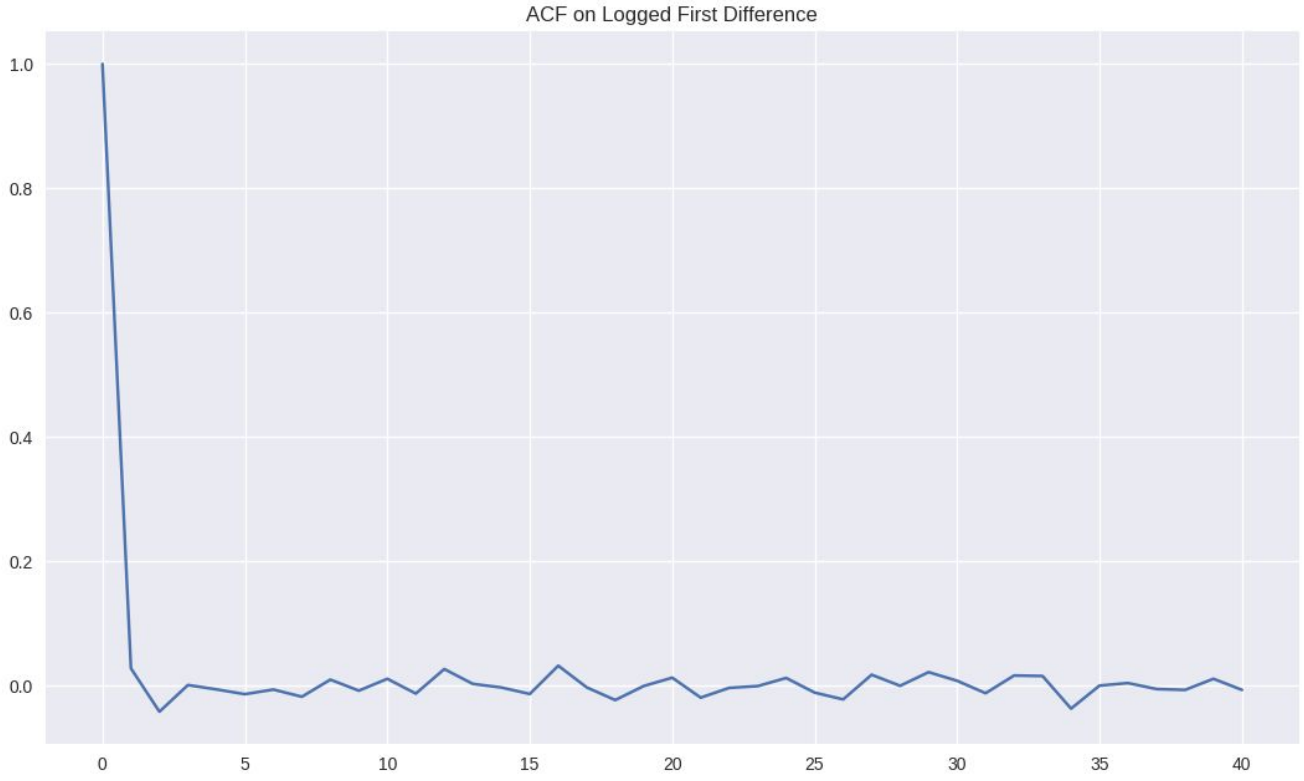
  Eg. Heavy Rain can be in both good and bad sense.

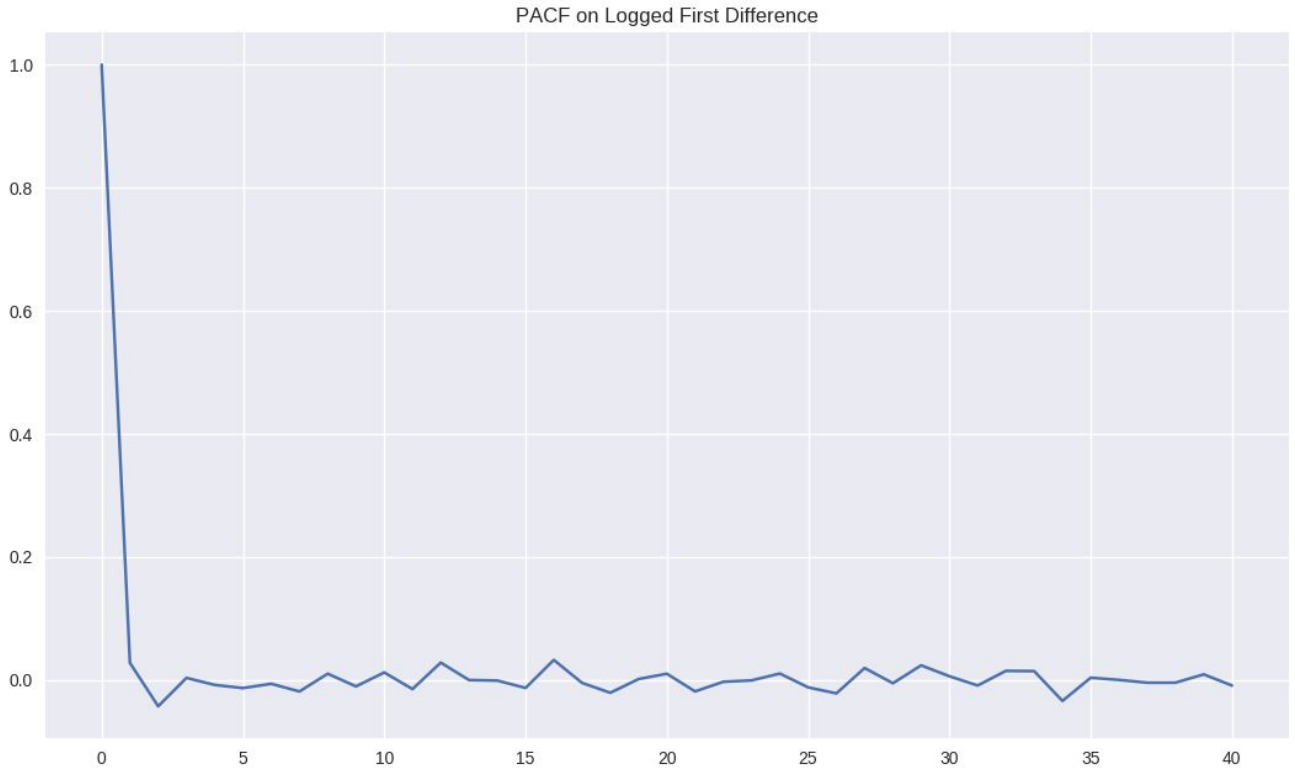
- Solution:Writing a context sensitive Grammar to capture meaning of whole sentence.
- Approach we used

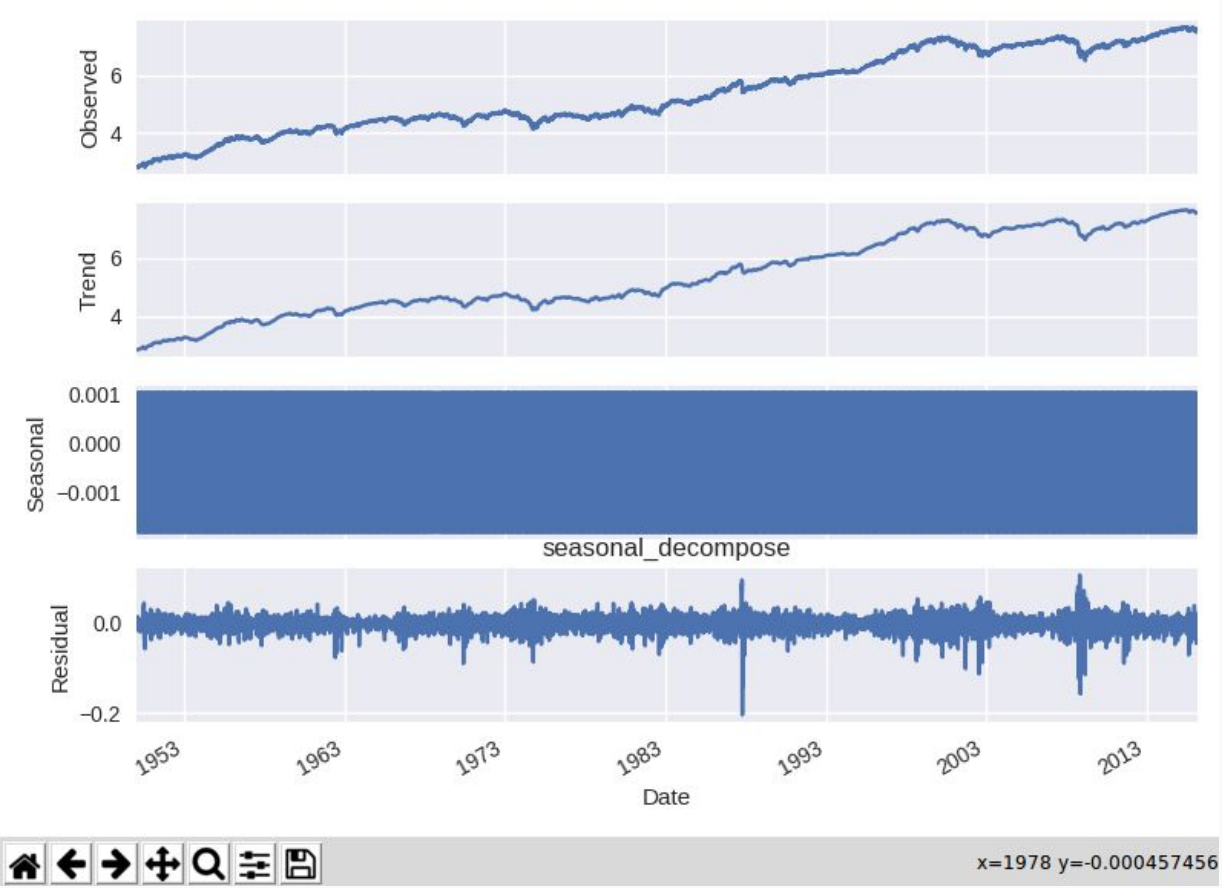
  ->We implemented it using arima model.

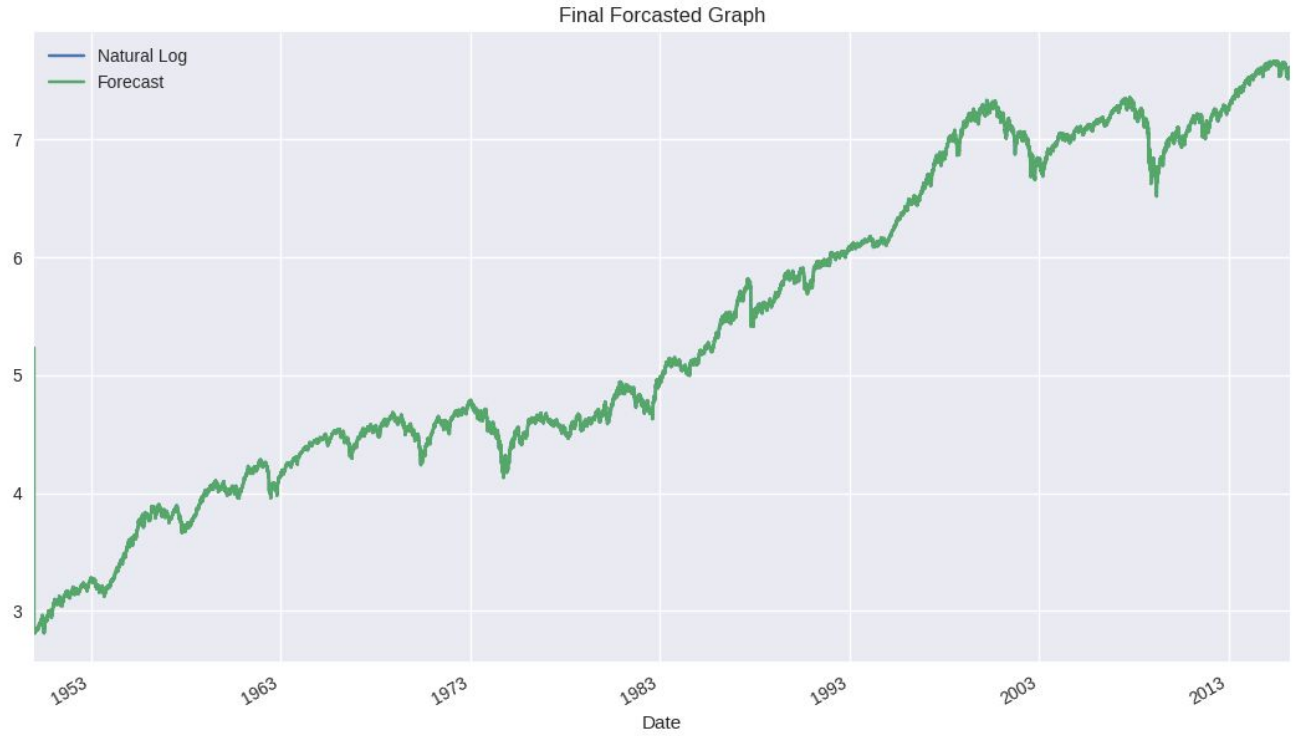

Original Variance
15000
12500
10000
7500
5000
2500
0
1953
1963
1973
1983
1993
2003
2013
Log Variance
Date
0.0175
0.0150
0.0125
0.0100
0.0075
0.0050
0.0025
0.0000
1953
1963
1973
1983
1993
2003
2013

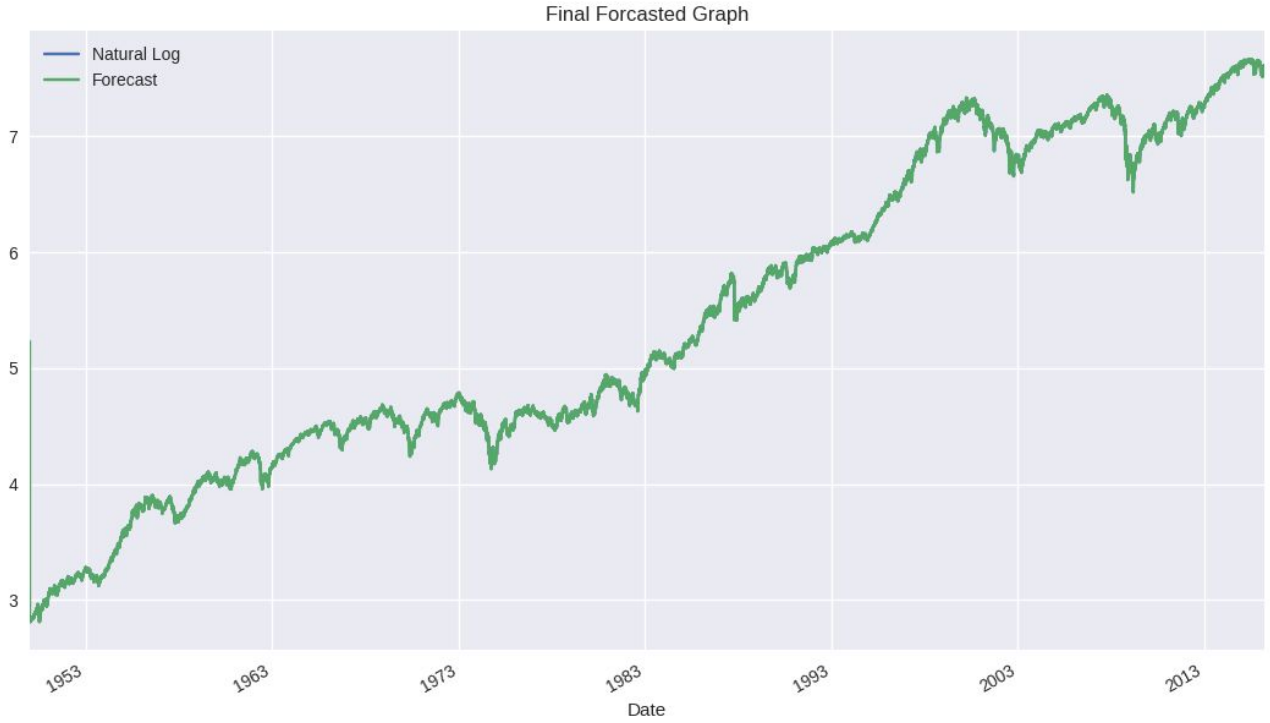

Lags : 20
pearsonr = 0.013; p = 0.096
0.10
0.05
0.00
−0.05
−0.10
−0.15
−0.20
Lag 20
−0.20
−0.15
−0.10
−0.05
0.00
0.05
0.10
Logged First Difference


ACF on Logged First Difference
1.0
0.8
0.6
0.4
0.2
0.0
0
5
10
15
20
25
30
35
40


PACF on Logged First Difference
1.0
0.8
0.6
0.4
0.2
0.0
0
5
10
15
20
25
30
35
40


Observed
6
4
Trend
6
4
Seasonal
0.001
0.000
−0.001
seasonal_decompose
Residual
0.0
−0.2
1953
1963
1973
1983
1993
2003
2013
Date
x=1978 y=-0.000457456


Final Forcasted Graph
Natural Log
Forecast
7
6
5
4
3
1953
1963
1973
1983
1993
2003
2013
Date


Final Forcasted Graph
Natural Log
Forecast
7
6
5
4
3
1953
1963
1973
1983
1993
2003
2013
Date

Conclusion

- We presented a novel way of defining and extracting events from a large news corpus.

- We predicted the stock price using arima model which can be extended to use event driven model.

References

Predicting Socio-Economic Indicators using News Events
http://cs.nyu.edu/~sunandan/event-predict.pdf

[ARIMA Models](http://people.duke.edu/~rnau/411arim.htm)

[SentiWord 3.0.txt]

https://github.com/ekanshpreet/data-portraits/blob/master/extlib/SentiWordNet/SentiWordNet_3.0.0.txt