Predicting Socio-Economic Indicators using News Events

Mohsin Mulani[1] Hemang Akbari[2]
Naveen Sharma[3] Rohan Adesara[4]

1. Abstract :

Many socio-economic indicators are sensitive to real-world events.Proper characterization of the events can help to identify the relevant events that drive fluctuations in these indicators. In this paper,we propose a novel generative model of real-world events and employ it to extract events from a large corpus of news articles. We introduce the notion of an event class,which is an abstract grouping of similarly themed events. These event classes are manifested in news articles in the form of event triggers which are specific words that describe the actions or incidents reported in any article. We use the extracted events to predict fluctuations in different socioeconomic indicators.

New Model : 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.

2. Introduction

Socio-economic indices such as commodity prices have shown high volatility in several countries over the last few years.Estimation of socio-economic indices naturally relies on information from multiple sources.However, in existing studies aimed at estimating such variables ,analysis and forecasting was usually done using structured data sources, considering only a handful of driving factors, which were also chosen manually.We have build model upon the basic observation that real-world events, which manifest themselves in unstructured text streams such as news, blogs and social media, can provide strong signals of the underlying factors that drive fluctuations in socio-economic indicators.

Our work fundamentally differs from prior work on two fronts:
(a) we explicitly assume no knowledge about the specific events that lead to fluctuations in any given socio-economic index and aim to automatically discover them.
(b) we do not use any external knowledge base or data to build the predictive model, relying solely on the real-world events extracted.

3. TERMS USED IN IMPLEMENTATION

For any generic document - topic are important.
For news articles - events are important.

We focus on modeling only the underlying essence of any event, i.e. 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 is a broad category of events 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.

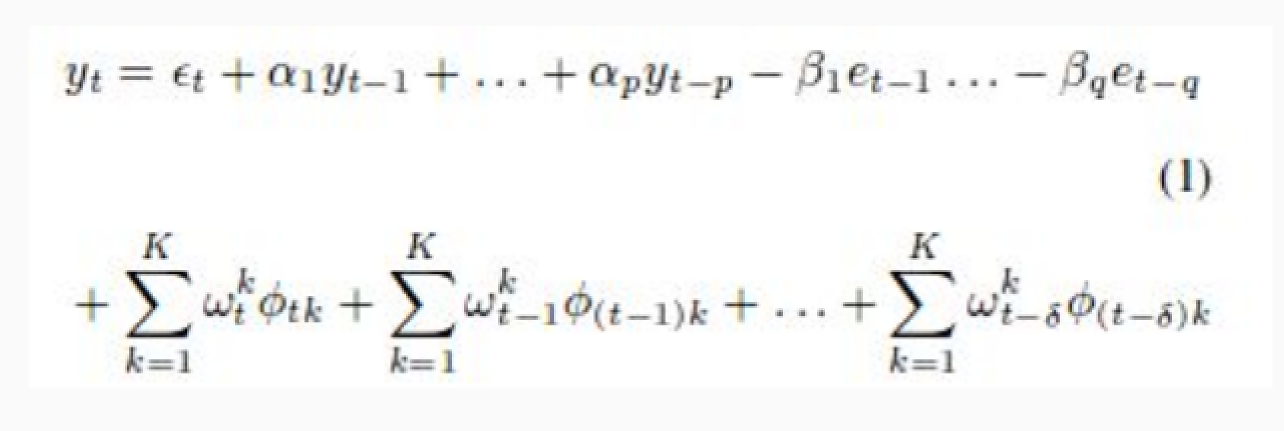
Based on the typical structure of a news article, the information to be conveyed to the readers is usually mentioned in the title and the lead (first) paragraph of the article.Thus, we consider the triggers found in the title or the lead paragraph to be an indicator of the underlying event class, the 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.

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.

4. Implementation

(1)Predict price (y):



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

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

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 φt:Dt→pow([0,1] ,K) that maps a collection of news articles published at certain time t, to a vector φt(Dt) = (φt1,φt2,...,φtK) 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 ∈ [K] :={1,2,...,K} in corpus D.

2) Spike prediction:

A spike is defined as a sudden change in the value of yt from its previous value yt−1.
For that here SVM based binary classifier is used

5. METHODOLOGY

Consider a corpus of news articles D and let the total number of articles be D . Suppose the set of all event triggers extracted from the news corpus be given by U = {u1,u2,...,uM} and let C ={C1,C2,...,CK} denote the event classes, where each Ci ⊂ U is a collection of event triggers and Ci ∩ Cj= ∅ and ∀i =j.

(1) Generative Model of News Articles : Identify which event class (Cd) the article d belongs to, by observing the words belonging to the set Ud and their positions,where Ud represents an event trigger present in article d.

(2) Event Trigger Extraction : We implemented a conditional random field (CRF) based supervised method to extract the event triggers from the news articles.CRFs are probabilistic models for labeling sequence data.In our setting, the training data is in the form of labeled sentences in news articles, where each word in the sentence has a label T if it is a trigger word or NT otherwise.

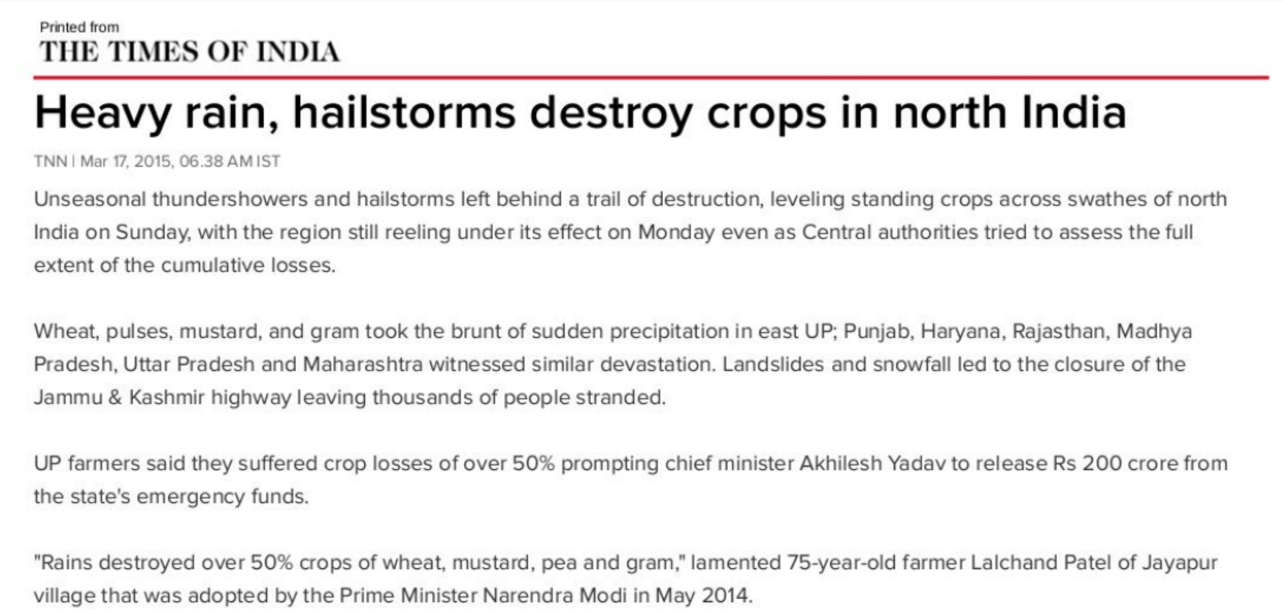
(3) Constructing the Event Class : The idea is to cluster “similar” news articles and obtain the event triggers that describe events belonging to the same event class.We use a neural network based language model to learn an embedding of each word.This technique embeds each word (or phrase) from a large corpus of text into a vector space where words appearing in very similar contexts are placed in the vicinity of each other.

(4) Event-driven prediction : We obtain a posterior distribution for the hidden event class C_d of each news article d.We assign C_d to the MAP estimate, i.e. choose the event class that has the maximum posterior probability for article d given the entire news corpus.Then, we define the proportion or intensity of event K at time t and it is called event vector φ_{tk} .The general formulation of the predictive model is:

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

where y_t represents the socio-economic indicator whose value is being predicted using the extracted event classes

6. RESULTS



Printed from
THE TIMES OF INDIA

Heavy rain, hailstorms destroy crops in north India

TNN | Mar 17, 2015, 06.38 AM IST

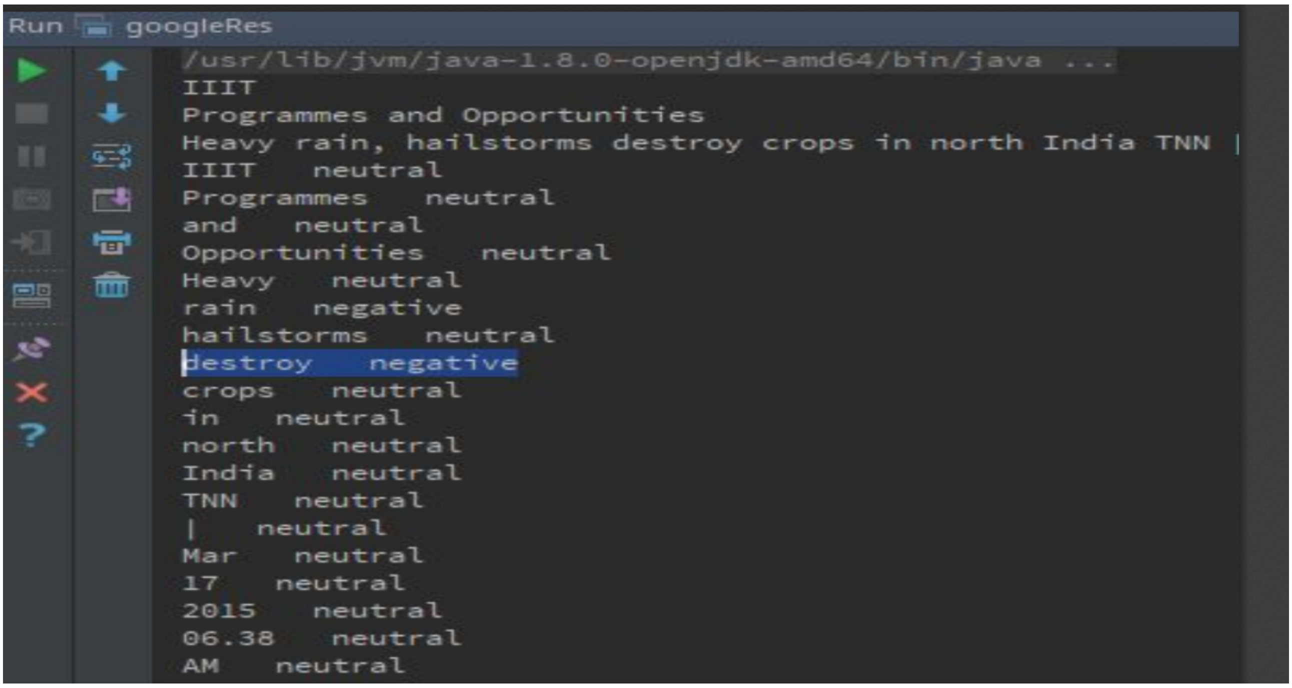
Unseasonal thundershowers and hailstorms left behind a trail of destruction, leveling standing crops across swathes of north India on Sunday, with the region still reeling under its effect on Monday even as Central authorities tried to assess the full extent of the cumulative losses.

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.

Figure : Sample news Article



```
/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
```

Figure : Each word is classified as positive,negative or neutral

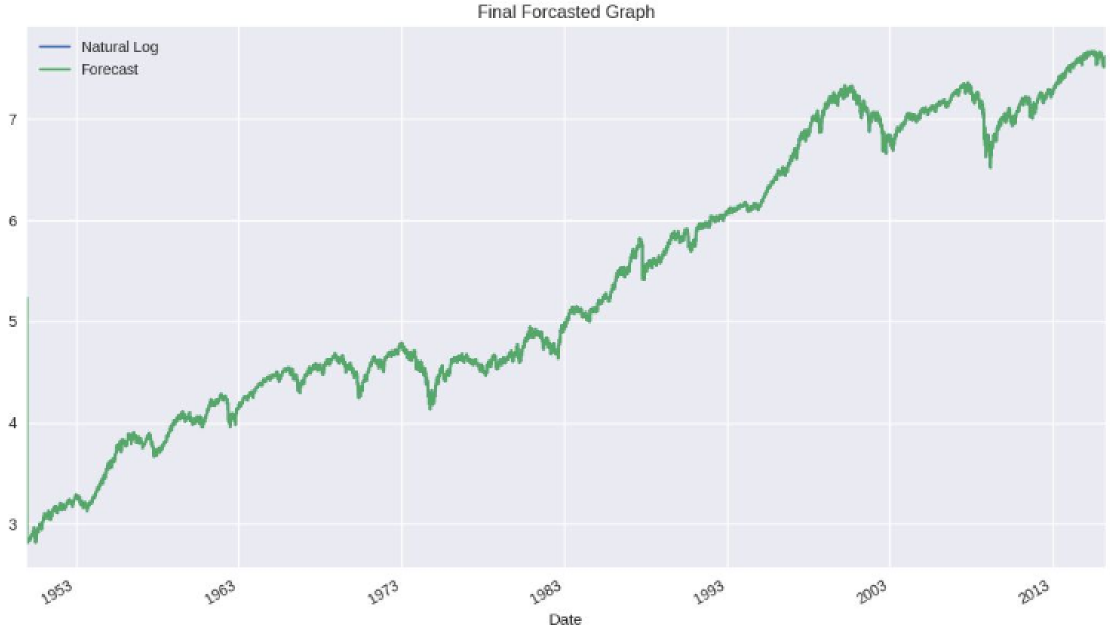


Figure : Final forecasted Graph