Predicting Socio-Economic Indicators using News Events

Sunandan Chakraborty[1,2], Ashwin Venkataraman[1], Srikanth Jagabathula[3],
Lakshminarayanan Subramanian[1,2]
[1]Department of Computer Science
[2]Center for Technology and Economic Development, NYU Abu Dhabi
[3]Leonard N. Stern School of Business
New York University, New York, USA
{sunandan, ashwin, lakshmi}@cs.nyu.edu, sjagabat@stern.nyu.edu

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 socio-economic indicators. Specifically, we focus on food prices and predict the price of 12 different crops based on real-world events that potentially influence food price volatility, such as transport strikes, festivals etc. Our 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. We also predict sudden increases in the food prices (i.e. spikes) using events as features, and achieve an average 5-10% increase in accuracy compared to baseline models, including an LDA topic-model based predictive model.

1. INTRODUCTION

Socio-economic indices such as commodity prices have shown high volatility in several countries over the last few years [13]. Unpredictable price fluctuations, especially of essential commodities, can have adverse effects such as food shortage, uneven distribution and consumption, reduced income among producers, supply chain issues etc. Understanding the factors that impact volatility of socio-economic indicators is a fundamental problem of interest for policy-making [28] and financial institutions [16][35].

Estimation of socio-economic indices naturally relies on information from multiple sources [14]. However, in existing studies aimed at estimating such variables [26], analysis and forecasting was usually done using structured data sources, considering only a handful of driving factors, which were also chosen manually. There may be unknown factors playing an important role in the indicators' volatility that may not be captured by such a manual process of selection. In addition, most known prediction mechanisms that have leveraged news content have primarily relied on predefined domain-specific features or market sentiment extraction mechanisms to predict variations in specific indices [14].

This paper builds 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 [16]. Numerous existing studies [31][35][39][40] have shown evidence of the impact of financial news on stock market prices, further supporting the observation. Motivated by this observation, in this paper we aim to predict fluctuations in socio-economic indicators by automatically extracting (representations of) real-world events from unstructured news sources. More specifically, the question we address is: Given a structured time series about some socio-economic index of interest and a large corpus of real-world events extracted from news sources, can we learn a predictive model for describing the fluctuations in the socio-economic index by identifying the relevant events that influence it? In the process, we also identify relationships between real-world event occurrences and variations in socio-economic indices – for instance, that festivals in India are associated with an increase in tomato and onion prices, or transport strikes called in a region are early indicators of rise in food prices in the region.

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.

In this paper, we propose a novel generative model to represent events surfacing in the news media, and define several event-driven predictive models for predicting fluctuations in socio-economic indicators. The models that we define are also flexible in the sense that they can be used together with existing mechanisms for forecasting socio-economic indices, such as time series analysis, to improve the prediction performance. We present one specific application in section 7. Our event model is based on the assumption that most news articles talk about only one central (or main) event, a fact observed in prior work [27]. Further, an event is usually associated with various entities, such as people, organizations, topics[1] as well as metadata like location and time. We introduce the notion of an *event class*, which represents an abstract grouping of similar events agnostic of spatio-temporal, entity or topic based features. In other words, once we strip all the extraneous details from an event, whatever is left describes the essence of the event and is termed as

[1] We use topic in an abstract sense here and not the technical sense in which it is defined in the topic modeling literature

KDD '16, August 13-17, 2016, San Francisco, CA, USA
© 2016 ACM. ISBN 978-1-4503-4232-2/16/08…$15.00
DOI: http://dx.doi.org/10.1145/2939672.2939817

the event class of the event. This is motivated by our belief that the factors that drive fluctuations in socio-economic indicators are more general, and specific details might bias the predictions. As an example, a transport strike happening in a region is already indication that a possible price hike might take place. The actual entities involved such as people names and other named entities, do not necessarily provide further evidence. Of course, it is possible that because of the presence of certain entities, such as political parties, the resultant effect of the strike is different because of certain steps/measures taken by the specific entities. But then those steps/measures would also be captured as a different event (more precisely event class) in our model, and therefore would influence the prediction in the end. Grouping the various events together also helps in addressing the sparsity issue mentioned in prior work [29] and helps in generalizing to unseen events more naturally.

Our event model can capture any kind of events, provided that there exists at least one article in the corpus that is about that event. So, our model overcomes two main limitations of existing frameworks – lack of flexibility and reliance on external knowledge bases and ontologies [29]. As our model is capable of extracting generic events from any corpus which are not specific to any domain or predefined class, our event-driven predictive models are also not restricted to predicting domain-specific variables and can, in principle, be used for any socio-economic indicator that is influenced by real-world events. This makes our framework broadly applicable to many scenarios.

We evaluate our event-driven predictive models to predict fluctuations in food prices for 12 popularly consumed food crops in India. Our 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. We also evaluate prediction of sudden increases in the food prices (spikes) using the extracted events as features. Our results show an average $5-10\%$ increase in accuracy compared to baseline models, including an LDA topic-model based predictive model.

2. RELATED WORK

Recent advances in text mining techniques have produced many low-dimensional representational schemes for text documents. Topic models such as Probabilistic Latent Semantic Indexing (PLSI) [15] and Latent Dirichlet allocation (LDA) [2] can be used to represent a large corpus of documents with a vocabulary $\sim$ 100,000 words using just a few hundred topics. These models have made knowledge acquisition from natural language text easier and more effective. In particular, online news articles are a popular source for mining real-world events using these low-dimensional representations. Trend analysis model (TAM) [17] and Temporal-LDA (TM-LDA) [38] model the temporal aspect of topics in social media streams like Twitter. MedLDA [43] uses a maximum margin classifier jointly modeled with topics to build predictive models for categorical and continuous variables. Vaca et al. [37] used a collective matrix factorization method to track emerging, fading and evolving topics in news streams. Shahaf et al. [32] developed a scheme to connect related news articles to enable better understanding of news stories.

There have also been works that have explicitly focused on prediction of real-world events by mining different kinds of corpora. Radinsky and Horvitz [29] proposed a framework to predict future real-world events from News and Web data. They designed automated abstraction techniques that are able to generalize from specific entities to broader classes of observations and events. The work by Rudin et al. [30] involves predicting the next event in sequentially organized data such as a customer's online shopping cart or the winners in each round of a sports tournament, using association rule mining and Bayesian analysis. Amodeo et al. [1] proposed a hybrid model consisting of time-series analysis, to predict future events using the New York Times corpus. More recently, there has been work which has focused on relationships and dependencies between structured data streams and real-world events. FBLG [5] focused on discovering temporal dependency from time series data and applied it to a Twitter dataset mentioning the Haiti earthquake. Similar work by Luo et al. [21] showed correlations between real-world events and time-series data for incident diagnosis in online services. In most of these works, events and/or topics have just been used as a tool for knowledge acquisition or information extraction, whereas our goal is to use the extracted events to predict fluctuations in socio-economic indicators. To the best of our knowledge, there has been no previous attempt to combine such events or topics from unstructured text streams with structured data (such as a time-series) to characterize and forecast fluctuations in socio-economic indices.

However, there does exist work in predicting specific variables from news data, such as stock price. Hagenau et al. [11] proposed a new scheme to include context from financial news and market feedback to better predict stock prices. Ming et al. [24] used daily news articles from the WSJ corpus and sparse matrix factorization techniques to predict stock price movement while Gidofalvi [10] used financial news to predict volatility of stock prices based on a naive Bayes classifier. Other works which focus on predicting stock prices include [3][8][41]. Si et al. [33] propose a technique to leverage topic based sentiments from Twitter to help predict the stock market. Similar works have been proposed for political indicators [9]. Our work differs from the aforementioned in that we propose a framework, where we specifically connect real-world events extracted from text data (news articles) to predict external variables or indicators, driven by the assumption that there exists a dependency between these variables and real-world events. In a sense, our framework is similar to Google Correlate [25], in terms of general applicability, however we focus on identifying real-world events in news sources whereas Google Correlate focuses on web search logs to determine queries that are correlated with real-world phenomenon. In addition, we propose a novel generative model to identify events that drive fluctuations in socio-economic indicators, whereas Google Correlate uses standard correlation metrics.

3. PROBLEM DEFINITION

We consider the following setup. Consider a corpus $\mathcal{D}$ of news articles indexed by time t, so that $\mathcal{D}_t$ is the collection of news articles published at time t. The granularity of the time index t depends on the particular application - it can be a day, week, month etc. and our formulation is agnostic to the actual granularity. 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. We discuss how to identify these events given a corpus of news articles in section 4. In addition, suppose that there exists a function $\phi_t : \mathcal{D}_t \to [0,1]^K$ that maps a collection of news articles published at certain time t, to a vector $\phi_t(\mathcal{D}_t) = (\phi_{t1}, \phi_{t2}, \ldots, \phi_{tK})$ that specifies the "intensity" of each of the K events at time instant t. In other words, larger the value of ϕ_{tk}, more is the proportion of event $k \in [K] := \{1, 2, \ldots, K\}$ in corpus $\mathcal{D}_t$. Note that $\phi_t(\mathcal{D}_t)$ will be a *sparse* vector with most of the entries being zero, corresponding to the fact that only a few events will be mentioned in the corpus $\mathcal{D}_t$. We call $\phi_t(\mathcal{D}_t)$ the *event vector* at time t and simply refer to it as ϕ_t in the remainder of the paper, with the collection $\mathcal{D}_t$ being defined implicitly by the time t. For the purposes of defining the problem, we suppose that the mapping ϕ_t is already provided and discuss how to construct such a mapping from unstructured news text in section 5.

Our objective is to build a model that can predict or forecast socio-economic indicators based on real-world event occurrences. Let $\boldsymbol{y}$ denote a time series of some socio-economic indicator of interest such that y_t represents its value at a certain time t. We consider two different prediction tasks:

- To predict $\boldsymbol{y}$ we can use standard time-series models such as ARIMA [4], that try to predict future values based on previously observed values. In our context, in addition to previously observed values of $\boldsymbol{y}$ we also take into account the real-world events extracted (the mapping ϕ_t), to predict future values of $\boldsymbol{y}$. Our proposed model has additional parameters ω_t^k denoting the weight of each event $k \in [K]$ at time t as well as a time lag δ. The new forecasting equation is given by:

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

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

  where e_t represents the moving average component of the ARIMA model and ϵ_t represents the error.

- We can also predict "spikes" in $\boldsymbol{y}$ where a spike is defined as a sudden change in the value of y_t from its previous value y_{t-1}. In this case, suppose there is training set of m examples: $\mathcal{T} = \{(\phi_{t_i}, s_{t_i}) \mid 1 \leq i \leq m\}$ where each example corresponds to a pair of the event vector (ϕ_{t_i}) and whether or not a spike ($s_{t_i} \in \{0, 1\}$) was observed at some time instant t_i in the past. We learn a standard SVM-based [6] binary classifier to predict spikes in the socio-economic indicator $\boldsymbol{y}$.

The above formulation can in principle use any function ϕ that maps a collection of news articles into a representation of real-world events. Indeed, our evaluation section looks at different ways to construct the mapping ϕ and compare the predictive performance of each of these methods. However, one major contribution of our work is the construction of a specific mapping ϕ – the *event class* model, which we describe in detail in the next section.

4. EVENT CLASS MODEL

For any generic document, ranging from academic articles to blog posts, topics can be a very important and useful feature for understanding its content. However, for specialized documents like news articles whose purpose is to report stories and incidents, we claim that *events*, which are aimed at capturing the main "actions" or "incidents", are more informative than *topics*, which essentially capture the main themes in the document. We envision topics as being part of the event *description* but there are additional aspects of events that need to be captured separately.

We now define the major components of our proposed event model. As mentioned in section 1, an event is an amalgamation of many components – entities, topics and metadata like location and time. 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. This is in contrast to topic models that consider all the words mentioned in a document. These collection of words/phrases are termed as *triggers* and we define an *event class* as a collection of related or similar triggers. More concretely, we have the following:

Definition 1. ***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.).*

Definition 2. ***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. We assume that any news article reports *one* central or main event which is drawn from a finite set of event classes. This event class is identified through the triggers present in the article. Again note that this differs from a standard topic model like LDA which assumes each document is a distribution over different topics. 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 [27]. 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. This specific event also involves location(s), topics (e.g. car accident or an air crash) and other properties, such as, actors, objects etc. A particular event (drawn from an event class) with specific values of actors, objects, location is manifested in one or more news articles. However, the main essence of this event class is carried by the *event triggers*, which is the primary action that describes the event class. 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.

The description of an event can be further enriched by identifying other information that frequently accompanies this event class. For example, consider the title from a New York Times article after the bombing at the Boston marathon – "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. This example can be generalized to assume that every central event covered in a news article is typically associated with other events, which are also mentioned in the article. Let us call these additional events as *subsidiary events*, defined as,

Definition 3. ***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.*

Now, going back to our original assumption that any news article is reporting *one* central event and if we use a collection of news articles to build an event model, an *event* can be defined as having one central event (i.e. triggers drawn from an event class) and mixture of closely associated events or subsidiary events (also, different triggers for each subsidiary event drawn from separate event classes). When a news article is published, we claim that it is an instance of an event, generated from the event model. With this, we define an event as:

Definition 4. ***Events*** *are a combination of one central theme and a mixture of subsidiary themes, where these themes are represented as event triggers drawn from separate event classes.*

Table 1: Model variables and parameters

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

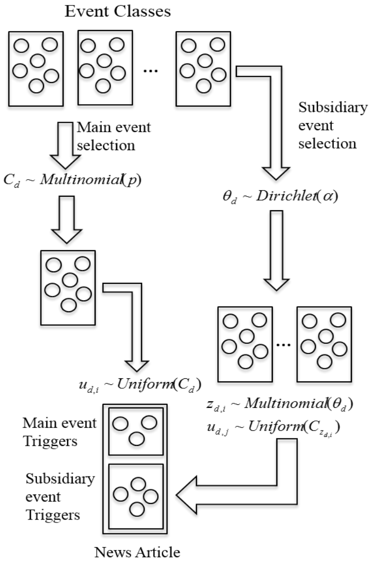
Event Classes
...
Main event selection
Subsidiary event selection
$C_d \sim Multinomial(p)$
$\theta_d \sim Dirichlet(\alpha)$
...
$u_{d,i} \sim Uniform(C_d)$
$z_{d,i} \sim Multinomial(\theta_d)$
$u_{d,j} \sim Uniform(C_{z_{d,i}})$
Main event Triggers
Subsidiary event Triggers
News Article

Figure 1: Event class model and generation of news articles. Event classes (boxes) are represented as collection of event triggers (small circles)

Finally, a news article is generated as an observable instance of an event sampled from this event model, along with additional information (e.g. location, timestamp, entities), which completes the description of a particular event instance.

4.1 Generative Model of News Articles

In spite of the dissimilarities, our generative model is motivated by LDA. Consider a corpus of news articles $\mathcal{D}$ and let the total number of articles be $D := |\mathcal{D}|$. Suppose the set of all event triggers extracted from the news corpus be given by $U = \{u_1, u_2, \ldots, u_M\}$ and let $C = \{C_1, C_2, ..., C_K\}$ denote the event classes, where each $C_i \subset U$ is a collection of event triggers and $C_i \cap C_j = \emptyset \ \ \forall \ i \neq j$. As stated previously, our assumption is that every news article describes only *one* main event and consequently we can associate each news article with a single generative event class. Suppose $C_d \in C$ denotes the (latent) event class of the main event of any news article $d \in \mathcal{D}$. Note that for ease of notation, we represent the event class of article d as a set instead of an index in $[1 \ldots K]$. Next, suppose that each $d \in \mathcal{D}$ consist of a disjoint union of two sets of words: $U_d \cup W_d$ where, each $u_{di} \in U_d$ represents an event trigger present in d. The remaining non-trigger words are represented by the set W_d. The set of trigger words is further divided into two types: $U_d = U_d^e \cup U_d^s$ where U_d^e represents the trigger words occurring in the title and the lead paragraph of article d (viz. the event class triggers) and U_d^s is the set of subsidiary event triggers occurring in the body of article d. The words in W_d represent other features of the news article such as topics, entities, locations, people etc. They can be considered as a description of the event, like *where* it happened, *when* it happened, *people*, *organizations* involved etc. but we assume that they are independent of the events mentioned in the news article [19]. Given the above, the generative process of the articles given a set of event classes is as follows (also shown in Figure 1).

1. For each article $d \in \mathcal{D}$, its event class C_d is sampled from a categorical distribution with parameter $\boldsymbol{p}$:

   $$C_d \sim \text{Categorical}(\boldsymbol{p})$$

   where $\boldsymbol{p} = (p_1, p_2, \ldots, p_K)$ denotes the prior probability distribution of the event classes in the corpus $\mathcal{D}$.

2. For each event class trigger $u_{di} \in U_d^e$ in the article d, sample an event trigger uniformly at random from the triggers belonging to sampled event class:

   $$u_{di} \sim \text{Uniform}(C_d)$$

   where $\text{Uniform}(C_d)$ is the uniform distribution over the event triggers in the set $C_d \in C$.

3. After the main event class along with the main event triggers are generated, the subsidiary events are sampled. The subsidiary events generation is similar to topic generation in LDA [2]. First sample a distribution of subsidiary events according to a Dirichlet distribution using an assumed Dirichlet prior $\boldsymbol{\alpha}$:

   $$\boldsymbol{\theta}_d \sim \text{Dirichlet}(\boldsymbol{\alpha})$$

   Then for each subsidiary event trigger $u_{dj} \in U_d^s$, first sample its event class:

   $$z_{dj} \sim \text{Multinomial}(\boldsymbol{\theta}_d)$$

   and finally sample an event trigger uniformly at random from the sampled event class:

   $$u_{dj} \sim \text{Uniform}(C_{z_{dj}})$$

   Note that the above process differs from LDA in that the distribution for u_{dj} is assumed to be uniform instead of a multinomial distribution in the case of LDA.

The posterior is given by $\Pr[C_{1:D}, \boldsymbol{\theta}_{1:D}, \boldsymbol{z}_{1:D} \mid \boldsymbol{p}, \boldsymbol{\alpha}, \boldsymbol{U}_{1:D}]$ where $\boldsymbol{z}_d, \boldsymbol{U}_d$ are vectors. In this model, we can observe the variables $\boldsymbol{U}_d$ corresponding to (all) the event triggers in article d and we assume that $\boldsymbol{p}$ and $\boldsymbol{\alpha}$ parameters are also known. The rest of the variables are latent. A summary of the parameters and variables of this model is presented in Table 1. The posterior distribution is used to infer these variables given the news corpus. We used Markov Chain Monte Carlo (MCMC) approximation method to compute the posterior. Our goal is to identify which event class (C_d) the article d belongs to, by observing the words belonging to the set U_d and their positions. In the remainder of this section, we discuss how the set of event triggers U is extracted from a corpus of news articles, followed by how the event classes are constructed from the set of extracted event triggers.

4.2 Event Trigger Extraction

The objective of Automatic Content Extraction (ACE) [7] is to develop automated content extraction techniques to support processing of natural language text from a variety of sources such as newswire, broadcast conversation, and weblogs. ACE has a specific task for event extraction from news sources – which defines *event triggers* as the words (or phrases) in a sentence that specify the occurrence as well as the type of the event. For example, in the news article headline – *FIFA Officials Arrested in Corruption Case* – the word ***arrested*** is the trigger word for the event mentioned in the article. Usually event triggers are verbs or nouns present in the sentence that describe some notion of "action" or "incident". Event triggers can have different forms – verbs (*Traders* ***protest*** *over FDI in retail*), nouns (***Burglary*** *in police station leaves cops red-faced*) and sometimes the verb or noun themselves cannot form the trigger but a combined phrase is the trigger (***Number of AIDS patients go up*** *in MP*).

In a standard event extraction task, triggers are extracted at a sentence level to understand the type of event mentioned in the sentence. Our goal is to understand the "best" event type that describes an entire news article. Therefore, we identify which triggers in a news article collectively describe the central event of the article. Typically, a news article is organized as follows – a title or headline which contains a one line overview of the main event; followed by the lead (or first) paragraph of the article which contains a brief description of the main event; then the rest of the article presents details of the central event along with follow-up actions of the main event. Based on this standard flow of a news article, we assume that the triggers appearing in the title and lead paragraph of the article are representative of the main event and consequently, form part of the set of all event class triggers.

Every word in the title and the first paragraph can be classified either as a trigger word or not. As discussed above, a trigger can have any part-of-speech and can consist of single or multiple words. Also, in our setting, a trigger may not be restricted to a 8-class event type/33-class event sub-type specification as proposed in ACE. Thus, existing methods [20][42] are not suitable for our setting.

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 and have been used in a variety of tasks such as POS tagging, speech detection etc. Lafferty et al. [18] define the the probability of a particular (output) label sequence $\boldsymbol{o} = (o_1, o_2, \ldots)$ given an (input) observation sequence $\boldsymbol{x} = (x_1, x_2, \ldots)$ to be a normalized product of real-valued *potential functions*, each of the form

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

where $t_j(o_{i-1}, o_i, \boldsymbol{x}, i)$ is a *transition* feature function of the entire observation sequence $\boldsymbol{x}$ and the labels at positions i and $i+1$ in the output sequence; $s_k(o_i, \boldsymbol{x}, i)$ is a *state* feature function of the label at position i and the observation sequence $\boldsymbol{x}$; and λ_j and μ_k are parameters to be estimated from training data. The probability of the label sequence $\boldsymbol{o}$ is modeled as

$$\Pr(\boldsymbol{o} \mid \boldsymbol{x}; \boldsymbol{\lambda}, \boldsymbol{\mu}) = \frac{1}{Z(x)} \exp(\sum_i \log F_i(\boldsymbol{o}, \boldsymbol{x})) \tag{3}$$

where $Z(x)$ is the normalizing constant and $F_i(\boldsymbol{o}, \boldsymbol{x})$ is defined above. The parameters $\boldsymbol{\lambda}, \boldsymbol{\mu}$ can be estimated given a training dataset of input and output sequence pairs using maximum likelihood estimation technique. 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. For example, the article title – "Blasts at Boston Marathon Kill 3 and Injure 100" would be labeled as,

> (Blasts-T) (at-NT) (Boston-NT) (Marathon-NT) (Kill-T) (3-NT) (and-NT) (Injure-T) (100-NT)

where, each (word-label) pair represents (x_i, o_i) where the input sentence $\boldsymbol{x} = (x_1, x_2, \ldots, x_N)$ and the trigger label sequence $\boldsymbol{o} = (o_1, o_2, \ldots, o_N)$ and N is the number of words in the input sentence. The triggers were manually labeled by an annotator to obtain the training, testing and validation datasets. We use a variety of features – both textual and non-textual – to build the CRF model for detecting triggers. Some examples include – POS tags, Named-entity tags, position of the word, preceding word, whether present in the title etc. Experiments were conducted using different feature sets and the best combination was chosen based on the performance on the validation dataset.

4.3 Constructing the Event Class

Now that we have identified the event triggers present in any news article, the next step is to determine the event classes using the extracted event triggers. The idea is to cluster "similar" news articles (describing similar events) and obtain the event triggers that describe events belonging to the same event class. For example, articles talking about a bomb explosion are likely to have triggers like, *explosion, blast, bombing* etc. So in our model, an event class is represented by event triggers that are similar to each other. To cluster the event triggers, we need to define a suitable notion of similarity between any two triggers. We use a neural network based language model [23] 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. The learned representations have been shown to have a number of interesting linguistic properties and can be used to cluster words having similar meanings or those used in very similar contexts. In our context, the event triggers extracted from the news corpus (see section 4.2) were embedded in a vector space of dimension 100 and clustered using K-means with the cosine similarity metric [22] used for computing distance between any two trigger words. To get the optimum number of event classes, we varied the number of clusters K and used the "elbow" method [34] to determine a suitable value of K. We ended up with $K = 250$ event classes for our corpus consisting of articles in a 7-year duration (see section 6 for more details of the corpus).

5. EVENT-DRIVEN PREDICTION

Our goal is to forecast various socio-economic indices using our proposed event model. In this section, we describe how we construct the mapping ϕ_t (see section 3) from the event class model. After running the MCMC inference, 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

$k \in [K]$ at time t as:

$$\phi_{tk} = \frac{\sum_{d \in \mathcal{D}_t} \mathbb{1}[C_d = k]}{|\mathcal{D}_t|} \quad (4)$$

where $\mathcal{D}_t$ is the collection of articles published at time t. Once we have the event vector ϕ_t, the general formulation of the predictive model is:

$$y_t = \omega_t^0 + \sum_{k=1}^{K} \omega_t^k \phi_{tk} + \epsilon_t \quad (5)$$

where y_t represents the socio-economic indicator whose value is being predicted using the extracted event classes. While the above formulation fully captures the relationship between y_t and the co-occurring events appearing in news articles, it is somewhat unwieldy given that K can be very large. However, we make the observation that in practice, only a subset of the K events actually influence any socio-economic indicator. We extract the relevant events based on co-occurrence patterns according to the method described next.

5.1 Identifying Relevant Events

Our goal is to find a subset of $M_y \leq K$ events, that have a strong "association" with the socio-economic indicator $\boldsymbol{y}$, in other words, they have discriminative power in predicting fluctuations in $\boldsymbol{y}$, such as a sudden rise in the price of a particular crop. More formally, suppose that at any time t, if y_t is at least 10% higher than y_{t-1}, then we define a spike $s_t = 1$, otherwise $s_t = 0$. This gives us a spike signal $\boldsymbol{s}(\boldsymbol{y})$ derived from the underlying signal $\boldsymbol{y}$. We use the likelihood ratio test [12] to identify the events (event triggers) that co-occur with spikes in the signal $\boldsymbol{s}(\boldsymbol{y})$. Specifically, we use the top 5% event triggers according to the likelihood ratio to construct the set $\phi(M_y)$ of events that have strong association in predicting fluctuations in $\boldsymbol{y}$. Given an event vector ϕ_t at time t, we define the mapping $\phi_t^{\boldsymbol{y}} : \mathcal{D}_t \rightarrow [0,1]^{M_y}$ as $\phi_t^{\boldsymbol{y}} = (\phi_t)_{\{k \in \phi(M_y)\}}$ and employ it in the event-driven predictive model above.

There can be many variations of the general event-driven predictive model introduced above. In the rest of this section, we present two specific models.

5.2 Historical Event Model

In the formulation so far, the value of y_t was assumed to be influenced by only the events happening at time t. In reality, an event might have a delayed effect on $\boldsymbol{y}$ and only considering events at time t might result in loss of useful information for the purposes of prediction. To capture this dependency, we introduce an additional parameter δ into the model that measures the window of influence of historical events. In particular, we consider events reported in articles published at times $t - \delta, \ldots, t$. The linear model that best approximates y_t under this formulation is of the form:

$$y_t = \omega_t^0 + \sum_{j=0}^{\delta} \sum_{k=1}^{K} \omega_{t-j}^k \phi_{(t-j)k} + \epsilon_t \quad (6)$$

where ω_{t-j}^k denotes the weight (or impact) of event k at time $t - j$ on the current value of the indicator y_t.

5.3 Topic Based Prediction

We discuss here how topics can be used to predict socio-economic indices. A set of topics learned from a large corpus of documents represents the main *themes* contained in the corpus. For news articles – which can be assumed as a source of events – topics can be thought of as the main themes of events covered in the news corpus. Therefore, suppose we learn an LDA topic model [2] over the entire news corpus $\mathcal{D}$, with number of topics as K'. The LDA topic model supposes that each article is a mixture of topics, each appearing with different proportions. We construct the mapping ϕ_t (refer to section 3) as follows: given a collection of articles $\mathcal{D}_t$ at time t, let θ_{di} represent the proportion of topic $i \in [1, \ldots, K']$ in article $d \in \mathcal{D}_t$. Then we define the topic-based mapping as

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

where

$$\theta_t^i := \max_{d \in \mathcal{D}_t} \theta_{di} \quad (7)$$

where, $\theta_{d,i}$ represents the proportion of topic i in article d appearing on day t. This is done to take the maximum proportion of a topic in a day's articles to be representative of the day's topic proportion. Alternatively, the average proportion $\sum_d \theta_{di} / |\mathcal{D}_t|$ from all the articles on day t could have been used to represent θ_t^i. Our design choice is based on the intuition that large presence of a topic in an article has more value than appearing in small proportions in many articles.

6. QUALITATIVE ANALYSIS

The design and the goals of the event model presented in this paper are different from existing event extraction tasks. Thus, it is difficult to evaluate our model using existing gold standard data which are not suitable for our purposes. So, we evaluated the model qualitatively on a corpus of news articles. Our data consists of the archive of the `Times of India`, the most circulated English daily published in India, between the years 2006 and 2012. The corpus had around 700,000 articles with a vocabulary of around 650,000 words. As we mentioned earlier, there were **250** distinct events classes extracted from the corpus. Table 2 presents 9 randomly picked event classes that were extracted from the corpus along with a subset of the subsidiary events associated with each event class. Here, we present 6 randomly chosen trigger words associated with the event class. The corpus contained articles mostly from India, and hence, some words extracted are specific to India. For example, *aila* was a cyclone that hit the eastern coast of India in 2009, *dharna* is a Hindi word for protest or agitation. From the table, we can observe that for every event class, the subsidiary events are in fact talking about related or associated events. For instance, in the event class corresponding to disease outbreak, most of the subsidiary events describe what follows an outbreak – *treatment*, *admitting* patients to hospitals and plans for *prevention*. However, there are also some subsidiary events that are precursors to an event – *announce*, *nominate* happen before an election. The present version of our model is incapable of differentiating between the subsidiary events that happen *before* or *after* an event. A possible future direction can be to classify the subsidiary events among these 2 types, which will provide better insights about each event class.

7. FOOD PRICE PREDICTION

In this section, we evaluate our event model by demonstrating its ability in predicting the value of socio-economic indicators. We have built event-driven predictive models to forecast the prices of 12 popular crops in India. Events extracted from the news corpus based on our event model (section 4) are used as features in these models. In this paper, we specifically choose food price fluctuations as an example scenario to demonstrate how event-based predictive models can be built. However, our approach can easily be

Table 2: Examples of event classes extracted from the news corpus. For each event class, the event class triggers are shown in the left column. Event class triggers are assumed to be equally likely so there is no order amongst the different triggers. In the right column, the subsidiary events associated with each event class are shown.

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,

generalized to similar socio-economic indicators whose volatility are potentially influenced by events appearing in the news media.

As discussed in section 3, we consider two prediction scenarios – predicting the actual (real-valued) price of the crops and predicting spikes in prices. The second predictive model is a binary classifier, predicting the price on a particular day as a spike (1) or non-spike (0). We compared our approach with standard baseline systems. For the real-valued prediction, we compared the performance with a standard time series projection model – ARIMA [4]. We show that our event-driven predictive model built on top of the ARIMA model reduces the forecast error compared to the pure ARIMA model. In addition, our spike predictor demonstrates improvement over different baseline systems. We begin with a brief description of the data and then discuss the experiments performed for the evaluation.

7.1 Data

The food price data for the experiments was collected from the website of the Ministry of Agriculture of Government of India [2]. The Directorate of Marketing and Inspection under the ministry publishes daily prices of different crops which include the minimum, maximum and modal (the rate at which maximum sale was done) prices of these crops across many different markets in the country. In this work, we focused on 12 different crops. We considered the following crops – onion, potato, rice, wheat, maize/corn – which are among the most consumed agricultural products in India. Other crops include agricultural produce across different categories, like fruits (apple, mango etc.), vegetables (eggplant, cauliflower, cabbage etc.), cash crops (cotton, jute, sugar etc.) We used the daily modal prices of these crops and computed the average across different markets in the country to represent the daily price of the crop in India.

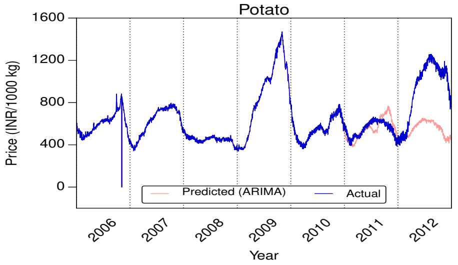

Figure 2: Price time-series of potato and projection using SARIMA. The model was trained till 2010 and tested for 2011 and 2012

7.2 Price Projection

Any time series data is usually made up of different components – trend, seasonality, cyclic and error. ARIMA (Auto-Regressive Integrated Moving Average) [4] models are the most general class of models for forecasting a time series. The ARIMA forecasting equation for a stationary time series is a linear (i.e., regression-type) equation in which the predictors consist of lags of the dependent variable and/or lags of the forecast errors. For the non-seasonal $\mathrm{ARIMA}(p, d, q)$ model, the projected value is computed by:

$$y'_t = \epsilon + \alpha_1 y'_{t-1} + ... + \alpha_p y'_{t-p} - \beta_1 e_{t-1} ... - \beta_q e_{t-q} \quad (8)$$

where y'_t is the differenced time series with d being the degree of first differencing, p represents the number of AR (Auto-Regressive) terms ($\boldsymbol{\alpha}$ are the parameters) and q represents the number of MA (Moving Average) terms ($\boldsymbol{\beta}$ are the parameters). However, due to inherent seasonal variation in food price data, the SARIMA[3](seasonal ARIMA) is a more suitable model to forecast food prices. We used the SARIMA (1,0,1)x(0,1,1) model, as it showed the best performance for our data. The seasonal difference is added by introducing a new variable y_{t-s} where $s = 365$ and represents the average daily price for one month period, for the month 365 days prior to t, to understand the previous year's trend. Figures 2and 3 show the forecasted prices for potato and wheat respectively by this model. The time series in the figures present the average daily modal value from different markets in India between January 1, 2006 and December 31, 2012. The model was trained with price values till December 31, 2010 and was tested on the predicted values between January 2011 and December 2012. The figures show that this model can detect the cyclic patterns of food prices across the years and the general trend but is unable to predict the local variations in the time series. We show next that incorporating event information can help in improving the accuracy of prediction.

7.3 Event-driven Prediction of Food Price

Volatility of food price can be attributed to many factors, including seasonal factors, inflation etc. However, there are certain incidents and events that can have a sudden impact on food

[2] http://agmarknet.nic.in/

[3] A seasonal ARIMA (SARIMA) model is classified as an ARIMA(p,d,q)x(P,D,Q) model, where P=number of seasonal autoregressive (SAR) terms, D=number of seasonal differences, Q=number of seasonal moving average (SMA) terms and p,d,q are the standard (non-seasonal) ARIMA terms

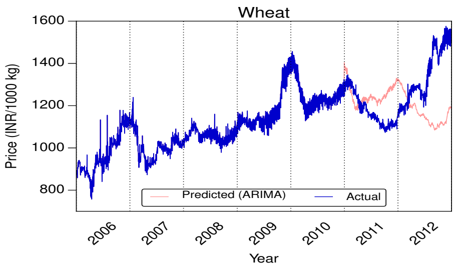

Figure 3: Price time-series of wheat and projection using SARIMA. The model was trained till 2010 and tested for 2011 and 2012

prices. Unfortunately, the existing models for time series projection do not account for such factors. Here, we present a modified ARIMA model where information about current events are incorporated within the model. Since not all events are equally relevant for a socio-economic indicator, we first identify the events that influence fluctuations in a specific indicator of interest based on the technique described in section 5.1. Table 3 presents the top 5% (according to the likelihood ratio) event class triggers for some of the crops. We see some commonalities across crops, like disease outbreaks, violence and movements, natural calamities, festivals etc. However, there are also some unique events for each crop. For example, transportation strike or any incident affecting the transportation system has a more profound effect on wheat and potato (example events being attack on railways, railway strike etc.). This is probably due to the fact that wheat and potato are not cultivated uniformly across the country and heavily rely on transportation for proper distribution. Any event that affects transportation of these crops has an effect on the supply, leading to a rise in price. On the other hand, festivals influence only onion and wheat prices. These two crops have a higher consumption during festivals leading to an increased demand and thereby increasing the price.

Table 3: The top 5% events (triggers) associated with each crop is shown in the order of their likelihood ratio value, for the four most popular crops (based on the data collected between 2006 and 2012)

Crop	Associated Triggers
Onion	Strikes/agitation/disturbances, increase , arise/occur exported, scam, celebrated flooding, importing, outburst, electing, paralysed, polluted
Potato	Strikes/agitation/disturbances, exported, scam, prevented scam, blockage, probing, hike, storing, raining
Wheat	strikes, importing, rise/hike, stabilize, dominated procure, rafting, sacked, save, disruption, gain
Rice	exported, scam, strikes, celebrated, riding, visited, hail, damaged, rise, acquire, transfer, join

7.3.1 Event Incorporated ARIMA

We use the model introduced in section 3 (Eq 1) with a slight modification: instead of using all the K events, we only use the subset $\phi(M_y)$ of events that are identified as relevant (see above) for the socio-economic indicator $\boldsymbol{y}$. This modified model adds event-based features on top of the standard ARIMA model to compute the value of socio-economic indicator at time t. The comparison of this model with the standard ARIMA model is shown in Table 4 where we report the RMSE of prediction. We observe that the residuals for the event-based ARIMA model is *lower* in most cases (apart from tomato and cabbage). In particular, the mean RMSE for the event-incorporated ARIMA is lower which supports our hypothesis that certain events can have impact on food prices. The real-world events are determined by external forces that need not have any specific (or repeating) patterns and hence, standard time series analysis methods cannot capture fluctuations caused by such factors.

Table 4: Performance comparison between ARIMA and event-incorporated ARIMA in predicting food prices. The table shows Root mean square errors (RMSE) for both models across all the crops

Crop	ARIMA	ARIMA + Events
Tomato	546.05	603.63
Wheat	195.70	139.21
Onion	710.24	331.52
Potato	322.09	199.93
Rice	147.44	121.64
Cabbage	288.76	403.87
Sugar	369.65	350.97
Apple	1511.94	903.71
Carrot	556.28	450.28
Maize	138.61	133.96
Brinjal	241.15	202.61
Jute	637.28	555.76
Mean	472.10	365.75

7.4 Prediction of Spikes in Food Prices

In this section, we discuss the performance in predicting food price spikes. We say that a spike in price is *true* if the value at time t is greater than the value at time $t-1$ by more than 10%. Our event model is built using a hierarchical process, involving extraction of event triggers as the first step and adding subsidiary events at a later stage. So, the first experiment we conducted was to gauge the advantage of adding subsidiary events to the event model. We consider a naive event-trigger based model (called `TRIGG`) that only utilizes the event class triggers occurring in any news article. To further understand the benefits of adding subsidiary events in our model against adding traditional topics, we implemented another model that combines `TRIGG` (event triggers) and LDA topics which we refer to as `TRIGG+LDA`. Finally, we consider a model that incorporates past events *and* current events for the purposes of prediction. We call this model `EVENT_HIST` (see section 5.2). As for the case of predicting the actual values in section 7.3.1, here also we only consider the subset of relevant events identified in section 5.1 for the purposes of predicting food price spikes.

Table 5: Accuracy in spike prediction for food prices

	Potato	Wheat	Rice	Onion
`TRIGG`	0.509	0.512	0.514	0.520
`LDA`	0.431	0.432	0.471	0.488
`TRIGG+LDA`	0.539	0.544	0.548	0.553
`EVENT_HIST`	**0.633**	**0.621**	**0.613**	**0.652**
`EVENT`	0.567	0.586	0.608	0.612

Table 5 summarizes the performance of the various event-driven predictive models defined above against other baselines and popular text mining techniques, where we report the accuracy of each method. Among individual crops, onion price prediction showed the best results, followed by rice, wheat and potato. A possible explanation of this observation is that the price volatility of onion

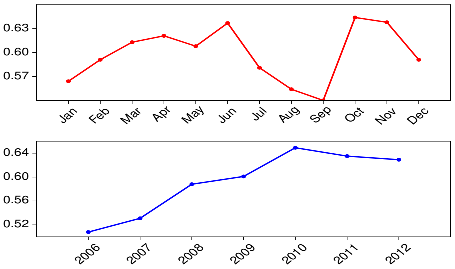

Figure 4: Average accuracy of EVENT model for all crops for each month across 7 years (upper) and average accuracy across 7 years [2006-2012] (lower)

is widely covered in the news articles compared to other crops, as onion price fluctuations have more adverse effects compared to the other crops [36]. Observe that *both* the EVENT and EVENT_HIST models have better performance compared to the baseline models. Further, the improved performance of TRIGG+LDA highlights the benefits of adding subsidiary events in our probabilistic event model. The poor performance of topics (LDA) is due to the ubiquitous nature of topics in news streams – almost all the topics appeared on each day and the per-day topic distribution was approximately uniform. On the other hand, events are much more sparse and display a non-uniform distribution over time. As a result, events displayed more discriminating power than topics in predicting the socio-economic indicators. To summarize, the results for EVENT and EVENT_HIST demonstrate better predictive power compared to standard methods, such as LDA, showing the benefits of events as features over topics. Moreover, the results clearly demonstrates the benefits of adding subsidiary events to the model and adding historical event information as additional features. In absolute terms, the accuracies are not high but are significantly better than the basic baseline of a random guess.

A finer analysis of the EVENT model showed some interesting characteristics of the prediction performance. Figure 4 (upper) shows the average accuracy of the model in predicting food price spikes for each month across the 7 years. We see that the performance goes down in the middle of the year (Jul-Sep) and again in December and January. In these months, the seasonal effect influences the food prices and the real-world events are not able to capture this. In particular, July and August are peak monsoon months and September often suffers from a lasting effect of the monsoon, including floods, disease outbreaks etc. Similarly, during winter (Dec-Jan) prices tend to go up. Tuning our model to remove such seasonal effects can increase the accuracy for these months. Figure 4 (lower) presents the variation in prediction accuracy across the different years. This plot shows an upward trend in the accuracy. The rise in prediction accuracy towards the end can be attributed to the fact that Times of India gradually increased the count of online articles. In 2006, the average number of daily articles were around 200, which increased to 600 in 2011-12. The increase in the number of articles meant more coverage of real-world events, which lead to better event-driven prediction.

8. DISCUSSION

The event model presented in this paper is based on the assumption that every news article has a main event, which is usually mentioned in the title and/or lead paragraph of the article. In the scenario where two (or more) types of events are discussed in the lead paragraph of the same article, we still preserve the assumption that there is only one main event in the article by choosing the event (i.e. the corresponding triggers) appearing *first* among the two (or more) events in the lead paragraph, as the main event of the article. This is based on the intuition that the preceding event has more importance and needs to be mentioned before introducing or explaining the other events in the lead paragraph.

In the present implementation, the subsidiary events in an article are assumed to be independent of the main event. This helps to keep our event model simple, though in practice there might be some inherent dependencies between the main and subsidiary events mentioned in news articles. For example, if the main event in the article is about a natural calamity, the subsidiary events typically will be related to its impact (e.g. deaths, injuries, rescue etc.). Hence, one of the goals in future work would be to redesign the model assuming a dependency between the main event and the subsidiary events in the articles.

For food price prediction, we implemented our event model on top of the seasonal ARIMA model. However, for the case of spike prediction, we did not consider the seasonal effects in the price spikes. Accounting for seasonality is a possible opportunity for future work. Moreover, in building the predictive models, we determined the events more likely to co-occur with the observed phenomenon (food price rise in this case) beforehand. Alternatively, all event types could have been used for this predictive analysis. This approach will involve many more parameters and will be slower to train. Also, preliminary experiments showed that the weights for most of the events were close to zero, making the adapted approach with fewer but more related events in the predictive model a better choice.

In future, we would like to augment our predictive models with more features apart from the event classes. There might be certain entities, such as people, location, organizations or topics that are also related to the observed phenomenon (food price rise or others) and incorporating them can lead to better prediction performance.

9. CONCLUSIONS

This paper is based on the premise that many socio-economic indicators are sensitive to real world events and this can be used to forecast fluctuations in their values. We presented a novel way of defining and extracting events from a large news corpus. We use these events to design and implement models to predict fluctuations in food prices. Our event-driven predictive model showed a 22% reduction in the RMSE when a standard ARIMA model is incorporated with event information. A separate model was built to predict spikes in the prices, where we observed that the event-based model outperformed the other models, including an LDA based predictive model. This paper has solely focused on predictive food prices using news events. However, our approach can easily be extended to predict other socio- or macro-economic indicators or phenomena, which are sensitive to events. One future direction is extending the evaluation to build predictive models for other scenarios such as stock prices, disease outbreaks, or forecasting effects of natural calamities like drought, flood, cyclone etc. The present work has only focused on finding association between events and an observable phenomenon. Another extension of this idea is to identify *causing* and *resulting* events of any phenomenon, which can be used for finer analysis and greater clarity in understanding the observed phenomenon.

10. ACKNOWLEDGEMENTS

We thank the NYU Abu Dhabi Research Institute and the Center for Technology and Economic Development (CTED) in NYU

Abu Dhabi for supporting Sunandan Chakraborty and Lakshminarayanan Subramanian on this project. We thank the NSF for supporting Ashwin Venkataraman in part on this project under the NSF CAREER grant CMMI 1454310. The authors would also like to thank Shankar Kalyanaraman for his valuable input on this paper, particularly for helping with the initial data collection.

References

[1] G. Amodeo, R. Blanco, and U. Brefeld. Hybrid models for future event prediction. CIKM '11, pages 1981–1984, 2011.

[2] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, Mar. 2003.

[3] J. Bollen, H. Mao, and X. Zeng. Twitter mood predicts the stock market. *J. of Computational Science*, 2(1):1–8, 2011.

[4] P. J. Brockwell and R. A. Davis. *Introduction to Time Series and Forecasting*. Springer, 2nd edition, Mar. 2002.

[5] D. Cheng, M. T. Bahadori, and Y. Liu. Fblg: A simple and effective approach for temporal dependence discovery from time series data. KDD '14, pages 382–391, 2014.

[6] C. Cortes and V. Vapnik. Support-vector networks. *Mach. Learn.*, 20(3):273–297, Sept. 1995.

[7] G. R. Doddington, A. Mitchell, M. A. Przybocki, L. A. Ramshaw, S. Strassel, and R. M. Weischedel. The automatic content extraction (ace) program-tasks, data, and evaluation.

[8] J. E. Engelberg and C. A. Parsons. The causal impact of media in financial markets. *J. of Fin*, 66(1):67–97, 2011.

[9] Y. Fang, L. Si, N. Somasundaram, and Z. Yu. Mining contrastive opinions on political texts using cross-perspective topic model. WSDM '12, pages 63–72.

[10] G. Gidofalvi. Using news articles to predict stock price movements, 2001.

[11] M. Hagenau, M. Liebmann, and D. Neumann. Automated news reading: Stock price prediction based on financial news using context-capturing features. *Decision Support Systems*, 55(3):685 – 697, 2013.

[12] A. Hald. On the history of maximum likelihood in relation to inverse probability and least squares. *Statist. Sci.*, 14(2):214–222, 05 1999.

[13] D. Headey and S. Fan. Anatomy of a crisis: the causes and consequences of surging food prices. *Agricultural Economics*, 39(s1):375–391, 2008.

[14] R. Heakal. Explaining the world through macroeconomic analysis. 2012.

[15] T. Hofmann. Probabilistic latent semantic indexing. SIGIR '99, pages 50–57, 1999.

[16] F. Hogenboom, M. de Winter, F. Frasincar, and U. Kaymak. A news event-driven approach for the historical value at risk method. *Expert Systems with Applications*, 42(10):4667 – 4675, 2015.

[17] N. Kawamae. Trend analysis model: trend consists of temporal words, topics, and timestamps. WSDM '11, pages 317–326, 2011.

[18] J. D. Lafferty, A. McCallum, and F. C. N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. ICML '01, pages 282–289, 2001.

[19] Z. Li, B. Wang, M. Li, and W.-Y. Ma. A probabilistic model for retrospective news event detection. SIGIR '05, pages 106–113, 2005.

[20] S. Liao and R. Grishman. Using document level cross-event inference to improve event extraction. ACL '10, 2010.

[21] C. Luo, J.-G. Lou, Q. Lin, Q. Fu, R. Ding, D. Zhang, and Z. Wang. Correlating events with time series for incident diagnosis. KDD '14, pages 1583–1592, 2014.

[22] C. D. Manning, P. Raghavan, and H. Schütze. *Introduction to Information Retrieval*. Cambridge University Press, 2008.

[23] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. NIPS '13, pages 3111–3119.

[24] F. Ming, F. Wong, Z. Liu, and M. Chiang. Stock market prediction from wsj: Text mining via sparse matrix factorization. ICDM '14, 2014.

[25] M. Mohebbi, D. Vanderkam, J. Kodysh, R. Schonberger, H. Choi, and S. Kumar. Google correlate whitepaper.

[26] S. Nallareddy and M. Ogneva. Predicting restatements in macroeconomic indicators using accounting info, 2014.

[27] B. O'Connor and D. Bamman. Computational Text Analysis for Social Science: Model Assumptions and Complexity. *public health*, 2011.

[28] A. M. Okun. Economics for policymaking. 2004.

[29] K. Radinsky and E. Horvitz. Mining the web to predict future events. WSDM '13, pages 255–264. ACM, 2013.

[30] C. Rudin, B. Letham, and D. Madigan. Learning theory analysis for association rules and sequential event prediction. *J. of Mach. Learn. Rsrch*, 14:3441–3492, 2013.

[31] R. P. Schumaker and H. Chen. Textual analysis of stock market prediction using breaking financial news: The azfin text system. *ACM Transactions on Information Systems (TOIS)*, 27(2):12, 2009.

[32] D. Shahaf and C. Guestrin. Connecting the dots between news articles. KDD '10, pages 623–632. ACM, 2010.

[33] J. Si, A. Mukherjee, B. Liu, Q. Li, H. Li, and X. Deng. Exploiting topic based twitter sentiment for stock prediction. In *ACL (2)*, pages 24–29. The Association for Computer Linguistics, 2013.

[34] C. A. Sugar, Gareth, and M. James. Finding the number of clusters in a data set: An information theoretic approach. *Journal of the American Statistical Association*, 98:750–763, 2003.

[35] P. C. Tetlock. Giving content to investor sentiment: The role of media in the stock market. *The J. of Finance*, 62(3):1139–1168, 2007.

[36] S. Tripathi. The importance of knowing one's onions, January 2011. [Online; accessed Feb-2016].

[37] C. K. Vaca, A. Mantrach, A. Jaimes, and M. Saerens. A time-based collective factorization for topic discovery and monitoring in news. WWW '14, pages 527–538, 2014.

[38] Y. Wang, E. Agichtein, and M. Benzi. Tm-lda: efficient online modeling of latent topic transitions in social media. KDD '12, pages 123–131. ACM, 2012.

[39] F. M. F. Wong, Z. Liu, and M. Chiang. Stock market prediction from WSJ: Text mining via sparse matrix factorization. *Arxiv preprint*, 2014.

[40] B. Wuthrich, V. Cho, S. Leung, D. Permunetilleke, K. Sankaran, and J. Zhang. Daily stock market forecast from textual web data. In *Systems, Man, and Cybernetics, 1998.*, volume 3, pages 2720–2725. IEEE, 1998.

[41] W. Zhang and S. Skiena. Trading strategies to exploit blog and news sentiment. *ICWSM*, 2010.

[42] D. Zhou, D. Zhong, and Y. He. Event trigger identification for biomedical events extraction using domain knowledge. *Bioinformatics*, 20:1587–1594, Jun 2014.

[43] J. Zhu, A. Ahmed, and E. P. Xing. Medlda: Maximum margin supervised topic models for regression and classification. ICML '09, pages 1257–1264, 2009.