IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received February 2, 2021, accepted February 10, 2021, date of publication February 12, 2021, date of current version February 25, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3059283

One Step Ahead: A Framework for Detecting Unexpected Incidents and Predicting the Stock Markets

ZIYUE LI, SHIWEI LYU, HAIPENG ZHANG, (Member, IEEE), AND TIANPEI JIANG
School of Information Science and Technology, ShanghaiTech University, Shanghai 201210, China

Corresponding author: Haipeng Zhang (zhanghp@shanghaitech.edu.cn)

This work was supported in part by the Shanghai Sail Program under Grant 19YF1433800, and in part by the Key Projects of Shanghai Soft Science Research Program under Grant 20692194300.

ABSTRACT Unexpected incidents can be destructive or even disastrous, affecting the financial markets. Incidents such as the 9/11 attacks (2001), the Fukushima nuclear disaster (2011), and the COVID-19 outbreaks (2019, 2020) severely shocked both local and global markets. For investors, it is crucial to quantify the key facts that affect the incidents' impacts, and to estimate the reactions of the markets accurately and efficiently for global event-driven investment strategies. Though Web data and other alternative data allow such a possibility, it is still very challenging to mine noisy and often biased heterogeneous data sources, and construct a unified framework for modeling global markets across across time and regions. As a first attempt, we build a framework that extracts incident facts globally based on a deep neural network, feeds them into models built on a global event database complemented with novel socioeconomic datasets (e.g. nightlight data from satellites), and predicts stock market directions in a simulated real-world setting with interpretable results that outperform various baselines. Specifically, we study terrorist attacks in three countries for over 20 years on average, as a first effort to systematically quantify the impact on stock markets at a large scale using novel indicators.

INDEX TERMS Satellite data, stock market prediction, terrorist attacks, unexpected incidents.

I. INTRODUCTION

Unexpected incidents such as terrorist attacks, natural disasters, armed conflicts, disease outbreaks, and accidents involved with human errors (e.g. nuclear disasters, aviation accidents) can affect not only people's daily lives but also various industries and even regional economic development [1], [2]. They can alter investors' expectations, causing them to increase risk premiums or even withdraw from the stock markets in panic, leading to changes in stock prices [3]. For instance, when the US markets reopened after the 9/11 attacks, the Dow Jones Industrial Index Average (DJIA) fell 7.1% with panic and lost over 14% (1.5 trillion US dollars in market value) in 5 days. After the Fukushima nuclear disaster, the Nikkei 225 index fell 15.6% in two days, while global markets plunged as well. The COVID-19 outbreaks have deeply disrupted the global markets. For instance, the SSE Composite Index in China dropped 8.7% when the market reopened after a nationwide lockdown took place during the Spring Festival holidays and the DJIA fell over 11% on March 16, suffering its worst day since the market crash in 1987. Therefore, it is very important for market participants to get detailed information of major events, estimate their impacts, and adjust asset allocation accordingly, all in a timely fashion.

Case studies in finance and economics research explore severe events and their consequences. Studies show that factors like casualty and attack intensity in a terrorist attack contributes more to its impacts on stock markets [4], [5]. As these studies rely on traditional event data curated by humans and updated at very low frequencies, they have not attempted systematic analysis and market prediction in a real-world setting at a large scale.

From a computer science perspective, research suggests that unstructured data (e.g. Web data) and unconventional data (e.g. satellite data) can help us quickly and precisely capture events and depict them from multiple angles. For example, researchers detect earthquakes and extract their

The associate editor coordinating the review of this manuscript and approving it for publication was Xianzhi Wang.

locations from Twitter data [6]. Nightlight data from satellites indicates regional economic development levels which can further quantify conflict intensity [7], [8]. Though these efforts show the possibilities, they have not been customized, integrated, or tested for the purpose of detecting events, modeling the markets, or predicting them in real time.

There is much research on using Web data to predict the stock markets. For example, researchers extracted aggregated public mood indices from Tweets and news to predict the market index [9], [10] while more recent work applies deep neural networks on Tweets and news articles to predict individual stock movements [11], [12]. In contrast, we are interested in discovering significant incidents and predicting how they will move the market. We choose a path that bridges the gap between finance research and computer science research on stock market prediction, such that investors know clear causal relationships between incidents and the predicted market movements with explainable models and transparent decision logic.

In this paper, we design a framework that:

1. detects a wide range of unexpected incidents globally, including terrorist attacks, natural disasters, and disease outbreaks from online news and social media with a compound classification model, and extracts comprehensive facts;

2. builds learning models to predict the markets' directions given unexpected incidents (we choose terrorist attacks), based on a historical event database complemented with socioeconomic indicators from sources including satellite images; and

3. feeds the incident details into the models to predict market movements, with interpretable decision logic.

Our analysis on terrorist attacks in three countries for over 20 years on average confirms their impact on the markets as well as the effectiveness of socioeconomic features such as the nightlight intensity from satellites. We evaluate the system and showcase its effectiveness and interpretability in a simulated real-world setting.

In the remainder of this paper, we first introduce the related work in Section II. We then present the framework in Section III, after which we perform the experiments and evaluation in Section IV. Finally, we conclude this paper in Section V.

II. RELATED WORK

Here we review the most relevant work on event detection, unexpected incident study and stock market prediction.

A. EVENT DETECTION

Existing event detection algorithms can be broadly classified into two categories: document-pivot methods and feature-pivot methods. The former detect events by clustering documents based on the semantic distance between documents or building a document classifier, while the latter study the distributions of words and discovers events by grouping words together. For the detection of unexpected incidents, Sakaki *et al.* use a probabilistic spatio-temporal model to detect earthquake events from Twitter [6], while Radinsky and Horvitz use a keyword dictionary to detect disease outbreaks and terrorist attacks from news [13]. More recently, Petroni *et al.* co-reference both newswire text and social media to extract seven types of critical events [14]. These inspire us to compile a dictionary and build a Convolutional Neural Network (CNN) based text classifier as an improvement. In order to extract incident information, deep neural networks [15] and rule-based methods [16] have been applied. Radinsky *et al.* use many pattern and grammatical relations to extract information from news titles [13]. The Global Database of Events, Language, and Tone (GDELT) project [17] automatically detects and labels global events from news media for general application scenarios. For the purpose of market prediction, we monitor a wide range of events from direct data sources with instant access and extract task-relevant information to better meet its specific requirement on data and information.

B. UNEXPECTED INCIDENT STUDY

Case studies demonstrate that incidents including terrorist attacks, natural disasters, armed conflicts, disease outbreaks, and accidents involved with human errors can affect the stock markets [18]–[20]. Among these incidents, terrorist attacks occur frequently [21], often with considerable impact [22]. Sandler and Enders show prevailing terrorist attacks in smaller economies (specifically, Israel, Colombia, and Spain) affect the economy more [23], by impacting people's daily activities and hindering productivity. Aslam and Kang and Drakos discuss different factors (severity and the level of psychosocial effects) that determine the incidents' impact on the markets and economy [4], [5]. Quantitative research extends the subjects to longer time spans. Eldor and Melnick study data spanning over 13 years for Israel, showing that more casualties lead to a more negative impact on the stock market index [24]. Novel datasets have been sought to depict the incidents. For instance, Levin *et al.* use changes in Flickr photos, nightlight intensity and news items as indicators of intensity of conflicts [7]. Though we aim at market prediction in a real-world setting utilizing unstructured and unconventional data, these static studies shed light on choosing research subjects, constructing features and enriching them with unconventional data.

C. STOCK MARKET PREDICTION

Traditionally, stock market prediction relies on structured or semi-structured data such as historical stock prices, trading volumes, order book structures [25], and company disclosure statements [26]. In recent years, more and more work has been augmenting the predictive models with Web data. Bollen *et al.* and Gilbert and Karahalios extract the aggregated public sentiment from Twitter and blogs to predict stock market indices [9], [27]. Schumaker and Chen use textual features from financial news articles to predict stock price with a support vector machine (SVM) classifier [28]. Though

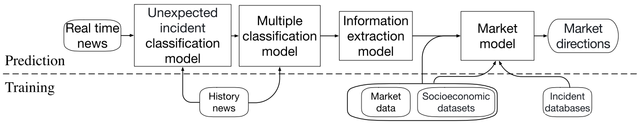


FIGURE 1. Framework architecture.

these studies explore text for signals, we are not aware of work that has detected unexpected incidents to predict the stock market.

From the methodology perspective, regression, neural networks, SVM, and decision tree are tools for stock market prediction [12], [28], [29]. With the success of deep learning in computer vision and Natural Language Processing tasks, there have been efforts on applying them to market and textual data (news and Tweets) to predict the stock market. Zhang *et al.* apply Recurrent Neural Network (RNN) on market data to make long and short term stock price predictions [30]. When applied to textual data, the assumption is that the unstructured text contains information that drives the stock prices. Ding *et al.* propose a deep learning method for event driven stock market prediction [31]. Xu and Cohen present a novel deep generative model jointly exploiting text and price signals to predict stock price [11]. Hu *et al.* design Hybrid Attention Networks to predict stock trends based on the sequence of recent related news [12].

Compared with application domains such as facial recognition and machine translation, deep learning's lack of explainability and interpretability is a concern for critical domains like health and finance where single wrong decisions can endanger lives [32] or cost significant monetary loss [33]. Relying on a data-driven trading system with incomprehensible reasoning can be risky. Especially when the market changes, it is difficult, if not impossible, to find out in limited time why a failed 'black-box' strategy failed, as an effort to reduce the loss. This makes us consider explainable models for more transparent market predictions.

III. A FRAMEWORK FOR INSTANT EVENT DETECTION AND MARKET PREDICTION

We aim to build a framework that instantly detects unexpected incidents, extracts key facts, and feeds them into a market model to predict the market movements with transparent and explainable decision logic.

Figure 1 shows its architecture. First, we cascade two classifiers to decide whether a piece of text reports an incident and which category of incidents it belongs to. According to the category, an information extraction model extracts facts such as time, location, casualty, victims, weapon (for terrorist attacks), magnitude (for earthquakes), and category (for hurricane). To predict the market, we build a model on market data and historical event data cross-referenced with novel socioeconomic datasets such as satellite data. In a real-world setting, the system receives pieces of text and outputs facts about incidents, market direction predictions, and demonstrations of decision logic.

In the following subsections, we describe the data types this framework relies on and the modules for the tasks mentioned above.

A. DATA TYPES

The framework is primarily based on four types of data.

Instant textual data from the Web in the form of news articles and micro-blogs which may contain instant reports of incidents. They are used for building text classification and information extraction models. In a real-world setting, they would be obtained instantly and fed into the framework for real time market predictions.

Market data is the price and trade-related data with timestamps for a financial instrument which can be a stock or a stock market index (an aggregation of the market).

Historical incident databases depict incidents with attributes such as time, location, severity, people involved, damage, and cause, upon which we build the market models. They are usually manually curated with relatively low update frequencies such that they can be considered static. High-quality databases include the Global Terrorism Database (GTD), the Emergency Events Database, and the Armed Conflict Location and Event Data.

Socioeconomic databases include attributes that measure regional development levels of economy, education, culture, and politics. The intuition is that two identical incidents that happen in different places (e.g. one in a well-developed area and one in an underdeveloped area), may have different degrees of impact. Attributes such as GDP, household income, infant mortality rate, and population density can be cross-referenced by locations to depict the incidents comprehensively. Commonly used databases include CIA World Factbook and data from NASA's Socioeconomic Data and Applications Center (SEDAC).

As mentioned in Section I, unconventional data such as nightlight intensity has been proved a reliable and consistent indicator of economic development across all regions with a fine granularity [34] while not all the regions have GDP data and not all the GDPs are calculated under a universal measurement.

In this paper, we choose to experiment on a specific type of incidents – terrorist attacks. Accordingly, we utilize the Global Terrorism Database as the source of historical incidents and socioeconomic data including nightlight intensity, population density, greenhouse gas emission, and facts about locations from Wikipedia. We will elaborate on these datasets in Section IV-A.

B. INCIDENT DETECTION

In order to make sense of news data, we start with detecting unexpected incidents. Similar to the work of Petroni *et al.* [14], we first use a list of incident keywords to filter out many irrelevant ones. The keywords are these appearing much more in incident-relevant articles than in regular ones. The remaining articles will be classified into their corresponding incident categories. We observe that the numbers of occurrences for different incident types vary greatly, resulting in imbalanced training sets. To address this, the news is further classified through two models: a binary classifier based on a CNN to decide whether the news is about an unexpected incident, and a decision tree for specific categories of unexpected incidents, such as flood, blizzard, disease outbreaks, and terrorist attacks. The design consideration for this two-stage strategy is that deep learning approaches usually have better performance when there is enough training data and therefore we use CNN in the first stage as a binary classifier on the incident VS non-incident classification task where the training data is more balanced. For the second stage, though some categories such as 'H5N1' and 'flood' have very limited training data, their news titles contain highly distinguishable key words according to our observations. For instance, 74.0% occurrences of the word 'flood' are in more 83.0% of the flood incident news titles. This makes decision tree a fit in our scenario.

1) UNEXPECTED INCIDENT CLASSIFICATION MODEL

We propose a CNN-based classification model built on a news corpus with binary labels (incident or not incident) to detect the unexpected incidents. The model architecture is adapted from a shallow-and-wide CNN [35], which achieves generally good performance on text classification. We improve it by adding an attention mechanism with the intuition that it would help better incorporate contextual relationships. We begin with a tokenized piece of text and then convert it to a matrix, of which each row is a word vector representation for each token. These rows output from pre-trained GloVe [36] models. An attention layer is added after the convolution layer. The model details are described in Appendix D.

2) MULTIPLE CLASSIFICATION MODEL

We use a decision tree trained on news text labeled with specific incident categories to further classify the incidents. Each piece of text is represented using a vector space model.

C. INFORMATION EXTRACTION

We extract critical information about the incidents, such as locations, casualties, severity for natural disasters, and weapons for terrorist attacks, from news titles. Though previous studies have dealt with similar tasks [13], we attempt to use less human labor and expert knowledge to automate the process.

For locations and time, the extraction rules are straightforward. Locations are extracted by putting thresholds on the Jaccard similarity coefficients between the title words and the location names in the geographical database GeoNames.[1] Since the incident time rarely appears in titles, we use the publish date as an approximation.

For death numbers and injury numbers, the extraction rules are not obvious to compose and even experts may fail to cover all the rules. We learn the rules from the perspective of dependency grammars with heuristics. To create training samples, we label news titles with victims that are dead or injured with numbers if there are any. For instance, the sentence '5 children killed and 3 injured on Line 4' is labeled as (dead: 5 children; injured: 3). We then use the lexical database WordNet[2] to extract the candidate victims (the ones in the 'person' category) with numbers. In this example, the candidates are '5 children', '3', and '4'. After parsing them with Stanford CoreNLP [37], we obtain dependency tuples ('5 children', 'direct object', 'killed'), ('3', 'direct object', 'injured'), and ('4', 'numeric modifier', 'Line'). With the first two tuples being matches for deaths and injuries, it can be generalized that dependency pairs ('direct object', 'killed') and ('direct object', 'injured') respectively indicate that the death numbers and injury numbers are found. Therefore, we calculate the probability of matching the desired information for each dependency pair in the training samples and pairs with high probabilities are deemed extraction rules.

Besides casualties, we extract information related to specific incident types. For instance, earthquakes and typhoons have structured ways of representing magnitudes and levels which can be extracted using predefined rules [16]. For terrorist attacks, weapon types (explosives, firearms, incendiary, melee, and others) are extracted by a text classifier. As the labeling of training data often involves a lot of expert knowledge, we use incident databases compiled by experts to mitigate this problem. The GTD summarizes the attacks and can be engineered into a labeled training set. Specifically, for each weapon type, we compute the most representative words from the summaries using TF-IDF. Text classifiers (e.g. SVM) are built upon features from the incident database and evaluated with unlabeled news titles.

D. MARKET MODELS

Here we formalize the prediction problem, propose features and introduce the predictive models.

1) PROBLEM STATEMENT

We use market data and incident-related information to predict market movements. We consider it as a binary classification problem. Typically, the market return for stock S on day t is defined as

$$r_{\text{reg}}(t) = \frac{S_c(t) - S_c(t-1)}{S_c(t-1)} \tag{III.1}$$

where $S_c(t)$ represents stock S's closing price on day t. However, for day t, if an incident happens after the market closes, its impact on S would not be reflected in $r_{\text{reg}}(t)$. Thus, we propose another definition of the daily return which measures the ratio of increase using the day $t+1$'s opening price instead of day t's closing price, to ensure that incidents that happen on day t would always be included in between the two market price timestamps. The pre-market and after-hours trading and the call auction mechanism for many markets to

[1] https://www.geonames.org

[2] https://wordnet.princeton.edu/

obtain the opening prices, would have by design priced in much of the information during the non-trading hours. For given day t and stock S, the return on day t is defined as

$$r(t) = \frac{S_o(t+1) - S_c(t-1)}{S_c(t-1)} \tag{III.2}$$

where $S_o(t)$ and $S_c(t)$ represent day t's opening and closing prices.

The market movement label $y(t)$ is set to 1 if the sign of Equation (III.2) is negative. Otherwise, it is set to 0. Denoted by $f \in \mathbb{R}^N$, the N-dimensional feature vector depicts the market and the incidents by attributes such as market returns, disaster severity, number of deaths, and nightlight intensity. The problem is then stated as: predicting market movement $y(t)$, either 1 or 0 on day t, with corresponding feature vector f.

2) CANDIDATE FEATURES

In this part, we propose potential features that could be extracted from the data mentioned in Section III-A. We then describe how they are further selected.

Previous studies give us hints about features that can predict market movements. The past performance of the market itself with that of other markets may indicate its future performance, which autoregressive models for financial time series forecasts are based on [38]. As an influential global economy, the US stock market indices sometimes correlate with or even lead other countries' markets [24], [39]. As discussed in Section I, the incidents' attributes may help us quantify their impact on the markets. For terrorist attacks, such attributes include casualty, level of psychosocial effect, and attack intensity [4], [5], while for natural disasters, economic losses are important [40].

Socioeconomic data helps us further understand the incidents and estimate the economic or psychosocial damage. For instance, a terrorist attack's damage may be magnified if it happens in an economic, political or religious center. This kind of facts can be looked up in regular datasets such as DBPedia and government data. However, these datasets can be sometimes incomplete, inaccurate or outdated as they rely on manual collection and compilation. Unconventional data can be a good replacement.

3) MARKET DIRECTION CLASSIFICATION

We build classification models with historical event databases, market data, and socioeconomic datasets and experiment on different feature combinations for improvement.

To keep the models updated with the most recent information, we re-train the models whenever a new sample comes in with a rolling window. The initial training set with the first N observations is adopted to train the model, and one-step-ahead prediction is produced at the $(N+1)$-th observation. Then, the training set is modified by adding one new observation and withdrawing the oldest one. Next one-step-ahead prediction is produced at the $(N+2)$-th observation. The live information stream is first simulated with the annually updated GTD dataset. To make it close enough to the real-world scenario, we instead get the incident information from news processed by the incident detection and information extraction modules. In addition, we attempt deep learning methods that work well on other stock market prediction tasks using much larger news corpora, namely a Hybrid Attention Network (HAN) based on Gated Recurrent Unit (GRU) [12] and an LSTM based method [41]. The HAN method consists of two attention layers, one for news attention and one for temporal attention, to analyze news articles in the sequential temporal context and pay more attention to critical time periods. Besides, by using the self-paced learning mechanism, it skips challenging training samples at early training stages, and incorporate them into later training stages. LSTM, as an artificial recurrent neural network architecture with feedback connections, is suitable for modeling timeseries data. The LSTM based approach converts news articles into distributed representations using Paragraph Vector and captures their temporal effects to predict stocks' opening prices.

E. EVALUATION METRICS

Methods described in Sections III-B, III-C, and III-D are evaluated using the following metrics.

1) BINARY CLASSIFICATION MODELS

For binary classification models including the incident detection and the market prediction models, we use precision and recall to evaluate the performance. Precision is defined as the proportion of correctly predicted positive cases (e.g. correctly predicted market downs) to all the cases predicted to be positive (e.g. predicted market downs) and recall is the ratio of the number of correctly predicted positive cases (e.g. correctly predicted market downs) to the total number of positive cases (e.g. total market downs). Additionally, F1 score is calculated by the harmonic mean of precision and recall. For incident detection models, we also compute accuracy as the proportion of correctly predicted cases to all the predicted cases. For the market direction prediction task, we only care to see whether unexpected incidents will cause the market to go downward. Therefore, we treat downward samples as positive samples and calculate precision, recall, and F1 scores only for them (accuracy will be equivalent to precision in this case).

2) MULTIPLE CLASSIFICATION MODELS

For the multiple classification model to classify the types of incidents, we randomly choose samples from results of the classifier and check whether the predicted type is correct. We evaluate the performance with precision for each incident type.

3) INFORMATION EXTRACTION

We randomly choose recent samples from results of the extraction model and manually check whether the extracted information is correct. We evaluate the performance by accuracy, defined as the proportion of the cases that are correctly extracted.

IV. EXPERIMENTS AND ANALYSIS

In this section, we evaluate the individual components of our framework, test the entire framework in a simulated real-world setting, and demonstrate the interpretability of our model with an example.

In Section IV-B, the proposed event detection models are evaluated with classification tasks against traditional classification models and convolutional neural network models. The results from the information extraction models are manually checked. We cover a wide range of incidents, including terrorist attacks, natural disasters, armed conflicts, and disease outbreaks.

For market modeling, we focus on a specific type of incident, terrorist attacks, in three countries – Israel, Colombia, and Spain. As mentioned in Section I and II, terrorist attacks often have impacts on stock markets and economies. According to the GTD, there are 6,580 attacks each year from 2000 to 2017 globally. For Israel, Colombia, and Spain, the yearly occurrences are 69, 124, and 25 respectively, and their stock markets are vulnerable to terrorist attacks. All these factors make them the subjects of our experiments.

Before evaluating the market models on historical event databases in Section IV-D, it is natural to ask whether the incidents have any impact on the markets and whether any features of the incidents are relevant to the degree of impact. In Section IV-C, we answer these questions by statistical tests on static historical terrorist attack data and stock market index data.

With individual components evaluated, we test the framework in a simulated real-world scenario in which we input instant news data and get the market direction predictions as outputs in Section IV-E, with a visual demonstration of the whole process and decision path. Based on this, we further analyze and interpret the results.

A. DATASETS FOR THE EXPERIMENTS

As mentioned in Section III-A, the framework is built upon four types of data. Here we describe the specific datasets used for our experiments.

1) NEWS DATA

The websites and Twitter accounts of news agencies are our sources of timely reports of unexpected incidents. We obtain news data from New York Times (NYT) API, Reuters, and Twitter. The NYT dataset contains titles, published date, URL, and labels of 1,989,151 articles published from 2000 to 2018. The Reuters data contains time, titles, and URLs of 9,529,834 articles from 2007 to 2018. Due to the availability of website data, the Reuters data from 11/2018 and 12/2018 are made up by news Tweets from its official Twitter account.

FIGURE 2. Nightlights for part of the US in 2016.

Though labeled, the NYT data cannot serve as positive training data because the incident labels do not distinguish articles that report the incidents (e.g. terrorist attacks) from articles that just talk about a general topic (e.g. condemning terrorism). Therefore, in order to obtain high-quality training data labeled with incident types, we scrape 16,331 news titles for 12 incident categories as references from Wikipedia pages. For instance, we use the reference news titles from the page *'List of terrorist incidents in 2008'* as samples labeled 'terrorist attack'.

2) MARKET DATA

We obtain major stock market index data for Israel, Colombia, and Spain, namely the Tel Aviv Stock Exchange Index (TA100), the Colombia Stock Exchange General Index (IGBC), and the Madrid Stock Exchange General Index (IGBM), from Thomson Reuters Eikon, with daily highest, lowest, opening and closing prices. They span 25, 14, and 23 years respectively, until 12/31/2018. As mentioned in Section III-D2, we obtain the US S&P500 index data as it may contain predictive information for other markets. For the S&P500, its daily return is closing price change as a percentage of the previous day's closing price. We deal with time zones and make sure that information from S&P500 is always historical comparing with the target market to be predicted.

3) THE GLOBAL TERRORISM DATABASE (GTD)

The GTD[3] is an annually updated open-source database of terrorist attacks that documents more than 180,000 events from 1970 to 2017 worldwide with 135 attributes. It helps us characterize the events and build up market models. Besides the basic features such as time and place of occurrence with geo-location, number of casualties, attack type, weapon used and attacking target are also included.

4) SOCIOECONOMIC DATA

We use four socioeconomic datasets to depict the regions at fine granularities. Details on data processing are in Appendix C.

1. NASA's global nightlight intensity data[4] to measure the economic development for areas where the incidents happen. Figure 2 shows the nightlights for part of the US in 2016.

2. SEDAC's gridded global population density data[5] to estimate the number of people directly affected by each incident.

[3] https://www.start.umd.edu/gtd/

[4] https://earthobservatory.nasa.gov

[5] https://sedac.ciesin.columbia.edu/data/set/spatialecon-gecon-v4

TABLE 1. Binary classification results for incident detection.

Model	Precision	Recall	F1	Acc
BERT	87.0%	55.6%	67.8%	75.9%
TextCNN	85.9%	53.2%	65.7%	82.6%
Word shallow-and-wide CNN	89.6%	54.4%	67.7%	82.1%
Proposed model *2-(2,3,4)*	86.9%	72.3%	79.1%	82.4%
Proposed model *6-(3)*	90.2%	69.6%	78.6%	83.9%

3. SEDAC's gridded global greenhouse gas emission data estimated by the Intergovernmental Panel on Climate Change.[6] It is used by various studies to model population, GDP development, and land-use changes [42], [43].

4. Factual data about locations from Wikipedia. Whether they are within national or provincial capitals is used as a political label. Besides, we manually label the cities with 'cultural center' and 'religious center' according to Wikipedia, to address other attributes that may be overlooked in the above mentioned datasets.

B. INCIDENT DETECTION AND INFORMATION EXTRACTION

This section shows how we train, apply, and evaluate our incident detection and information extraction models. Due to data availability, we experiment on news titles, instead of full text. As suggested by previous research [44], news titles usually contain the most critical information, such as time, location, casualty, and severity, which makes them a sufficient source of information.

1) INCIDENT DETECTION EXPERIMENTS

After initial matching with an incident keyword list, the candidate news is first fed into a binary classifier to decide whether it is related to any unexpected incidents. The relevant news is further classified into 12 specific types, covering terrorist attacks, disease outbreaks (cholera, H5N1, and Ebola), and natural disasters (avalanche, blizzard, earthquake, flood, freezing rain, hurricane, landslide, tornado, and tsunami).

Specifically, we use the incident news titles from Wikipedia and the general titles for NYT to construct the incident keyword list described in Section III-B. This list is then applied to all data from Wikipedia, NYT, and Reuters to filter out irrelevant titles. The proposed binary classifier in Section III-B1 is trained and tested on the resulting 13,644 Wikipedia incident news titles (positive samples) and 16,246 NYT titles (negative samples). The negative samples are the titles matching incident keywords but not actually about incidents. The NYT data comes with labels indicating broad topics and the titles without incident relevant labels are kept as negative samples. We test hyperparameter combinations and compare with TextCNN [45], Word shallow-and-wide CNN [35], and BERT [46]. As shown in Tabel 1, the proposed model using three filter region sizes: 2, 3, and 4, each of which has 2 filters (2-(2,3,4)), hit the highest F1 score at 79.1% while the proposed model using six filters with region size being 3 (6-(3)) achieves the highest precision, at 90.2% with a slightly lower F1 score at 78.6%. Their F1 scores outperform other methods in comparison by at least 10.8%. Besides, their accuracy scores are within top 3. Though TextCNN's accuracy which ranks second is 0.2% higher than the proposed model with 2-(2,3,4) setting, its F1 score is largely outperformed by 13.4%. We choose the 2-(2,3,4) setting with a more balanced performance for upcoming experiments. We notice that BERT does not perform as well as our models. A possible reason is that the pre-trained BERT (bert-base-uncased) does not contain enough tokens – its size is 30,000 while our model relies on GloVe, which includes 2.2M tokens. For instance, city names and person names in news titles can be informative for deciding whether the articles are related to incidents, and they may be missing in the pre-trained BERT.

TABLE 2. Event type classification results.

Incident Type	Original		Expanded	
	Precision	Support	Precision	Support
Terrorist Attack	92.5%	613	92.5%	613
Earthquake	85.0%	40	82.0%	50
Flood	91.7%	108	91.7%	108
Tornado	83.3%	24	86.0%	50
Volcano	94.4%	18	92.0%	50
Avalanches	100.0%	1	80.0%	15
Freezing rain	100.0%	1	80.0%	15
Hurricane	78.1%	96	78.1%	96
Landslide	75.0%	12	73.3%	30
Tsunami	61.1%	18	70.0%	50
Cholera	87.5%	8	90.0%	20
Ebola	72.6%	61	72.6%	61

The multi-classifier described in Section III-B2 is trained on the labeled Wikipedia data and applied to classify the unlabeled Reuters data (71,115 titles) into the 12 types. We manually examine 1,000 samples randomly selected from the results and calculate the precision for each incident type, as shown in Table 2. From this, the accuracy for the multi-classifier is calculated to be 88.5%. As we can see from the table, these samples are also very unbalanced. In order to better evaluate the model, we include more samples for these types with less than 50 samples by randomly selecting more samples for them to expand to 50. For types whose total predicted samples are below 50, we include all their samples. As we can see from the Expanded Precision column in the table, terrorist attack, flood, volcano, and cholera reach the precision of over 90% and the score is 92.5% for terrorist attack, which ranks first. Types including hurricane, landslide, tsunami, and Ebola, have precision scores below 80.0%, with tsunami being the lowest (70.0%). If there are more training samples for these types, the performance may further improve.

2) INFORMATION EXTRACTION EXPERIMENTS

From the results of the previous step, we extract key information according to incident types. For instance, we extract

[6] https://sedac.ciesin.columbia.edu/data/set/ipcc-gridded-emissions

weapon type for terrorist attacks, disease type for disease outbreak and severity for disaster. For all incidents, we extract number of deaths, number of injuries, specific region and country. To evaluate, we randomly select 1000 samples that are correctly classified as terrorist attacks and 500 correctly classified samples for other categories. If any value is wrong or missing, it marks a wrong extraction. The accuracies for casualties, locations and weapons are all above 88.0%, with deaths and injuries hitting 91.7% and 97.9% respectively. The overall accuracy for disease and disaster related attributes is lower, at 74.0% and the reason might be that their rare occurrences in the corpus make it harder to construct universal extraction rules.

C. STATISTICAL ANALYSIS AND FEATURE SELECTION

To construct features for predicting the market, we perform empirical analysis with statistical tests to explore whether the incidents have impact on the market movements, whether any feature is related to the degree of such impact and whether the US market provides information for predicting the targeting markets. Based on this analysis, we construct the features and further select them.

1) STATISTICAL ANALYSIS

First of all, we check whether the terrorist attacks influence the market. For trading days with returns no more than x, we calculate y as the percentage of these days with terrorist attacks. For Israel (Figure 3(a)), the trend is that for the days when the market drops more, larger portion of them tend to have terrorist attacks. For instance, point B indicates that for days when the returns are no more than -3.4%, 10.3% of them have terrorist attacks while for point A, which corresponds to days when the returns are no more than -4.1%, the ratio goes up to 12.9%. This suggests that terrorist attacks lead to drops in market returns.

We are then curious to see whether any attribute of terrorist attacks affects the impact on the market. Intuitively, numbers of deaths and injuries can be direct indicators of the severity. We conduct a statistical analysis of the market index returns corresponding to different numbers of deaths and casualties. We analyze the mean and variance homogeneity in return by T and F-tests respectively. For Israel, we compare the returns on days without deaths from terrorist attacks and the returns on days with a least 1 or 2 deaths. When comparing days with 0 death and days with at least 1 death, the mean difference in return is statistically significant with p-value being 0.0335. When compared with days with least 2 deaths, the p-value is much smaller, which is 0.0011. Tests for Columbia and Spain show similar results. This suggests that terrorist attack severity has impact on the market movements.

Besides casualties, we explore the relationships between the attack location's nightlight intensity and the market performance. For a day with attack(s), we record the highest nightlight intensity value among the locations. We compute the portions of market downs for days with nightlight intensity values above thresholds. As we can see in Figure 3(b), incident days with larger nightlight intensity values have higher portions of market downs. For instance, point A shows that 49.3% of the incident days with nightlight intensity above 100 have market downs while the ratio increases to 53.1% for point B that represents incident days with nightlight intensity above 230. This convinces us of nightlight's usefulness.

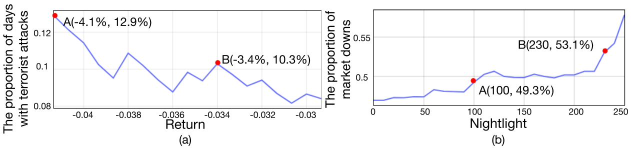


FIGURE 3. Empirical analysis on the impact of terrorist attacks and their locations' nightlight intensity on the market.

Studies show the US market may lead markets in other countries and it still holds with our data for the three countries according to Granger causality tests. This indicates that S&P500's information can be composed into features for predicting these markets.

2) FEATURE COMPOSITION

After compilation, we have 143 candidate features. 135 of them are from the GTD, including number of casualties, attack type, weapon used and attacking target. We also use historical returns of S&P500 and the target market, nightlight intensity, population density, greenhouse gas emissions data and political/cultural/religious labels to compose the features. We aggregate these features at a daily level, in case that multiple incidents happen on a same day. Details about the aggregation are in Appendix E.

3) FEATURE SELECTION

We apply a random forest to select features. With a threshold of 0.01 on variable importance, we get the following features in descending order: S&P500 trend indicator, number of injuries, number of deaths, nightlight intensity, number of terrorist attacks, province, historical trend of the target market index, population density, greenhouse gas emission, political indicator (capital or not), religious indicator and cultural indicator.

There may exist collinearity across the three socioeconomic features, namely nightlight intensity, greenhouse gas emissions and population density. Linear regression results suggest that the greenhouse gas emission feature can be well explained by the other two features and we therefore discard this feature. Details about the regressions can be found in Appendix F.

D. ANALYSIS OF HISTORICAL MODEL

In this section, we build classification models to predict whether the targeting market will go downward, given current and past incident and market information. The precision and recall here are calculated only for downward movements. For evaluation, two baselines are constructed. The first one is

TABLE 3. The performance of different feature combinations on the full sample set (Exp-FS) for Israel.

	Exp-FS for Israel				
Feature	Precision	Recall	F1	Support	Total
Market-only	58.4%	52.3%	55.2%	453	1000
+ Terrorist related	60.4%	54.3%	57.2%		
+ Socioeconomic (= Full-feature)	64.5%	55.6%	59.7%		

[1] The plus sign indicates that we add new features to the previous row of features.
[2] The precision and recall of random guess baseline are 45.3% and 47.4% in Exp-FS.

TABLE 4. Breakdown of the results from Exp-FS for Israel. The statistics are computed for days with and without terrorist attacks.

	Days with terrorist attacks			Days without terrorist attacks		
Features	Precision	Recall	F1	Precision	Recall	F1
Market-only	51.6%	47.0%	49.2%	59.5%	53.2%	56.2%
Full-feature	64.8%	53.0%	58.3%	64.4%	56.1%	60.0%

TABLE 5. Overall results from models trained on days with attacks (Exp-Terr) for three countries.

Algorithms	Precision	Recall	**F1**	Support	Total
Decision tree	70.6%	57.6%	63.4%	250	513
Logistic regression	61.2%	60.0%	60.6%		
Random forest	63.0%	54.4%	58.4%		
SVM	52.3%	50.0%	51.1%		

[1] The overall precision and recall of random guess baseline are 48.7% and 49.0%.

random guess according to the data distribution, with precision being proportion of market downs in the test set and recall being the proportion of market downs in the training set. The second baseline model is fitted with market data, denoted by market-only.

With regard to different features, we intend to see whether incident features and socioeconomic features help. We then compare the learning algorithms described in Section III-D3. The experiments conducted in this section only involve static data for the purpose of verifying, analyzing and improving. To be specific, we treat the incidents from the GTD as if they were captured instantly, even though the GTD is updated annually. Compared with news, the GTD information is much more complete with reliable manual compilation and verification which makes it a more suitable data source for the aforementioned purpose.

To train and validate the models, we split the data into a training set and a testing set for each country. The training set is further split to obtain a validation set for tuning the hyperparameters. Details of this practice are described in Appendix G.

The experiment starts with the full sample set (Exp-FS). We compare different combinations of features, including market features, incident features and socioeconomic features using a decision tree model for Israel. As shown in Table 3, when only market features are used, it achieves a precision of 58.4% and a recall of 52.3%, comparing with the random guess baseline of 45.3% and 47.4%. When the incident features are added, the precision increases by 2.0%, to 60.4% and the recall increases by 2.0%, to 54.3%. We then further add in socioeconomic features, with all features combined, the model reaches a precision of 64.5% and a recall of 55.6%, both outperform the market-only features. For the other two countries, we see similar improvement. When we combine the results for three countries together, we get a precision of 68.7% and a recall of 60.6% using all features, as opposed to 66.8% and 57.1% using market data only.

We further breakdown these results to see how incident features and socioeconomic features improve the performance. For Israel, we calculate the precision and recall for the 162 days with terrorist attacks and the 838 days without terrorist attacks independently. Table 4 shows that when the incident features and socioeconomic features are added, the performance for the days with terrorist attacks has greater improvement, with precision increased by 13.2% to 64.8% and recall increased by 6.0% to 53.0%. Meanwhile, for the days without attacks, it also shows moderate improvement. A possible explanation is that for 17.8% of the days without attacks, there are attacks in the past two days. This information recorded in the lagged features may help the predictions. For the other two countries, we get similar observations.

The observation that days with attacks have better performance improvement suggests solely training and testing on data from days with attacks (Exp-Terr). This gives us 320 days for training, 95 days for validating and 162 days (66 market down instances) for testing. The model achieves a higher precision of 68.9% and a lower recall of 47.0%, comparing with 64.8% and 53.0% from the model trained on the full samples. When we combine the results for the three countries, we also see improved precision. The overall precision and recall are 70.6% and 57.6% with a support of 250, comparing 59.5% and 50.0% for market-only features. Results outperform both baselines by large margins, especially for precision (by 21.9% and 11.1% respectively).

We compare decision tree, random forest, logistic regression, and SVM with the same setting for Exp-Terr. Decided by grid-searches, we use the limited-memory BFGS method for logistic regression and the Gaussian kernel functions for SVM with optimized parameters. Table 5 suggests that decision tree is a better choice with the highest precision and F1 score at 70.6% and 63.4% respectively. As discussed earlier, interpretability is preferable here. The decision tree provides not only insights on feature importance and contribution, but also a clear visualization of the logic. Figure 4 shows such a sample to be interpreted in Section IV-F.

Besides these methods with careful feature engineering, we also attempt the end-to-end deep learning methods mentioned in Section III-D. We use the Reuters data for Israel and TA100 data (open, high, low, and close prices) from 2007 to 2015 as training data, and the data from 2016 to 2018 as test data. For the HAN model by Hu *et al.* [12], we use a learning rate of 0.001 and a dropout rate of 50%. For Akita *et al.*'s LSTM based model [41], we keep its original parameters. The results from both models show higher precision compared with the random guess baseline (60.0% and 53.9% compared

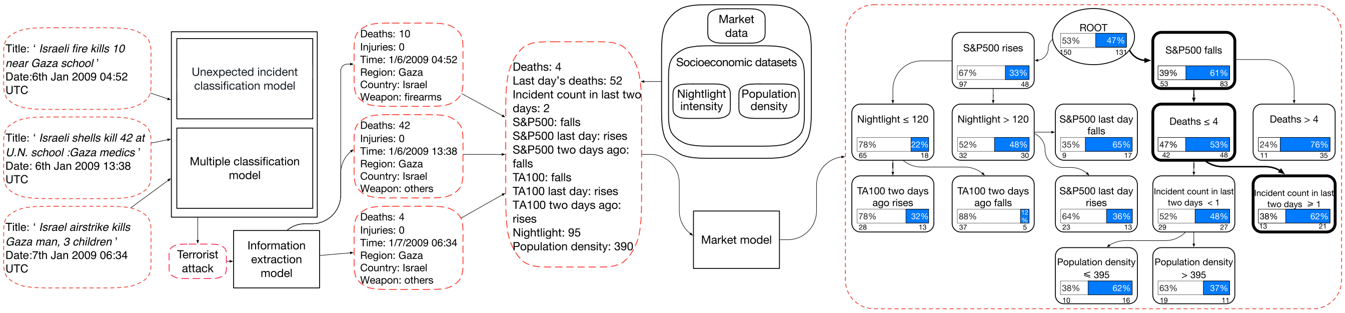


FIGURE 4. A sample decision tree outcome. In each tree node, the percentage and number on the left correspond to market ups in the training set and these on the right correspond to market downs.

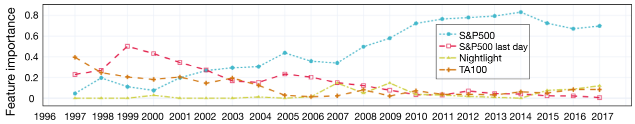


FIGURE 5. Trends in feature importance scores of top four features between 1997 and 2017, training in a moving window of 5 years using decision tree.

with 53.6%) while the recall are extremely low (5.7% and 8.6%). The differences in tasks and data might have caused the drop. These models were originally used for predicting individual stocks with much larger news datasets which may contain enough information for predicting one stock's movements in short time windows. A stock index representing a basket of stocks is affected by various factors and some of them, such as macroeconomic factors [47], are not always reflected in the news.

E. A REAL-WORLD SCENARIO

We further test the entire framework by feeding it with news data which can be instantly obtained in a real-world scenario. The predictions are made by a model trained on the most updated knowledge, with news being the sole source of new incident information.

We expect the dynamic models would capture the most updated knowledge in an ever changing world. Figure 5 illustrates how feature importance scores change over time. In a moving window of 5 years, we tune a decision tree model that best predicts the market in the 6-th year and compute the feature importance (Gini importance) scores. In Figure 5, for the purpose of illustration, we show top four features according to their importance scores. The trend is that the SP500 feature and the TA100 feature from the previous day are becoming less important while SP500 feature from the current day and the nightlight feature are gaining importance. This indicates the necessity of incorporating the most updated information into the models.

As described in Section IV-A1, we prepare a news corpus containing 9,529,834 samples from 2007 to 2019. 17,303 samples are classified as terrorist attacks. The attacks' information is further extracted. With manual verification, the accuracies for the classification and extraction are 91.3% and 93.7%. For Israel, the scores are 96.0% and 93.7%. We then apply the extracted information to market models for one-step-ahead predictions.

In the prediction process, the model is trained on the dataset with a fixed size rolling window of 281 days. For incident features, the initial training set entirely relies on the GTD data, market data, and the socioeconomic datasets. To keep the models dynamically updated, after each prediction, the information extracted from the news is added to the training set. Other setting is the same as that described in Section IV-D.

Similar to Exp-FS, there is an increasing trend in precision and recall as we add in more features for the three countries, all outperforming the market-only baselines. The overall precision and recall for the three countries are 68.4% and 54.5% and for Israel, it achieves a precision of 71.2% and a recall of 48.8%, compared with a precision of 66.1% and a recall of 45.3% for the baselines.

F. DISCUSSION AND RESULT INTERPRETATION

When we use the GTD to simulate the instant information (Section IV-D), the overall results show a precision of 70.6% and a recall of 57.6%, outperforming the baselines by large margins (21.9% and 8.6%), while these metrics become slightly lower when we use the news data instead of the GTD, as described above. We hope this can be improved when we have multiple paralleled sources of news.

Figure 4 shows such a decision tree from the results in Section IV-E. On 1/6/2009 and 1/7/2009, three news titles are detected to be about terrorist attacks, with key facts including countries (Israel), regions (Gaza), deaths (10, 42, 4), injuries (0, 0, 0), and weapons (firearms, others, others). They are cross-referenced with market and socioeconomic datasets and aggregated into a vector for 1/7/2009. The market is predicted to drop because: 1. S&P500 index drops today; 2. though the number of deaths is no more than 4, there are attacks in the last two days. This is in line with common experience: 1. local market is affected by the US market; 2. recent attacks may still impact the current market. An experienced investor may make a same prediction given this information. Other paths in the tree also shed light on interpretable rules. For instance, though the S&P500 goes up today, if the nightlight intensity is large enough and S&P500 goes down yesterday, the model will predict that the market will go down, meaning that an incident that happens in a more developed area has downward impact on the market.

V. CONCLUSION AND FUTURE WORK

In this paper, we demonstrated our framework for detecting unexpected incidents and predicting the market directions with explainable decision logic. We experimented on terrorist attack incidents in 3 countries for over 20 years on average and nightlight was proved to be an effective indicator of market impact. The incident detection and extraction modules hit accuracy of 91.3% and 93.7%, respectively. In all scenarios, the market prediction module outperformed the baselines. For instance, in the historical model setting, it achieved an overall precision of 70.6%, outperforming the baseline by 21.9%. We open source our implementation.[7]

There are many possible future directions. First, we can predict the impacts of more categories of incidents such as natural disasters and disease outbreaks in more regions. Second, we can mine the news article content to extract more information and use external knowledge bases to infer missing attributes. Finally, as it is hard to capture transient market impact at current data granularity (daily) and our experiments are restricted to market direction predictions, we hope finer-grained market data would allow better measurement or even price predictions.

APPENDIX SUPPLEMENT

We open source the implementation at https://github.com/one-step-ahead-result/one-step-ahead-code, with sample data and hyperlinks for obtaining the complete data.

A. PYTHON PACKAGES

We use Pytorch[8] package to implement the deep neural network model. For the decision tree model, logistic regression model random forest model and SVM model, we use the scikit-learn package.[9]

B. COMPUTING RESOURCE

For all experiments, we use the same computing resource which has 56 CPU cores, 256GB of RAM, and 4 RTX2080 GPUs.

C. SOCIOECONOMIC DATASET DETAILS

We provide details on the socioeconomic datasets described in Section IV-A4.

1) NASA's GLOBAL NIGHTLIGHT INTENSITY DATA

The original grayscale data obtained from NASA is a single snapshot, made by aggregating the nightlight intensity values measured in 2016 and mapping them into corresponding pixels. It has a 3 km by 3 km resolution and each pixel's value range is from 0 (dark) to 255 (max intensity). When using it as feature for a certain region, we choose the nightlight intensity value of the pixel closest to the center of that region.

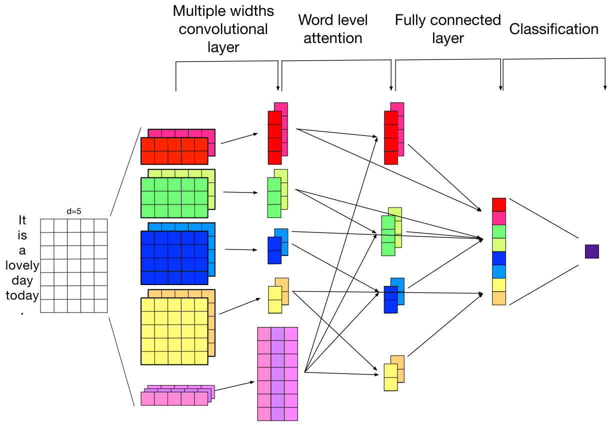


FIGURE 6. The architecture of the binary classification model.

2) SEDAC's GLOBAL POPULATION DENSITY DATA

The dataset records the population density values for 111 km by 111 km grids. To determine the population density of an incident's location, we find the grid whose center is the closest to the location and use its population density value.

3) SEDAC's GREENHOUSE GAS EMISSION DATA

The dataset contains the grid emissions of four greenhouse gases (CH4, CO, NOx, and NMVOC) for 1990. As it is of the same resolution (111 km by 111 km) as the population density data, we process it similarly and sum the emission values for 4 gases for each grid.

D. DEEP NEURAL NETWORK SETUP FOR INCIDENT CLASSIFICATION

1) MODEL STRUCTURE

We elaborate the model in Section III-B1 here. Similar to the TextCNN model, we add an attention mechanism for the output of each CNN. Figure 6 shows the architecture of our model. We denote the dimension of the word vectors by d. Suppose the length of a given sentence is s. Let $x_i \in \mathbb{R}$, which corresponds to the i-th word in the sentence, be a k-dimensional word vector. A piece of text of length n (zero-padding applied if necessary) is represented as:

$$x_{1:n} = x_1 \oplus x_2 \oplus \ldots \oplus x_n$$

where $\oplus$ is the concatenation operator. In general, let $x_{i:j}$ refer to the concatenation of words $x_i, x_{i+1}, \ldots x_j$. An h-gram input $x_{i:i+h-1}$ is transformed though a convolution filter $w_s \in \mathbb{R}^{h\times k}$:

$$c_{si} = f(w_s \cdot x_{i:i+h-1} + b_c)$$

where s is the s-th filter, $b_c \in \mathbb{R}$ is a bias term, and f is a non-linear ReLU function. An input x_i through a convolution

[7] https://github.com/one-step-ahead-result/one-step-ahead-code
[8] https://pytorch.org/
[9] https://scikit-learn.org/

filter $w_{a,i} \in \mathbb{R}^k$:

$$t_{ci} = f(w_{a,i} \cdot x_i + b_t)$$

where i is the i-th filter. This produces a *feature map* $C_s \in \mathbb{R}^{n-h+1}$, and a *feature map* $T \in \mathbb{R}^{A\times n}$, where n is the number of tokens in the sentence. We input c and a through a fully connected layer $\hat{w} \in \mathbb{R}^{(h\times A+1)}$ to combine their information and calculate an attention value for the convolution filter c_{si}:

$$l_{si} = f(\hat{w} \cdot (c_{si} \oplus t_{1i} \oplus t_{1(i+1)} \oplus \ldots \oplus t_{A(i+h-2)} \oplus t_{A(i+h-1)}) + b)$$

where A is the number of convolution filter. This produces a *features map* $L_s \in \mathbb{R}^{n-h+1}$. We then input L_s and C_s through neural network $\widetilde{w} \in \mathbb{R}^s$, we dot the attention value with its corresponding convolution layer's output:

$$\hat{c_s} = f(\widetilde{w} \cdot (L_s \cdot C_s) + b)$$

This processes a feature corresponding to this particular filter. These features are then concatenated:

$$g = \hat{c}_0 \oplus \hat{c}_1 \oplus \ldots \oplus \hat{c}_s$$

Finally, a fully connected layer is applied:

$$\hat{y} = f(w_y \cdot g + b_y)$$

2) EXPERIMENT CONFIGURATION

Here we detail the configuration for the incident classification experiment described in Section IV-B1. For the Wikipedia incident news titles (positive samples) and the NYT data (negative samples), we use a 9:1 train-test split. We also fix word vector dimension to 50, dropout rate to 0.5, and sentence length to 20. We compute the loss with binary cross entropy criterion and use Adam Optimizer on minibatches of size 1,000 to train the model.

E. FEATURE AGGREGATION

As mentioned in Section IV-C2, we aggregate the features at a daily level. We sum up the population density values as an estimate of total number of people affected on that day. Similarly, we sum up greenhouse gas emission values for the incident locations. Since the nightlight intensity has a range from 0 to 255, adding the values up across different incident locations would not have much physical meaning. For nightlight intensity, we choose the highest value across the incident locations. Administrative divisions for the locations, such as city and province are processed with one-hot encoding. Attributes including number of deaths, number of injuries, number of terrorist attacks, administrative divisions and political/cultural/religious labels are all accumulated on a daily basis with lagged attributes that reflect the past two days. Following a common practice to include past market information to predict its future movements, we construct the binary market direction features for the past two days using Equation (III.2) for targeting countries and Equation (III.1) for the US market. Because the markets do not trade on weekends, when we calculate a Monday's return using Equation (III.2), we subtract Friday's closing price from Tuesday's opening price. This return actually reflects the information during the three days. Thus, we combine these three days into one day as a proximate, to be consistent with the return on Monday.

F. COLLINEARITY ANALYSIS FOR FEATURE SELECTION

As mentioned in Section IV-C3, collinearity may exist among features. From the three socioeconomic features (nightlight intensity, greenhouse gas emissions and population density), we intend to discard the ones that do not bring much new information to improve feature vector quality. To do this, we use linear regressions to see whether one feature can be predicted by the other two for grids in Israel and for all grids (global). To align them to the same measurement scale (resolution), we downsample the nightlight intensity dataset to a map with 1 degree by 1 degree grids. The regression results suggest that we discard the greenhouse gas emission feature because it can be well explained by the other two features, with R^2 being 0.869 for Israel and 0.489 for global.

G. DATA SPLIT FOR MARKET MODELING

We describe how we split the data for market model training and validation, mentioned in Section IV-D. We train and validate the models for Israel on data from 10/23/1992 to 4/12/2010 (3,802 days in total, of which 415 days have terrorist attacks) and test the models on Israeli data from 4/13/2010 to 12/28/2017 (1,000 days in total, of which 162 days have terrorist attacks), with an approximate 4:1 ratio. The data for Colombia from 11/2004 to 12/2017 is similarly split while for the highly biased Spain data from 7/1994 to 12/2017 the train-test ratio is adjusted to about 1:1 such that there are enough days with terrorist attacks in the test set. As the data is sequential, we avoid using cross-validation. Instead, we further split the training set to get a temporary training set and a validation set (300 days for Israel) which is temporally after the new training set. After being used for tuning the hyperparameters with grid-searches, it is merged with the new training set to build models for testing. There are two reasons for doing this. One reason is to increase the size of training set and another reason is that intuitively, the validation set's temporal adjacency to the testing set may help achieve better performance as it carries the latest information.

ACKNOWLEDGMENT

The authors thank the anonymous reviewers for their constructive suggestions.*(Ziyue Li and Shiwei Lyu contributed equally to this work.)*

REFERENCES

[1] A. Abadie and J. Gardeazabal, "Terrorism and the world economy," *Eur. Econ. Rev.*, vol. 52, no. 1, pp. 1–27, Jan. 2008.

[2] E. Cavallo and I. Noy, "Natural disasters and the economy—A survey," *Int. Rev. Environ. Resource Econ.*, vol. 5, no. 1, pp. 63–102, 2011.

[3] K. P. Arin, D. Ciferri, and N. Spagnolo, "The price of terror: The effects of terrorism on stock market returns and volatility," *Econ. Lett.*, vol. 101, no. 3, pp. 164–167, Dec. 2008.

[4] F. Aslam and H.-G. Kang, "How different terrorist attacks affect stock markets," *Defence Peace Econ.*, vol. 26, no. 6, pp. 634–648, Nov. 2015.

[5] K. Drakos, "Terrorism activity, investor sentiment, and stock returns," *Rev. Financial Econ.*, vol. 19, no. 3, pp. 128–135, Aug. 2010.

[6] T. Sakaki, M. Okazaki, and E. al, "Earthquake shakes Twitter users: Real-time event detection by social sensors," in *Proc. 19th Int. Conf. World Wide Web (WWW)*, 2010, pp. 851–860.

[7] N. Levin, S. Ali, and D. Crandall, "Utilizing remote sensing and big data to quantify conflict intensity: The arab spring as a case study," *Appl. Geography*, vol. 94, pp. 1–17, May 2018.

[8] N. Levin and Y. Duke, "High spatial resolution night-time light images for demographic and socio-economic studies," *Remote Sens. Environ.*, vol. 119, pp. 1–10, Apr. 2012.

[9] J. Bollen, H. Mao, and X. Zeng, "Twitter mood predicts the stock market," *J. Comput. Sci.*, vol. 2, no. 1, pp. 1–8, Mar. 2011.

[10] X. Li, H. Xie, L. Chen, J. Wang, and X. Deng, "News impact on stock price return via sentiment analysis," *Knowl.-Based Syst.*, vol. 69, pp. 14–23, Oct. 2014.

[11] Y. Xu and S. B. Cohen, "Stock movement prediction from tweets and historical prices," in *Proc. 56th Annu. Meeting of Assoc. Comput. Linguistics (COLING)*, 2018, pp. 1970–1979.

[12] Z. Hu, W. Liu, J. Bian, X. Liu, and T. Y. Liu, "Listening to chaotic whispers: A deep learning framework for news-oriented stock trend prediction," in *Proc. 11th ACM Int. Conf. Web Search Data Mining (WSDM)*, 2018, pp. 261–269.

[13] K. Radinsky and E. Horvitz, "Mining the Web to predict future events," in *Proc. 6th ACM Int. Conf. Web search data mining - WSDM*, 2013, pp. 255–264.

[14] F. Petroni, N. Raman, T. Nugent, A. Nourbakhsh, Ž. Panić, S. Shah, and J. L. Leidner, "An extensible event extraction system with cross-media event resolution," in *Proc. 24th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Jul. 2018, pp. 626–635.

[15] Y. Chen, L. Xu, K. Liu, D. Zeng, and J. Zhao, "Event extraction via dynamic multi-pooling convolutional neural networks," in *Proc. 53rd Annu. Meeting Assoc. Comput. Linguistics 7th Int. Joint Conf. Natural Lang. Process.*, vol. 1, 2015, pp. 167–176.

[16] H. Tanev, J. Piskorski, and M. Atkinson, "Real-time news event extraction for global crisis monitoring," in *Proc. 53rd Annu. Meeting Assoc. Comput. Linguistics, 7th Int. Joint Conf. Natural Lang. Process.*, vol. 1, 2008, pp. 167–176.

[17] K. Leetaru and P. A. Schrodt, "GDELT: Global data on events, location, and tone, 1979–2012," in *ISA Annual Convention*, vol. 2, no. 4. 2013, pp. 1–49.

[18] T. Tavor and S. Teitler-Regev, "The impact of disasters and terrorism on the stock market," *Jàmbá J. Disaster Risk Stud.*, vol. 11, no. 1, pp. 1–8, Jan. 2019.

[19] C. Kollias, S. Papadamou, and A. Stagiannis, "Armed conflicts and capital markets: The case of the Israeli military offensive in the gaza strip," *Defence Peace Econ.*, vol. 21, no. 4, pp. 357–365, Aug. 2010.

[20] D. L. Pendell and C. Cho, "Stock market reactions to contagious animal disease outbreaks: An event study in Korean foot-and-mouth disease outbreaks," *Agribusiness*, vol. 29, no. 4, pp. 455–468, 2013.

[21] G. LaFree and L. Dugan, "Introducing the global terrorism database," *Terrorism Political Violence*, vol. 19, no. 2, pp. 181–204, Apr. 2007.

[22] A. H. Chen and T. F. Siems, "The effects of terrorism on global capital markets," in *The Economic Analysis of Terrorism*. Evanston, IL, USA: Routledge, 2007, pp. 99–122.

[23] T. Sandler and W. Enders, "Economic consequences of terrorism in developed and developing countries: An overview," in *Terrorism, Economic Development, and Political Openness*, vol. 17. Cambridge, U.K.: Cambridge Univ. Press, 2008.

[24] R. Eldor and R. Melnick, "Financial markets and terrorism," *Eur. J. Political Economy*, vol. 20, no. 2, pp. 367–386, Jun. 2004.

[25] C. Cao, O. Hansch, and X. Wang, "The information content of an open limit-order book," *J. Futures Markets*, vol. 29, no. 1, pp. 16–41, Jan. 2009.

[26] H. Lee, M. Surdeanu, B. MacCartney, and D. Jurafsky, "On the importance of text analysis for stock price prediction," in *Proc. LREC*, 2014, pp. 1170–1175.

[27] E. Gilbert and K. Karahalios, "Widespread worry and the stock market," in *Proc. Int. AAAI Conf. Web Social Media*, 2010, pp. 1–8.

[28] R. P. Schumaker and H. Chen, "Textual analysis of stock market prediction using breaking financial news: The AZFin text system," *ACM Trans. Inf. Syst.*, vol. 27, no. 2, p. 12, 2009.

[29] H. Wang, Y. Jiang, and W. Hui, "Stock return prediction based on Bagging-decision tree," in *Proc. IEEE Int. Conf. Grey Syst. Intell. Services (GSIS)*, Nov. 2009, pp. 1575–1580.

[30] L. Zhang, C. Aggarwal, and G.-J. Qi, "Stock price prediction via discovering multi-frequency trading patterns," in *Proc. 23rd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2017, pp. 2141–2149.

[31] X. Ding, Y. Zhang, T. Liu, and J. Duan, "Knowledge-driven event embedding for stock prediction," in *Proc. COLING*, 2016, pp. 2133–2142.

[32] A. Holzinger, C. Biemann, C. S. Pattichis, and D. B. Kell, "What do we need to build explainable AI systems for the medical domain?" 2017, *arXiv:1712.09923*. [Online]. Available: http://arxiv.org/abs/1712.09923

[33] W. Samek and K.-R. Müller, "Towards explainable artificial intelligence," in *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*. Cham, Switzerland: Springer, 2019, pp. 5–22.

[34] K. Shi, B. Yu, Y. Huang, Y. Hu, B. Yin, Z. Chen, L. Chen, and J. Wu, "Evaluating the ability of NPP-VIIRS nighttime light data to estimate the gross domestic product and the electric power consumption of China at multiple scales: A comparison with DMSP-OLS data," *Remote Sens.*, vol. 6, no. 2, pp. 1705–1724, Feb. 2014.

[35] Y. Kim, "Convolutional neural networks for sentence classification," 2014, *arXiv:1408.5882*. [Online]. Available: http://arxiv.org/abs/1408.5882

[36] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2014, pp. 1532–1543.

[37] C. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. Bethard, and D. McClosky, "The stanford CoreNLP natural language processing toolkit," in *Proc. 52nd Annu. Meeting Assoc. Comput. Linguistics: Syst. Demonstrations*, 2014, pp. 55–60.

[38] T. Terasvirta and H. M. Anderson, "Characterizing nonlinearities in business cycles using smooth transition autoregressive models," *J. Appl. Econometrics*, vol. 7, no. S1, pp. S119–S136, Dec. 1992.

[39] C. Floros, "Price linkages between the US, Japan and UK stock markets," *Financial Markets Portfolio Manage.*, vol. 19, no. 2, pp. 169–178, Aug. 2005.

[40] C. Kousky, "Informing climate adaptation: A review of the economic costs of natural disasters," *Energy Econ.*, vol. 46, pp. 576–592, Nov. 2014.

[41] R. Akita, A. Yoshihara, T. Matsubara, and K. Uehara, "Deep learning for stock prediction using numerical and textual information," in *Proc. IEEE/ACIS 15th Int. Conf. Comput. Inf. Sci. (ICIS)*, Jun. 2016, pp. 1–6.

[42] A. Grübler, B. O'Neill, K. Riahi, V. Chirkov, A. Goujon, P. Kolp, I. Prommer, S. Scherbov, and E. Slentoe, "Regional, national, and spatially explicit scenarios of demographic and economic change based on SRES," *Technol. Forecasting Social Change*, vol. 74, no. 7, pp. 980–1029, Sep. 2007.

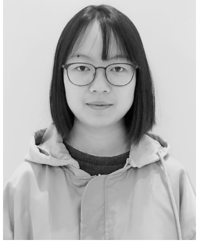
[43] J. Dams, S. Woldeamlak, and O. Batelaan, "Predicting land-use change and its impact on the groundwater system of the Kleine Nete catchment, Belgium," *Hydrol. Earth Syst. Sci.*, vol. 12, no. 6, pp. 1369–1385, 2008.

[44] K. Radinsky, S. Davidovich, and S. Markovitch, "Learning causality for news events prediction," in *Proc. 21st Int. Conf. World Wide Web (WWW)*, 2012, pp. 909–918.

[45] Y. Zhang and B. Wallace, "A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification," 2015, *arXiv:1510.03820*. [Online]. Available: http://arxiv.org/abs/1510.03820

[46] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*. [Online]. Available: http://arxiv.org/abs/1810.04805

[47] M. J. Flannery and A. A. Protopapadakis, "Macroeconomic factors do influence aggregate stock returns," *Rev. Financial Stud.*, vol. 15, no. 3, pp. 751–782, 2002.



ZIYUE LI received the B.S. degree in computer science and technology from ShanghaiTech University, Shanghai, China, in 2019, where she is currently pursuing the M.S. degree in computer science and technology with the School of Information Science and Technology, under the supervision of Prof. Haipeng Zhang. Her research interests include data mining, natural language processing, and recommender systems.

SHIWEI LYU received the bachelor's degree in software engineering from the School of Software, Northeastern University, China, in 2019. He is currently pursuing the master's degree with the Prof. Haipeng Zhang's Financial Intelligence Laboratory, School of Information Science and Technology, ShanghaiTech University, China. His research interests include data mining and deep learning.

HAIPENG ZHANG (Member, IEEE) received the B.E. degree in software engineering from Nanjing University, in 2009, and the Ph.D. degree in computer science from Indiana University under the supervision of Prof. David J. Crandall, in 2014. He was an exchange student at HKUST, in 2007. From 2010 to 2013, he did research internships with the National Institute of Informatics, Tokyo, eBay Research Labs, San Jose, CA, USA, Microsoft Research, Cambridge, U.K., and Samsung Research North America, San Jose. From 2014 to 2018, he worked with IBM Research and China Financial Futures Exchange on Data Science and Fintech. He joined ShanghaiTech, as a Tenure-Track Assistant Professor, PI, in August 2018. He leads the Financial Intelligence Laboratory and has established Fintech research collaborations with core financial institutes in China including China Foreign Exchange Trade System, Shenzhen Stock Exchange, and UnionPay. He publishes in venues, including WWW and WSDM. His work received media coverage from New Scientist magazine and the Communications of the ACM website.

TIANPEI JIANG received the B.S. degree in mathematics from the South China University of Technology, in 2013, and the M.S. degree in financial engineering and the Ph.D. degree in statistics from the University of Western Ontario, in 2014 and 2017, respectively. From 2017 to 2018, he worked as a Data Scientist with Bank of Communications. He joined the Prof. Haipeng Zhang's Financial Intelligence Laboratory, ShanghaiTech, as an Assistant Researcher, in 2019.