Detecting, Quantifying and Accessing impact of News events on Indian Stock Indices

Ishan Verma
TCS Innovation Labs
Tata Consultancy Services Ltd.
New Delhi, India
ishan.verma@tcs.com

Lipika Dey
TCS Innovation Labs
Tata Consultancy Services Ltd.
New Delhi, India
lipika.dey@tcs.com

Hardik Meisheri
TCS Innovation Labs
Tata Consultancy Services Ltd.
New Delhi, India
hardik.meisheri@tcs.com

ABSTRACT

The impact of different types of events reported in News articles on stock market is a widely accepted phenomenon. Market analysts rely heavily on technology to combine data from different sources and generate appropriate insights for predicting stock movements. With plethora of sources reporting news on plentitude of events happening across the world, a combination of text mining techniques and predictive technologies can play a significant role in this arena. In this paper we have presented methodologies to identify and quantify the presence of different types of information that can affect the market from a multitude of web sources, and finally use the information for predicting stock movement direction. We propose the use of PESTEL factors to categorize market-impacting information. We have analyzed large volumes of past available data using Granger causality to understand how these categories impact the market. We propose a paragraph-vector based information classification mechanism. We also present Long-Short term memory Network (LSTM) based prediction model to investigate the prediction capabilities of the information components. The proposed system outperforms state of the art linear SVM on data from different stock indices.

KEYWORDS

News Event Detection, Granger Causality, Stock Prediction

ACM Reference format:
Ishan Verma, Lipika Dey, and Hardik Meisheri. 2017. Detecting, Quantifying and Accessing impact of News events on Indian Stock Indices. In *Proceedings of WI '17, Leipzig, Germany, August 23-26, 2017,* 8 pages.
DOI: 10.1145/3106426.3106482

1 INTRODUCTION

There has been a fairly large body of studies by economists to unravel the associations between world events and stock returns. A review of related literature reveals that most of the earlier works have concentrated on working with financial News articles - reporting events related to the stock market and business organizations. Though economists agree that stock markets can show abnormal behavior as a reaction to socio-political events, acts of terrorism or natural disasters etc. also, there has been no formal effort to track such events in an automated fashion and use them for prediction. There have been isolated efforts in analyzing twitter sentiments or multi-nation activities involving a popular sport etc. -but to the best of our knowledge, no consolidated effort exists in studying the effect of a broad category of events reported in news over stock market movements. This may be due to the fact that it is not possible to build a pre-computed repository of all possible relevant events that are likely to impact stock markets in the near future. Such events can only be learnt in hindsight.

Acknowledging the fact that a priori identification of all events that can possibly affect stock movements is not feasible, in this paper we propose a different approach to track a broader category of events encompassing various categories and analyze their effects on stock indices. Like many of our predecessors, we also analyze stock-market performance and News articles jointly to identify strong and weak causal relationships among different kinds of world events and stock price movements. However, rather than extracting specific events, we assign the News articles reporting these events to six labeled categories, termed together as PESTEL - Political, Economic, Social, Technological, Environmental and Legal.

The PESTEL framework [5, 22] had been proposed for analyzing general environments of a business organization. It highlights six critical factors that can help managers to identify potential opportunities and threats for the organization. The PESTEL framework focuses on analyzing the effects of broader macro-environmental trends on an organization, business or an industry segment as a whole and thereby assess sustainability and profitability. Though PESTEL had not been proposed to study the effects of macro-economic events on stock market, we propose to use this framework for analyzing and predicting stock indices of a region as a whole, since these indices can be thought of as cumulative indicators of business environments for different sectors of a geographical region.

The significance of each factor is described as follows:

- Politics plays a critical role in business since it defines and regulates systems of control. The aggregated political News content of a period goes up whenever events of political significance occur.
- Economic factor in News repositories go up whenever there are announcements related to fiscal policies or high-value business transactions like mergers and acquisitions, partnerships or investment declarations etc. These can

WI '17, Leipzig, Germany

DOI: 10.1145/3106426.3106482

determine, measure and assess the economic health of a given region.
- Social factors or demographic factors takes into consideration the mental states of individuals or consumers within a business environment. Consumer sentiments and thoughts are central and critical to determining market conditions.
- Technology is a key factor that drives business today. An organization's use or projected use of technology plays an important role in building its brand image in the consumer's minds.
- Environment is another important factor in today's global business environments. Companies that flout environmental policies are penalized by both consumers and governments. Environment-related News may have both direct and indirect impacts on individual stock indices either as a reaction or in anticipation. Environment related policy announcements as well as information about their violations or possible violations can affect the market in both positive and negative ways.
- Legal events have always had business impact on the involved companies in different ways depending on which side of the law a company is in. These effects percolate through and impact the overall market indices.

It is well accepted today that events of significance in any category spreads over the web very fast and wide. Automated systems can not only capture such information but also gauge the significance of a piece of information based on the volume and velocity at which it spreads. The distinguishing features of this work are highlighted below:

(i) The proposed system makes use of News articles captured from the Web to determine the aggregated presence of each category of information within a time-window in near real-time. We have used the Distributed Memory model of Paragraph vectors, described in [1] to quantify the presence of different categories of information in a News article using a multi-labeled classifier.
(ii) We propose the use of Granger causality measure [13] to determine causal relations among the macro-environmental factors defined earlier and stock index movements.
(iii) We propose the use of machine learning techniques like SVMs and LSTMs to predict stock movements based on the macro-environmental events reported in News. The problem is formulated as a binary classification problem that predicts whether a stock index will show positive or negative movements based on available news information.
(iv) We have conducted rigorous experiments with data from the Indian stock market and related News spanning over four years. Our experiments reveal several interesting insights related to the diverse causal factors that can affect different sectors. For example, we observed that the Energy sector is heavily affected by Political events while the Banking sector is more affected by Technology events. LSTMs are found to out-perform state of the art linear SVMs in predicting stock movements.

These studies are very significant in today's knowledge-driven economy, where investment decisions are heavily dependent on the ability to collect and analyze the right information at the right time. Investment decisions today depend not only on traditional forecast models that use historical stock values, but also on the unstructured text data that is out there on the Web in the form of News articles, social media content, expert viewpoints and so on.

2 REVIEW OF RELATED WORK

There has been a fairly large body of studies by economists to unravel the associations between stock returns and world events. They have been studying the effects of global and regional economic, political and social events reported in News on stock prices. In [4], Cutler et al. attempted to estimate the fraction of the variation in stock market that can be attributed to economic news. They show that about a third of all variations in stock returns can be attributed to various types of economic news. Studies shown in [8, 9, 12, 14] have examined the effects of scheduled macroeconomic announcements releasing information about gross national product, inflation rate, unemployment rates etc. on trading and volatility indices. Implied volatilities estimated from stock index options prices are used to investigate stock market uncertainty. Implied volatility can be interpreted to be a market's expectation of the average stock's return volatility over the remaining life of the option contract, as Merton (1973) shows. In [19], Nikkinen and Sahlstrom investigated how the macroeconomic news from USA, the world's largest economy has an effect on the uncertainty in a foreign stock exchange. This has important implications for multinational investors since the uncertainty directly affects stock and options valuation.

In [16], Kim and Mei proposed to employ a components-jump volatility filter to investigate the possible market impact of political risk. The filter identifies jump return dates and associates them with political events, thereby producing a measure of the effects of political announcements on the market return and volatility effects. This paper shows that political developments in Hong Kong have a significant impact on its market volatility and return. It was indicated that the results have some interesting implications for option pricing and political risk management.

Studies have also shown that stock market returns and investor sentiments are causally related to each other. While investor sentiments themselves may be dictated by occurrences of certain global events, certain politically or socially significant events are known to affect the stock market. The range of such causal events may be very large. In 2007, Edmans, Garcia, and Norli [10] showed that a strong association exists between results of soccer games and local stock returns. They investigated 39 stock markets and found the existence of an asymmetric effect, where losses have a significant negative effect in the losing countries local markets, whereas victories do not have significant effects. Kaplansky and Levi [15] take this further to predict the aggregated effects of soccer sentiments from numerous local markets on the U.S. market. Based on the fact that relatively large proportion of local investors of many countries also invest internationally, and mainly in U.S., they show that during the World Cup period there is a global negative effect induced by all losing countries fans at an international level.

In more recent times, a number of studies have analyzed texts from social network services (SNS), blogs and news to analyze correlations between stock prices and public emotion as reaction to

social events and news [2, 3, 20, 21, 23]. Most of these works analyze only articles in specific categories like financial sections. Public emotions are categorized as positive or negative. In [25] Wong and Ko a novel approach was presented to determine public emotions as a numeric score by applying emotion analysis to daily news articles. This work considered all types of News articles. It utilized an emotion analysis lexicon generated from crowd-annotated news for the purpose.

In [18] Luss and Aspremont show that information extracted from news articles can be used to predict intraday price movements of financial assets using support vector machines. Multiple kernel learning is used to combine equity returns with text as predictive features to increase classification performance. This paper also addresses the kernel-learning problem efficiently. The authors state that while the direction of returns is not predictable using either text or returns, text features can predict the size of returns better than historical returns alone.

In [6], Ding et al. utilized Open Information Extraction (Open IE) techniques to extract structured events from web-scale data. In the same paper, they proposed the use of both linear and non-linear models to investigate the hidden and complex relationships between events and the stock market. It is shown that the accuracy of S&P 500 index prediction is 60%, and that of individual stock prediction can reach around 70%. The same authors proposed deep learning based methods in [7] to analyze event-driven stock market prediction. Events extracted from news text are represented as dense vectors that are trained using a novel neural tensor network. A deep convolutional neural network is then used to model short-term and long-term influences of events on stock price movements. This work also uses market simulation to analyze the system's capability to make profits. In both the above work, events are extracted from news titles only to avoid noise.

Our work stands apart from the rest since most of these works consider financial News articles. None of them consider all kinds of News articles, which may contain information about events that are non-financial in nature. Events in these categories may contain political, social, technological or other significant information that may influence investor's decisions.

3 NEWS INFORMATION PROCESSING

Figure 1 shows the information processing pipeline followed for acquisition and processing of documents which follows the framework presented in [17]. News articles are collected using RSS feeds provided by a number of agencies. All gathered content is uniquely identified, time-stamped and stored. Stock indices data is captured from National Stock Exchange India website (https://www.nseindia.com/). Details of rest of the modules are presented in the coming sections.

3.1 News Classification into PESTEL categories

The next step in information processing pipeline is to classify the text documents into PESTEL categories. The PESTEL labels have already been defined in Section 1. The objective of this module is to identify and quantify information components that can qualify as PESTEL category elements. A single News article almost always contains information components belonging to multiple categories.

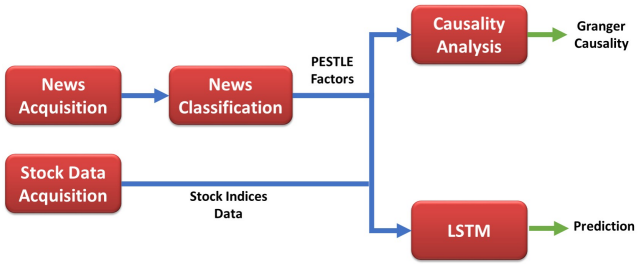

Figure 1: Information Processing Architecture

For example, an article reporting a political, economy or a technology event often refers to the social implications of the event. Hence we propose a multi-label classification mechanism that generates a quantified vector as output. The output vector signifies the relative proportion of each type of information in the article.

The problem can thus be formulated as follows: Given a set of classification labels the challenge is to identify the presence of those labels within a news document and assign a weightage to it. This is modeled as a multi-class classification problem where given a set of news documents and a set of classes, the task is to assign weightage to each class based on its presence in the news document.

The training set has been prepared using a mix of human-labeling and web-site labeled data. Several News websites provide News articles under specific labels of "Technology", "Environment" and "Economic". Archived articles obtained from these channels have been used as labeled samples for these categories. Political, social and legal articles have been collected from multiple global and regional news websites and manually labeled. Each annotated news article may be labeled with multiple labels by different annotators. These are retained in the training set as different instances.

We have used the Distributed Memory model of Paragraph vectors, described in [5], to represent the texts. Paragraph Vector is an unsupervised framework that learns continuous distributed vector representations of both paragraphs and words, using a stochastic gradient descent and gradients are learnt using back propagation.

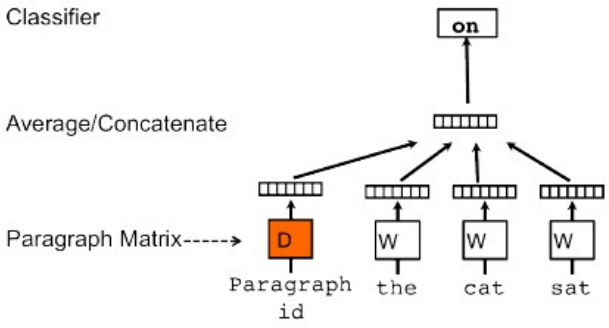

Figure 2: Paragraph vector learning framework [22]

The goal is to represent entire text document using a single vector, which can then be used as input to a supervised machine-learning algorithm to associate documents with labels.

The algorithm itself has two key stages:

(i) In the training stage word vectors W, softmax weights U, b and paragraph vectors D for seen paragraphs are learnt
(ii) During the inference stage - paragraph vectors D for new paragraphs are obtained keeping W, U, b fixed. D is used to make a prediction about some particular labels using a standard classifier. We have used logistic regression.

Output of the paragraph vector classifier is a six dimensional vector where each dimension corresponds to one of the PESTEL factors. The individual values associated to each factor in the vector is considered as the weight of presence of the factor in a document. The PESTEL factor categorization performance are shown in Results section.

Figure 3 shows a news article talking about Indian currency demonetization dated 8th Nov. 2016. Top-right corner shows the value of different PESTEL event obtained using the paragraph vector classifier. It can be observed that the presence of Economy. Political and Social event was higher than others in this news article. Presence of content related to stock futures, government policy change introduced by Indian Prime Minister Narendra Modi and impact on Indians is justifying the higher value for Economy. Political and Social events.

Figure 3: Sample News article with PESTEL classification values

The total presence of a PESTEL factor across different time-stamped news articles can be aggregated at various levels like day, week, and month and so on. These aggregated values can be further utilized to obtain time-series of PESTEL factors. The next section discusses how these time-series are utilized to analyze the impact of different factors on stock indices

3.2 Assessing Impact of PESTEL factors on Stock Movement

To validate whether the PESTEL factors identified in News articles impact stock indices or not, we chose to use Granger causality test. The Granger causality test is a statistical hypothesis test for determining whether a time-series X is useful in forecasting another time-series Y. Time-series X is accepted as impacting Y, if and only if, prediction of future values of Y improves after taking values of X into consideration.

In this case our null hypothesis is that 'PESTEL factors reported in News do not Granger-cause changes in indices. We apply Granger causality test using PESTEL event occurrence time-series and stock indices time-series.

$$y_t = \alpha + \sum_{i=1}^{n} \beta_i y_{t-i} + \epsilon_t$$
$$y_t = \alpha + \sum_{i=1}^{n} \beta_i y_{t-i} + \sum_{i=1}^{n} \gamma_i x_{t-i}$$

The first one tries to predict future values of Y from past values. The second model takes past values of both X and Y to predict future values of Y, where series X represents quantified PESTEL factors computed as mentioned in earlier section.

Past studies have established that the effect of reported events on stock indices are usually observed within a short span of time. However, studies do not agree on a unique value for the time period over within which the effect may be observed. While effect of some reported event is observed to be within hours, for others it could be days. Accordingly, we have experimented with different sets of delay factors ranging over 1, 2, and 5 days.

For Granger causality test, the p-value of the result of hypothesis test is considered. The critical value for rejecting null hypothesis is set to 0.95. This means that if p-value is observed to be greater than 0.95 then it can be safely assumed that X Granger-causes Y.

We conducted the Granger Causality test between Indian stock market indices and PESTEL factors by using a collection of 250000 News articles, published from January 2013 to December 2016 collected from trusted web News sources. Figure 4 shows that the comparative presence of each factor is more or less similar across years.

Figure 5 shows the results of Granger causality test which assessed the impact of the extracted PESTEL factors on India's volatility index. Volatility Index is a measure of market's expectation of volatility over the near term. Volatility is described as the "rate and magnitude of changes in prices". Volatility Index is a measure, of the amount by which an underlying Index is expected to fluctuate, in the near term. Predicting volatility is a challenging problem since volatility depends on a number of external factors and not on its past values alone.

Figure 5 shows that presence of legal events in News articles immediately impacts the stock market. This has been established by several earlier studies also, both by economists and data analysts. Our Granger causality experiments simply validate this. However, Figure 5 also reveals that political events have an even higher impact albeit at a delay of a day. This is an interesting aspect. We investigated this phenomenon and discovered that while political incidents get reported in News almost as soon as they occur, analysis and discussion around significant political events start after that and then percolates among masses and thereafter starts impacting the stock market. Hence the effects of these events are observed at a delay. Figure 5 also shows that effects of events containing high socio-political factor affects volatility indices for a longer period, though its effect decays gradually.

In section IV, we shall present some more interesting results of Granger causality tests for more fine-grained analysis of different sectors of the market. It is obvious that all sectors are not likely to be impacted by an event in identical ways. Our experiments

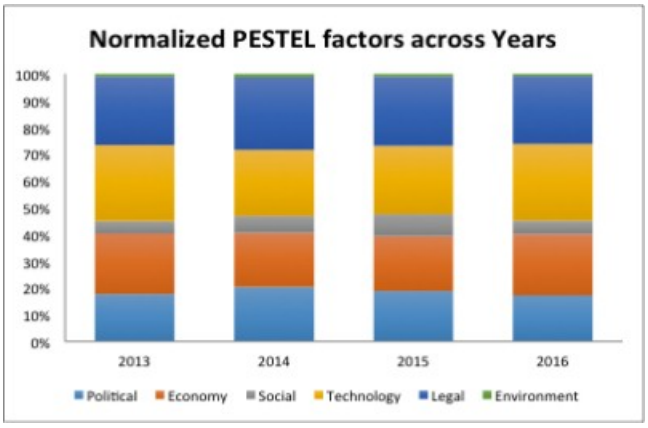

Figure 4: Distribution of PESTEL factors across years

corroborate this view. We have been able to discover some interesting causal relationships among the market indices for different industry sectors and PESTEL factors.

PESTEL Factors	India_Volatility		
	1 Day	2 Day	Week
Legal	0.97	0.83	0.58
Technology	0.59	0.36	0.24
Environment	0.53	0.42	0.28
Political	0.71	0.98	0.86
Social	0.29	0.72	0.86
Economy	0.29	0.16	0.14

Figure 5: Impact of PESTEL factors on India Volatility Index

3.3 Stock Movement Prediction using LSTM

Predicting whether stocks will fall or rise based on events or issues that have been reported in News in the recent past can be formulated as a binary-valued prediction problem. The aim is to predict whether there will be a movement in the stock indices or it will remain static. Again, the prediction can be done in multiple ways by varying the time-periods. Our aim is to see how well and how much in advance can the movement be predicted assuming that knowledge about recent PESTEL factors reported in News articles are available.

Long Short Term Memory networks, referred to as "LSTMs" are a special class of Recurrent Neural Networks (RNN). RNNs are capable of learning long-term dependencies and have proven to be effective in prediction task with respect to temporal domain. Recurrent architectures represent time in a recursive manner. Usually the hidden layer state is conditioned on a previous state [11]. This helps in modeling complex signals that are spread over longer period of time as the hidden state can also act as a memory cell. However RNNs suffer from vanishing gradient problem. The vanishing error problem refers to how the influence of past inputs decays quickly over time.

LSTMs handle the vanishing gradient problems by using a memory cell to capture long-term dependencies. These type of architectures are good for modeling time-series prediction problems where the input events change dynamically. Since LSTMs can capture sequence dependence among input variables very well and are explicitly designed to avoid the long-term dependency problem of simple RNNs. Remembering information for long periods of time is practically their default behavior, not something they struggle to learn!

These kind of dependencies are very essential to be captured for stock market prediction as they are dynamic and do not follow any pattern. We have used LSTMs to predict the fall/rise of stock value based on previous values of events.

3.3.1 Proposed LSTM Architecture.

Network Architecture

Figure 5 shows the network architecture that is used for prediction of stock index movement. There are three layers namely, Input, LSTM and output layer. Input layer has same dimension as that of input data. The number of input nodes depends on the number of days for which events are considered. Since we have experimented with time-periods of 1, 2 and five days, consequently the number of event input nodes are 6, 12 or 30 for the different situations.

Output of the network is expected to be 1 or 0, where 1 signifies that the stock index has risen and 0 signifies that the index has fallen. While the network is fed with real values indicating occurrence of different categories of PESTEL factors till t^{th} day the output of the network predicts what kind of movement is expected on $(t+1)^{th}$ day.

The input layer forwards the input values to LSTM layer without any change. The LSTM layer consist of 4 hidden states or cells. h_t^p is output of p^{th} cell at t^{th} time instant. This output is then fed to output layer which contains 2 neurons.

Figure 6 shows generalized cell of LSTM. <> represents vector, p represents LSTM, here $p \in [1, 2, 3, 4]$. There are four 3 gates namely, forget gate, output gate and input gate. In addition, there is memory cell which is present. Suffix *i, o, f, c* represents input gate, output gate, forget gate and memory cell respectively. t represents the input that tth time instance. LSTM is governed by the following equations:

$$
\begin{aligned}
i_t^p &= \sigma(w_{xi}^p x_t + w_{hi}^p h_{t-1}^p + w_{ci}^p c_{t-1}^p + b_i^p) \\
f_t^p &= \sigma(w_{xf}^p x_t + w_{hf}^p h_{t-1}^p + w_{cf}^p c_{t-1}^p + b_f^p) \\
c_t^p &= f_t^p c_{t-1}^p + i_t^p tanh(w_{xc}^p x_t + w_{hc}^p h_{t-1}^p + b_c^p) \\
o_t^p &= \sigma(w_{xo}^p x_t + w_{ho}^p h_{t-1}^p + w_{co}^p c_{t-1}^p + b_o^p) \\
h_t^p &= o_t^p tanh(c_t^p)
\end{aligned}
$$

where, w_{ij}^p refers to the weights from i to j for p^{th} cell. For example w_{xi}^p refers to weights from input xt to input gate. Tanh is 'tan hyperbolic' function and Sigm is sigmoid function. c_t^p represents memory element for t^{th} time instance for p^{th} cell. Output from each cell is propagated along with memory information. This helps in capturing long and short term dependencies.

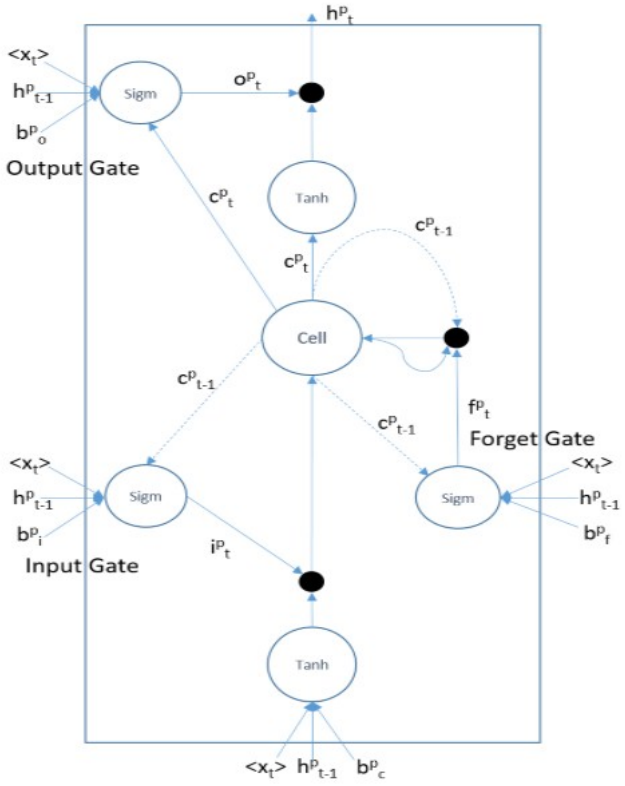

Figure 6: Generalized cell of LSTM

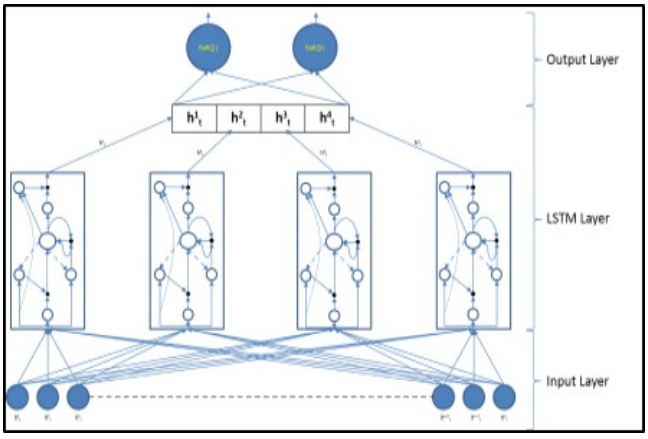

Figure 7: Network Architecture

Training in LSTM is done through truncated Back Propagation through time (BPTT) [24]. It is extension of back propagation used in the neural networks along the temporal axis. All the inputs from $t = 1 : T$, are used in forward pass and is stored with the output. Then the back propagation starts from $t = T$ all the way up to $t = 1$. Weight updation is done using stochastic gradient decent. Stochastic gradient decent is optimized using Adam optimizer which is explained below,

Adaptive Moment Estimation (Adam) calculates adaptive learning rates dynamically for each parameter. It also stores exponentially decaying average of past squared gradients (v_t) and exponentially decaying average of past gradients (m_t) as follows,

$$m_t = \beta_1 m_{t-1} + (1 - \beta_1) g_t$$
$$v_t = \beta_2 v_{t-1} + (1 - \beta_2) g_{2t}$$

m_t and v_t are estimates of the first moment (the mean) and the second moment (the uncentered variance) of the gradients respectively. The default values of 0.9 for $\beta 1$ and 0.999 for $\beta 2$.

Activation function for output layer is Softmax which is standard for binary classification. In addition to this for tuning the learning rate of the network, Adam Optimizer is used. We have used Binary Cross entropy as we are dealing with two class problem. The loss function is defined as follows,

Computes the binary cross-entropy between predictions and targets.

$$L = -tlog(p) - (1 - t)log(1 - p)$$

where t is target value and p is predicted value.

4 EXPERIMENTS AND EVALUATIONS

4.1 Data Description

A document collection has been built from news documents collected through RSS feeds provided by major news sources for India geography. The collection contains over 258144 documents collected for the duration of Jan-2013 to Feb-2017.

The overall data set is split temporally in training and test set for prediction purposes. Data from 01/01/2013 to 31/08/2016 are used for training the LSTM, while the data from 01/09/2016 to 31/01/2017 have been used for testing. Altogether, the data contained 912 instances of daily indices data, 820 of which is used for training and 92 for testing.

4.2 Indexes considered for evaluation

The S&P CNX Nifty (also known as NSE Nifty or 'Nifty') is a stock index in India. Nifty consists of 50 stocks representing 23 industry sectors. The constituents of the index change periodically, depending on liquidity, availability of floating stock, turnover and volume of transactions. Apart from India Volatility index which was described in section II, data from the following indexes are considered for evaluation

NIFTY 50- The NIFTY 50 is a diversified 50 stock index accounting for 13 sectors of the economy. It is used for a variety of purposes such as benchmarking fund portfolios, index based derivatives and index funds. The NIFTY 50 Index represents about 65% of the free float market capitalization of the stocks listed on NSE as on March 31, 2016.

NIFTY Bank - An index comprised of the most liquid and large capitalized Indian Banking stocks. It provides investors and market intermediaries with a benchmark that captures the capital market performance of Indian Banks. The index has 12 stocks from the banking sector which trade on the National Stock Exchange.

NIFTY Auto- The NIFTY Auto Index is designed to reflect the behavior and performance of the Automobiles sector which includes manufacturers of cars & motorcycles, heavy vehicles, auto ancillaries, tires, etc. The NIFTY Auto Index comprises of 15 stocks that

are listed on the National Stock Exchange.

NIFTY IT- NIFTY IT provides investors and market intermediaries with an appropriate benchmark that captures the performance of the IT segment of the market. Companies in this index are those that have more than 50% of their turnover from IT related activities.

NIFTY Energy - Energy sector is universally recognized as one of the most significant inputs for economic growth. NIFTY Energy index was developed to capture the performance of the companies in energy sector. NIFTY Energy Index include companies belonging to Petroleum, Gas and Power sub sectors.

4.3 Evaluations

Our experiments are carried out on three different time intervals: 1 day, 2 day and 1 week. We test the influence of events on predicting the polarity of stock indices change for each time interval. Our feature set include quantified event features and previous days stock indices values. We compare LSTM based prediction with state of the art linear SVM classifiers.

4.3.1 Prediction Measures.

We have used two assessment metrics to evaluate performance of the prediction algorithm. First, we chose accuracy measure since it is a standard and intuitive approach to measure the performance of classifiers. However, accuracy is known to be sensitive to data skew: when a class has a high frequency, the accuracy can be high using a classifier that makes prediction on the majority class. Stock value predictions typically use Matthews correlation coefficient (MCC) as an evaluation metric. MCC is used as a measure of the quality of binary classifications. It takes into account true and false positives and negatives and is generally regarded as a balanced measure which can be used even if the classes are of very different sizes. The MCC returns a value between âĹŠ1 and +1. A coefficient of +1 represents a perfect prediction, 0 no better than random prediction and âĹŠ1 indicates total disagreement between prediction and observation. Given TP, TN, FP and FN (True Positive, False Positive, True Negative and False Negative, respectively), it is computed as follows:

$$MCC = \frac{(TP*TN)(FP*FN)}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$$

4.4 Results

4.4.1 PESTEL factor Classification Result.

Here we present evaluation of the classification algorithm used for PESTEL events. News data for experimentation purposes is collected through RSS feeds from a chosen set of websites. These websites publish news about India geography. Our training dataset contained around 1200 labeled document equally distributed for each PESTEL class. Using paragraph vector classifier we achieved an accuracy of 91.94% for 10-fold cross validation over 70:30 split. Table 1 shows confusion matrix for the classification.

Table 1: Confusion Matrix for PESTEL Classifier

	Pol.	Eco.	Soc.	Tech.	Env.	Leg.
Political	187	0	0	0	0	0
Economy	0	174	0	0	0	0
Social	0	0	184	0	0	0
Technology	0	0	0	190	0	0
Environment	0	0	0	0	174	0
Legal	0	0	0	0	0	197

Index	No of Days	Political	Economy	Social	Technology	Environment	Legal
NIFTY 50	1 Day	0.52	0.02	0.54	0.39	0.4	0.6
	2 Day	0.31	0.04	0.84	0.34	0.04	0.38
	Week	0.47	0.16	0.94	0.67	0.06	0.53
NIFTY Auto	1 Day	0.35	0.02	0.92	0.32	0.55	0.47
	2 Day	0.29	0.04	0.85	0.36	0.26	0.4
	Week	0.43	0.17	0.8	0.66	0.34	0.49
NIFTY IT	1 Day	0.23	0.02	0.2	0.08	0.09	0.11
	2 Day	0.15	0.01	0.1	0.04	0.03	0.04
	Week	0.22	0.02	0.09	0.12	0.01	0.09
NIFTY Bank	1 Day	0.62	0.05	0.29	0.56	0.52	0.73
	2 Day	0.35	0.07	0.54	0.57	0.1	0.49
	Week	0.54	0.29	0.74	0.99	0.17	0.71
NIFTY Energy	1 Day	0.76	0.17	0.1	0.93	0.95	0.45
	2 Day	0.98	0.44	0.11	0.98	0.44	0.69
	Week	0.77	0.89	0.13	0.49	0.49	0.52

Figure 8: Granger Causality Results for different indices

4.4.2 Granger causality result.

Figure 8 shows Granger causality result across different stock indices and PESTEL events for the period 2013-16. The cells of the graph that are highlighted in yellow indicate the factors that are found to impact the sector significantly. The significant findings of Figure 8 are summarized as follows:

1. NIFTY_50 the global index is impacted the most by social events. The effects are long-term.
2. NIFTY Auto index is also impacted by social events.
3. NIFTY Bank index is highly affected by Technology related News.
4. NIFTY Energy sector is clearly the most susceptible one. It is affected by social, technological and Environment related reports.
5. NIFTY IT index did not show much causal correlation with any of the PESTEL factors. Talking to experts revealed that this sector in India is more dependent on People News rather than anything else.

These results clearly show that it is important to take a wide variety of events into consideration for predicting stock indices.

4.4.3 Prediction Results.

Table 2 shows prediction results for LSTM based prediction model. It also lists accuracy and MCC obtained with linear SVM model on the same dataset. It can be observed that LSTM outperforms SVM's in most of the cases under both scenarios. It is also interesting to observe that while for SVM the accuracy either remains same or goes down with increase in data size in the form of number of days, LSTM accuracy improves. The only sector which shows no

improvement is the IT sector, which as indicated earlier has no Granger-causality with PESTEL events. Hence this result is also valid.

The daily and weekly accuracy achieved using LSTM based algorithm is comparable with prediction accuracy reported by [6] Ding et. al. in their work on event based stock prediction of Standard & Poor 500 stock index. While their result is obtained using US News dataset, our collection is focused on News from Indian Geography with its impact on performance of stock indices pertaining to different Indian industrial sectors.

Table 2: Stock Index Prediction with PESTEL Events and previous Index Values

Index	No. of Days	SVM		LSTM	
		Accuracy	MCC	Accuracy	MCC
India Volatility Index	1	61.96	0.169	61.96	0.154
	2	56.52	0.071	**67.39**	**0.304**
	5	58.24	0.099	**64.83**	**0.250**
Nifty 50	1	53.26	0.014	**58.69**	**0.163**
	2	51.09	0.043	**52.17**	**0.098**
	5	52.75	0.035	**54.95**	**0.113**
Nifty Bank	1	**57.61**	**0.267**	51.08	0.019
	2	51.09	0.01	**55.43**	**0.114**
	5	54.95	0.049	**56.05**	**0.070**
Nifty Auto	1	**53.26**	0.063	50	**0.103**
	2	**52.17**	0.051	51.09	**0.146**
	5	**58.24**	**0.165**	51.65	0.105
Nifty IT	1	55.43	0.097	**56.52**	**0.118**
	2	51.09	0.036	**54.35**	**0.071**
	5	52.75	0.032	**56.04**	**0.122**
Nifty Energy	1	51.09	**0.188**	**58.69**	**0.188**
	2	54.35	**0.093**	**57.61**	0.079
	5	**59.35**	**0.133**	58.24	0.083

5 CONCLUSION

In this paper, we have presented a framework for extracting generic PESTEL factors from news articles, quantifying their presence and using it to predict the movement of different stock indices. Later we have shown that the quantified information can be used to predict stock movement with better accuracy than SVM's.

We intend to extend this work in future to cover a wider range of factors along with a more fine-grained representation of generic events. Events related to people, natural disasters etc. have high impact on stocks. In the present framework these are all clubbed together under one factor. We also intend to work on predicting stock movements for individual companies which is an even more difficult task since it depends on both generic world events as well as events very specific to the company.

REFERENCES

[1] Johan Bollen, Huina Mao, and Xiaojun Zeng. 2011. Twitter mood predicts the stock market. *Journal of computational science* 2, 1 (2011), 1–8.

[2] Johan Bollen, Huina Mao, and Xiaojun Zeng. 2011. Twitter mood predicts the stock market. *Journal of computational science* 2, 1 (2011), 1–8.

[3] Wesley S Chan. 2003. Stock price reaction to news and no-news: drift and reversal after headlines. *Journal of Financial Economics* 70, 2 (2003), 223–260.

[4] David M Cutler, James M Poterba, and Lawrence H Summers. 1989. What moves stock prices? *The Journal of Portfolio Management* 15, 3 (1989), 4–12.

[5] Crispin Dale. 2000. The UK tour-operating industry: A competitive analysis. *Journal of Vacation Marketing* 6, 4 (2000), 357–367.

[6] Xiao Ding, Yue Zhang, Ting Liu, and Junwen Duan. 2014. Using Structured Events to Predict Stock Price Movement: An Empirical Investigation.. In *EMNLP*. 1415–1425.

[7] Xiao Ding, Yue Zhang, Ting Liu, and Junwen Duan. 2015. Deep Learning for Event-Driven Stock Prediction.. In *Ijcai*. 2327–2333.

[8] Louis H Ederington and Jae Ha Lee. 1993. How markets process information: News releases and volatility. *The Journal of Finance* 48, 4 (1993), 1161–1191.

[9] Louis H Ederington and Jae Ha Lee. 1995. The short-run dynamics of the price adjustment to new information. *Journal of Financial and Quantitative Analysis* 30, 1 (1995), 117–134.

[10] Alex Edmans, Diego Garcia, and Øyvind Norli. 2007. Sports sentiment and stock returns. *The Journal of Finance* 62, 4 (2007), 1967–1998.

[11] Jeffrey L Elman. 1990. Finding structure in time. *Cognitive science* 14, 2 (1990), 179–211.

[12] Michael J Fleming and Eli M Remolona. 1999. Price formation and liquidity in the US Treasury market: The response to public information. *The journal of Finance* 54, 5 (1999), 1901–1915.

[13] Clive WJ Granger. 1969. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: Journal of the Econometric Society* (1969), 424–438.

[14] Campbell R Harvey and Roger D Huang. 1991. Volatility in the foreign currency futures market. *The Review of Financial Studies* 4, 3 (1991), 543–569.

[15] Guy Kaplanski and Haim Levy. 2010. Exploitable predictable irrationality: The FIFA World Cup effect on the US stock market. *Journal of Financial and Quantitative Analysis* 45, 2 (2010), 535–553.

[16] Harold Y Kim and Jianping P Mei. 2001. What makes the stock market jump? An analysis of political risk on Hong Kong stock returns. *Journal of International Money and Finance* 20, 7 (2001), 1003–1016.

[17] Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*. 1188–1196.

[18] Ronny Luss and Alexandre dâĂŹAspremont. 2015. Predicting abnormal returns from news using text classification. *Quantitative Finance* 15, 6 (2015), 999–1012.

[19] Jussi Nikkinen and Petri Sahlström. 2015. Impact of Scheduled US Macroeconomic News on Stock Market Uncertainty: A Multinational Perspecive. (2015).

[20] John R Nofsinger. 2005. Social mood and financial economics. *The Journal of Behavioral Finance* 6, 3 (2005), 144–160.

[21] Gabriele Ranco, Darko Aleksovski, Guido Caldarelli, Miha Grčar, and Igor Mozetič. 2015. The effects of Twitter sentiment on stock price returns. *PloS one* 10, 9 (2015), e0138441.

[22] John V Richardson Jr. 2006. The library and information economy in Turkmenistan. *IFLA journal* 32, 2 (2006), 131–139.

[23] Nadine Strauß, Rens Vliegenthart, and Piet Verhoeven. 2016. Lagging behind? Emotions in newspaper articles and stock market prices in the Netherlands. *Public Relations Review* 42, 4 (2016), 548–555.

[24] Paul J Werbos. 1990. Backpropagation through time: what it does and how to do it. *Proc. IEEE* 78, 10 (1990), 1550–1560.

[25] Chayanin Wong and In-Young Ko. 2016. Predictive Power of Public Emotions as Extracted from Daily News Articles on the Movements of Stock Market Indices. In *Web Intelligence (WI), 2016 IEEE/WIC/ACM International Conference on*. IEEE, 705–708.