Predicting Major Market Events from Media Headlines and Multimedia Items

Michael Di Amore, Yaqing Wang, Yitong Xu, Varun Aggarwal
Columbia University
In association with Capital One

October 27, 2017

I. Introduction

Since the inception of markets, predicting major market events has been an ambitious goal studied heavily by the finance community. Event based trading strategies have been explored a number of times in the literature, and whole hedge funds have been founded on its premise. The goal of this project is to predict rare market events using the GDELT database. With the intent of providing Capital One with a powerful tool for predicting changes in the market that may affect their business. The database GDELT is a publicly available database that catalogs events across the globe. It is touted as the "Global Database of Society" and is the main data-source of our project.

II. Goals

Our goal is to develop a tool to help predict major market events, we are currently exploring two approaches. The first, phrases the problem in terms of classification, predicting whether or not the market moves in excess of some threshold. *Michael* has currently made progress on this front and has employed both traditional and deep learning with promising results. The second, phrases the problem in terms of regression and we believe a fine-grained time series forecasting model will be extremely helpful. *Yitong* is currently working on developing a model which could assist in detecting major market changes. If we can accurately predict the market changes in a short time period, e.g. one or two weeks prior, then Capital One can make internal changes to position themselves accordingly. Although well-known autoregressive moving average model *ARMA* and its variants demonstrate their effectiveness in modeling time series data, recent advancements in *deep learning* and its powerful nonlinear approximations achieve the state-of-art accuracy in stock market predictions and weather forecasting [1]. We will discuss our preliminary result in section VI.

III. Data Description (Features)

The GDELT database consist of two primary sub-databases that are relevant to us, the "Events" database and the "General Knowledge Graph (GKG)". The GKG is quite large, with a size of over 2.5TB for just the last year alone. We have primarily been using the Events database for ease of use and to get a head-start on exploring and modeling the data. Should the need arise we should be able to use the GKG however that would require us to use big data tools and write some custom connectors in Spark.

For querying the data, *Michael* has written a query script that calls python wrappers to

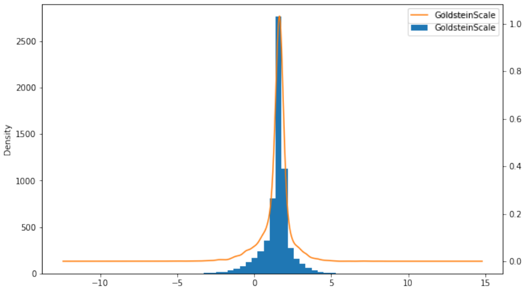


Figure 1: *Goldstein Scale Distribution*

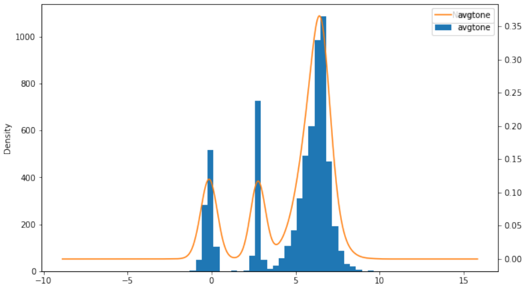


Figure 2: *Average Tone Distribution*

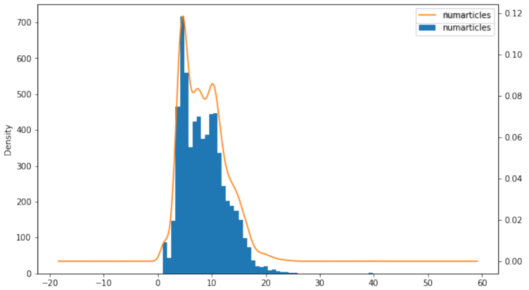


Figure 3: *Number of Articles Distribution*

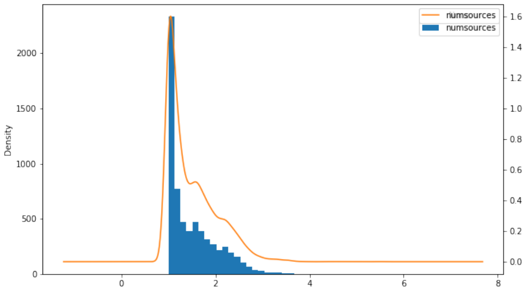


Figure 4: *Number of Sources*

Pandas/GoogleBigQuery methods in order to tap into the GDELT database. There are also wrappers for querying the Quandl API and Yahoo! Finance API.

As a baseline for modeling and exploration we focus on the GDELT Events Database for articles related to business and in the US for the period from 2000-01-01 to 2017-10-06.

We focus on the sentiment metrics and volume quantities associated with each article. This allows us to frame the problem as a supervised learning task. Sentiment metrics include the Goldstein Scale, which is a measure of sentiment from the perspective of stability. See Figure 1 for the distribution of the Goldstein scale in our dataset.

Average Tone, which is a measure of positive or negative sentiment. Average Tone values range from -100 to +100 to see the distribution in our dataset refer to Figure 2

For Volume metrics we look at number of articles, number of sources, and number of mentions. These tell us the frequency with which the news story was published. We present these together in Figure 3 and Figure 4.

We also use some financial metrics as features. Such as the VIX (the CBOE Volatility Index), unemployment rates, M1/M2 velocity, fed funds rate, and others.

These are all gathered from Quandl primarily through the FRED Database. We present some of their graphs below. Figure 5 is the plot of the effective federal funds rate. Figure 6 is

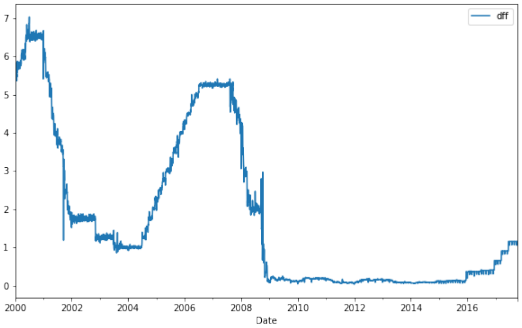


Figure 5: *Effective Federal Funds Rate*

the plot of the VIX over the time interval we are studying, as we can see there is an extremely high volatility regime in 2008 and a historically low volatility regime today.

Lastly we present the unemployment rate, which as expected is highest during the period immediately after the crisis and begins to taper off as we enter recovery

IV. Data Description (Targets)

What constitutes a market event? This is a fundamental question that we needed to answer. We focused on the idea of volatility defining an event, specifically volatility with relation to the SPX index. With the guidance of our mentors we tried a number of metrics, such as $metric = 1.5\sigma$ where σ is the 10year annualized standard deviation adjusted to be on a daily scale i.e. by diving by $\sqrt{(252)}$. We tried other metrics as well such as $metric = 6\sigma$ and $metric = 4\sigma_{roll}$ Where $sigma_{roll}$ is the half a year rolling average. Currently we are using $metric = \mu_{roll} + 3\sigma_{roll} \;< r \; or \; \mu_{roll} - 3\sigma_{roll} > r$ Where μ_{roll} is the rolling mean return and $sigma_{roll}$ is the rolling standard deviation

Using the last definition have 20 events over a 17 year period, which is a large imbalance of about 0.50% positive classes vs 99.50% negative classes as shown in Figure 8

We now examine and hope to draw insight from the plot of rolling volatility. In Figure 9, we present the rolling volatility with the events superimposed upon them. As we can see most of the mass of the events takes place in the 2008-2009 era, in fact if we looked at the statistics 5 market events take place in the period from the start of 2008 to the end of 2009, or 30% of our targets.

Let us now examine the distribution of rolling volatility overall as shown in Figure 10. We have what appears to be a bimodal dis-

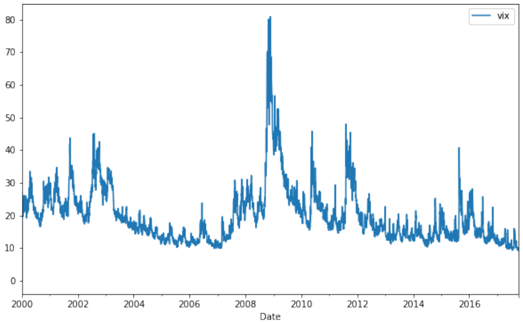


Figure 6: *VIX Level*

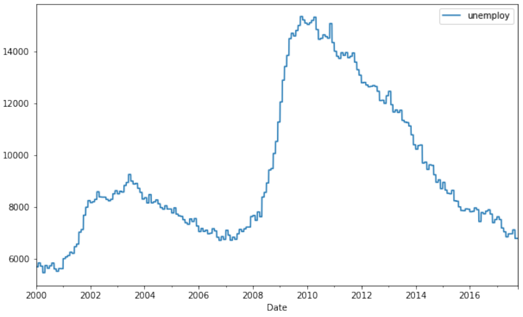
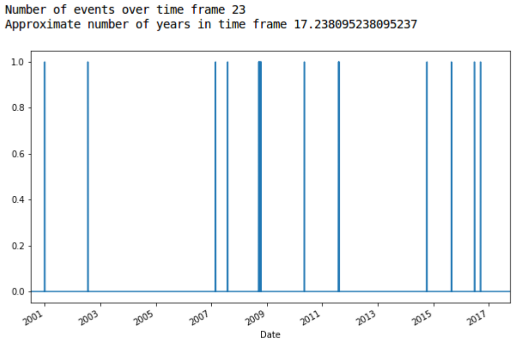


Figure 7: *Unemployment Rate*



```
Number of events over time frame 23
Approximate number of years in time frame 17.238095238095237
```

Figure 8: *Delta Functions of Events*

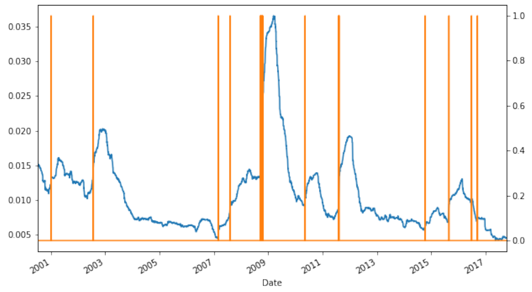


Figure 9: *Rolling Volatility with Events*

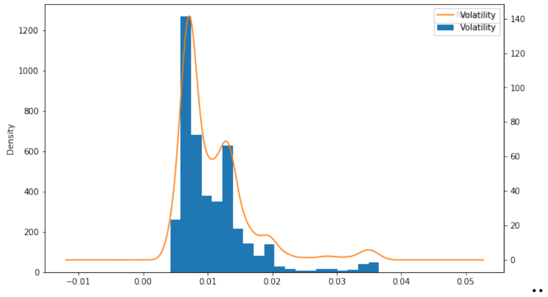


Figure 10: *Rolling Volatility with Events*

tribution with significant right tail. The KDE is a bit misleading here as it has negative values, when in actuality volatility can't become negative.

V. Data Summary

In this section we summarize the main points of our dataset, and speak to some modeling concerns

1 Data is heavily imbalanced, will require a clever method to deal with this. Perhaps some combination of oversampling and undersampling

2 Data is a time series, therefore we need to be weary of any autocorrelation present within the dataset and make sure to use proper cross validation methods where applicable

3 Our mentors have encouraged us to experiment with different definitions of a "market event" we should continue to explore this space

VI. Progress so far

After we developed the query mechanism for the data from the GDELT Databse, we started on data exploration. We also discussed what exactly a market event was and at the beginning we went back and forth on a precise definition. During this time our mentors suggested we experiment with different definitions, here are our preliminary findings.

Yaqing tried to model AAPL, as suggested by our mentors, to determine if we could predict possible market events such as the effect of the Iphone/Ipad release on AAPL stock. Figure 13 show some results form that effort. *Yaqing* also ran an XGboost regression (Gradient Boosted Trees) that proved somewhat fruitful.

Varun tried to model GOOG, and came across some interesting findings. He first examined if the sources of the news were reputable and found that the most of the stories came from a source called "contacto latino" as shown in Figure 11

Varun also examined the distribution of GDELT events database showing that the GDELT events database alone might not be enough to predict market outcomes and provided some insight into proper preprocessing techniques for scaling the GDELT features. His results can be found in Figure 12. The first plot is detrended google stock prices with a

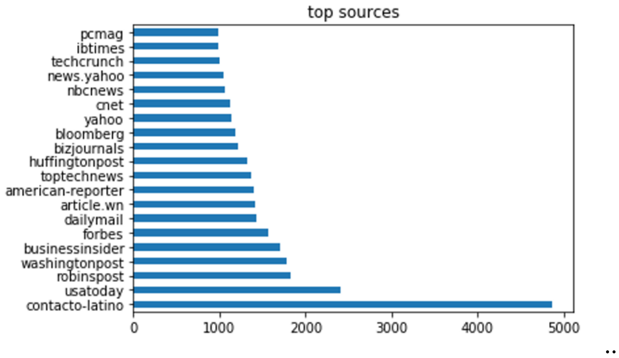


Figure 11: *Sources of news*

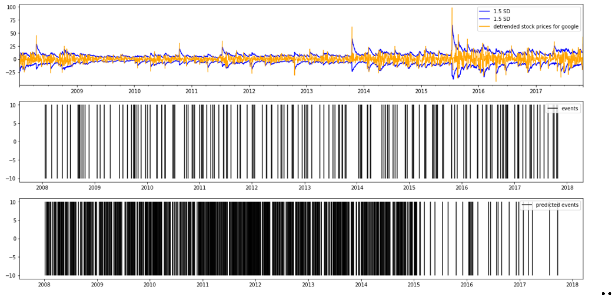
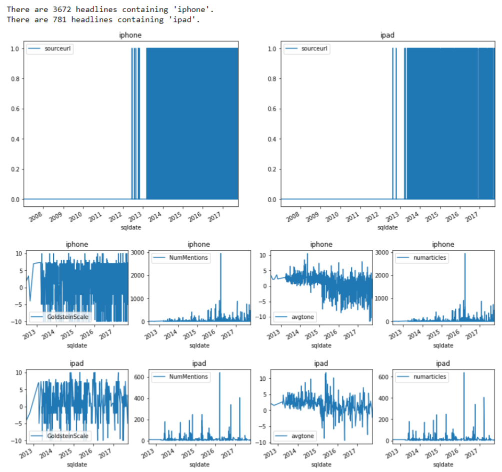


Figure 12: *Exploration of Google*



```
There are 3672 headlines containing 'iphone'.
There are 781 headlines containing 'ipad'.
```

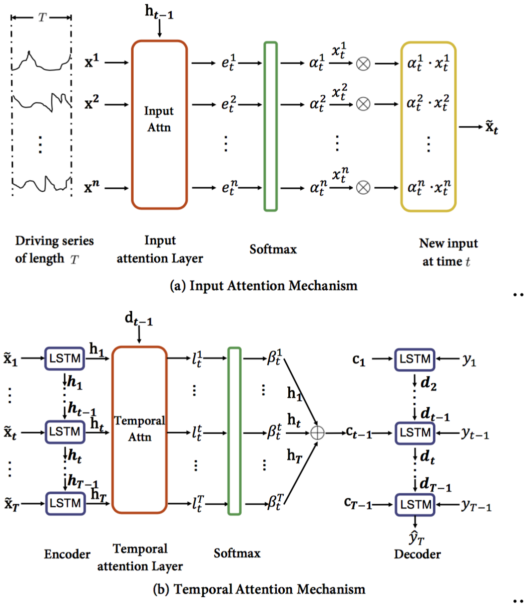
Figure 13: *Exploration of Apple*

1.5 standard deviation band. The second and third images are a comparison of actual events under this definition to predicted events.

Yitong focused on our continuous time-series model, and since we hadn't settled on a precise definition of a market event he used weather data as a proxy. In his modeling he used a Dual attention autoencoder-decoder. The detailed model structured can be found in Figure 14. We show some of our preliminary results here. As shown in Figure 15, we can see the model has powerful capacity to predict the weather changes, even some extreme weather conditions. Since we get a clearer idea on the definition of market events now. In the next step, we will do some fine tuning and then transfer the model to financial continuous data, and test on it.

We then looked to expand our feature space by taking other financial indicators such as the VIX (CBOE Volatility Index) as a feature along with other financial indicators. We also looked into using the Gold and Crude Oil futures as possible features, but we would have to deal with the "rolling" of these contracts to make them continuous.

With the expanded feature space and clearer market event definition, *Michael* posted some promising results from a second pass at Deep



(a) Input Attention Mechanism

(b) Temporal Attention Mechanism

Figure 14: *Dual Attention Time Series Encoder Decoder*

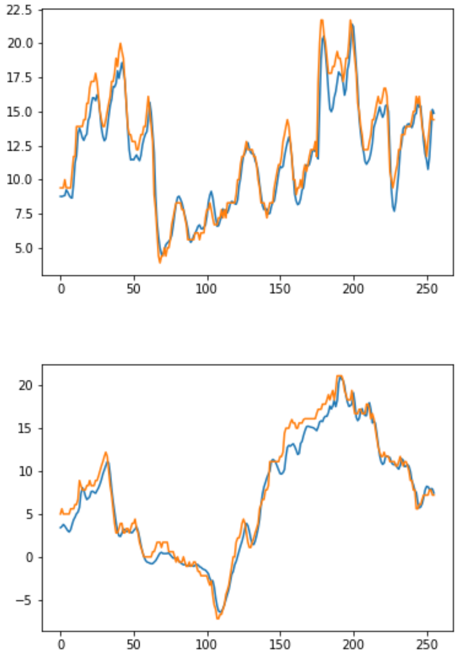


Figure 15: *Preliminary Results from Time Series Deep Learning Model*

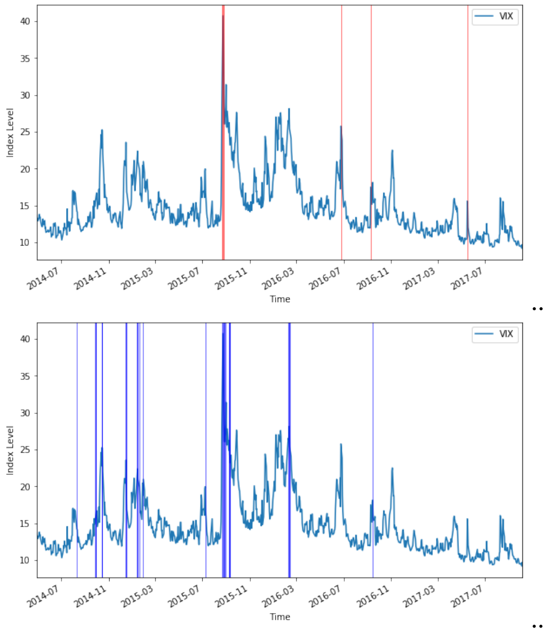


Figure 16: *Results on VIX for Deep Learning Model*

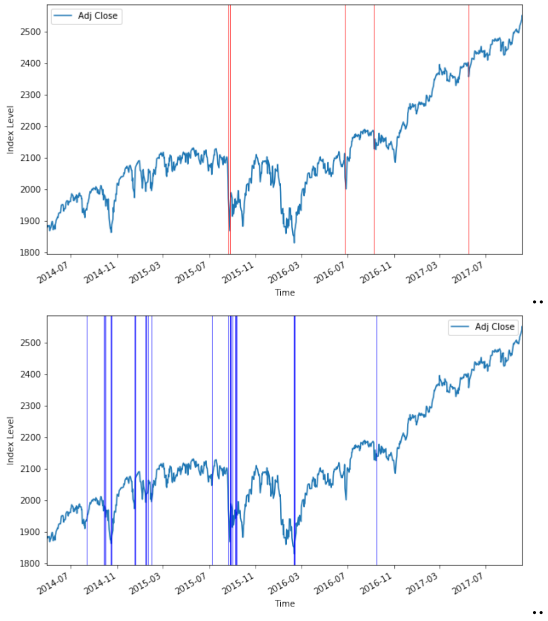


Figure 17: *Results on SPX for Deep Learning Model*

Learning. Shown in Figure 15 in regards to the VIX and Figure 16 in regards to the SPX. The red lines represent the actual dates of events while the blue lines represent the predicted events. As you can see the predictions are "on" for a period before and after an actual event, and even seem to pick up on smaller patterns not considered an event.

As it stands now we've completed most of our exploration of the data and are now taking serious steps toward modeling.

VII. Literature Review

We include some recent papers in our literature review about predicting market events from news sentiment. Perhaps the most similar to our goal is the paper by Xiong,Nichols,and Shen [2] titled "Deep Learning Stock Volatility with Google Domestic Trends" in this paper the authors use the information from google domestic trends to predict SPX Volatility. We draw some inspiration from this paper such as proper scaling of our inputs, what to consider our target, and how to construct the architecture of our Deep Net.

Another paper, "Forecasting abnormal stock returns and trading volume using investor sentiment: Evidence from online search" [3] aims to use ticker lookups as a proxy for sentiment. This is an interesting idea and is similar to the use of the GDELT metrics as a proxy for investor sentiment

A paper comparing GARCH(1,1) which is a commonly used time series model for estimating volatility, to using Google search results as a proxy for this estimate might also be of interest. Although Smith's paper [4] is focused primarily on foreign exchange markets this is an interesting take on the problem from the point of view of econometrics.

In classical economics event studies are used to assess how an event will impact the value of the firm. Though not entirely relevant to our current discussion of market events, we feel we should include the notion of "event studies" for posterity as it can be seen as a similar goal to ours. For information on event studies see Mackinlay [5] .

References

[1] S. Wang, J. Sun, B. J. Gao, and J. Ma, "A dual-stage attention-based recurrent neural network for time series prediction," *International Joint Conference on Artificial Intelligence (IJCAI)*, 2017.

[2] R. Xiong, E. P. Nichols, and Y. Shen, "Deep Learning Stock Volatility with Google Domestic Trends," *ArXiv e-prints*, Dec. 2015.

[3] K. Joseph, M. B. Wintoki, and Z. Zhang, "Forecasting abnormal stock returns and trading volume using investor sentiment: Evidence from online search," *International Journal of Forecasting*, vol. 27, no. 4, pp. 1116 – 1127, 2011.

[4] G. P. Smith, "Google internet search activity and volatility prediction in the market for foreign currency," *Finance Research Letters*, vol. 9, no. 2, pp. 103 – 110, 2012.

[5] A. C. MacKinlay, "Event studies in economics and finance," *Journal of Economic Literature*, vol. 35, no. 1, pp. 13–39, 1997.