Big Data Analytics Capstone

Project Paper

Damian Etchevest

St. Thomas University

CIS-627-170

Contents

Introduction

Trading and war are extremely opposite, but they share some communalities. For once, trading does not imply the use of weapons to oppress an enemy; however, someone always loses because of lack or presence of a weaker strategy. In this market driven world, information is power. Adverse selection refers to a scenario in which the parties on a trade have unequal information over the subject. This study acknowledges that institutional investors have access to news before the public does, but… What is it the impact of information on stock prices? Our research will pursue the development of an automated process which feeds from stock prices, from *"Dukascopy",* and news feeds, to determine the presence of informed traders, length, and strength of several classes of sources. Section 1 uses methods to heuristically describe the use of classification models to predict the presence of decompressed markets, a scenario in which the price information is non-redundant, in other words, the entropy is high.

Background Analysis

Introduction

These article serves as an extension of the work performed by Maximilian Frankie and Sean Mondesire [1]. On it, they explore the relationship between published news and the stock market reaction. Their research results achieve corresponding *Mean Square Error (MSE)* of 0.039 and an accuracy of 0.652. In contrast to their research, we explore the development of an information driven wave pre-news release and test the confirmation of the tendency on post-news release.

Section 1 - Information Driven Bars & Entropy

Section 1 deep dives into the development of an automated signal generator which performs with live feed stock data to determine when the presence of informed traders exists. This last term, informed traders, refer to the presence of traders that possess fundamental or alternative company's information and generate trades upon understanding its respective future market impact.

Different methods have been tested that can be useful to disclose the information contained in stocks prices. Georges Dionne and Xiaozhou Zhou [1] use *Limit Order Book (LOB)*, a type of

order to buy or sell an instrument at a determined price or better, to identify *High Price Impact Trades (HPTIs)*¸ trades that relate to informed traders which make transactions associated with large price changes relative to their volume proportion. Passive limit orders, from retail investment, are contrarian, consistent with the liquidity provision hypothesis. In contrast, the implementation of Fama-MacBeth regressions on order imbalance measures proposes that aggressive market orders (another form of retail investments) can predict future news, suggesting that retail investors are mainly informed about firm-level news, and that they are likely to have valuable private information [2].

On the other hand, Lopez de Prado [3], claims that with the advances of algorithmic trading, which takes parent orders and chops them into child orders, trading become more complex, since orders and traders are not congruent, and 87% of the executed child orders are passive, therefore blurring the underlying intentions of the parent order. In his study, Lopez de Prado, analyses market microstructure and examines the accuracy of three methods for classifying information driven trades, *the tick rule, the aggregated tick rule, and the bulk volume classification methodology*.

Methodology & Model Design

Introduction

The aim of this study is to determine the relationship between news events and the stock market. To accomplish this, we study the patterns of behavior on days which have a news event and test them against those days which do not have news releases on market hours. We are looking to determine the wave (pattern) on market parameters.

Data Sources

The data sources utilized for stock market data differ from those of news data. For the first ones, we use the API from Dukascopy, which allows us to pull historical data on a considerably number of instruments at a rate of 5000 quotes per call. The API requires an application, but it is essentially free and fast. On the other hand, for the later resource, we utilize various financial new sources including but not limited to Bloomberg, Wall street Journal, and the Financial Times.

As stated before, the purpose of this project is to test if market waves behave significantly different in days of events as on those days of non-news releases. Therefore, top of book market data, which provides order matched data, should be the most beneficial form of source. Nasdaq, NYSE, and IEX represent the list of stock exchanges where our instruments of interest trade. Our research assumes that institutional investors perform mostly on Nasdaq. However, the capital investment that it would require to get top of book market data makes it difficult for readers to replicate the research. Therefore, we utilize tick-by-tick data, which is free and consists on bid-ask, a two-way price quotation that indicates the best price at which a security can be sold and bought at a given point in time, and their respective volumes.

Our research focuses on the behavior of FAANG (Facebook, Amazon, Apple, Netflix, Google) stocks market. The reason behind this decision is because of the high frequency of news release and the volatility on their respective markets.

In order to develop a model that does not overfit, the rule suggests training with a good volume of samples. For this project, we utilize data from market hours in 2019. The dataset was built with a loop, which makes calls on market hours, through every day of that year. The call parameters include key, instrument, timeframe (tick), count (5000 as max), start and end (from 9:30 to 15:56).

Data processing

As stated before, the data pull is an automated batch process which makes calls every five minutes (in order to allow the market accumulate ticks), at a rate of 5000 quotes per call, which is the maximum allowed by the source. At every call, we process the data by developing volume and tick bars. This decision goes in hand with the assumption that time bars produce less accuracy [3], as they intend to produce information at a fix interval (time). Therefore, volume and tick bars provide both the benefit of synchronize sampling (at a pre-determined threshold of an estimated amount of bars per market session) with a proxy of information arrival, and of normalizing the returns, which is assumed on most machine learning algorithms.

This research assumes that partial information should be carried form bar to bar, by doing so, we focus on market trends while the intention is to avoid bar spikes (outliers). To do so, the last quartile of the ticks contained in the current bar are stored and added to the next bar. Further analysis is required to verify the significance difference of the method.

Several variables are calculated when developing the bars:

- Ticks inside the Bar
- Volume from bid-ask
- Price (different formulas)
- Spread of bid-ask quotes
- Tick Rule (price change from ticks)

Procedure

In order to determine if there is significantly different pattern of behavior developed on market days where a news catalyst occurs, we have centered the analysis on the aggregated tick and volume bars from the different instruments processed before. By doing so, we are certain that any analysis does not overfit a specific instrument, and therefore, be confident to make claims about the results which represent the real behavior of the market psychology relevant to any instrument.

Some relevant assumptions are to be tested in the research, which include that before news release, market movements significantly different from non-news release days express the intentions of informed investors, which might refer to institutional investors. Previous research use methods to find the probability of the presence of informed traders. To do it, they use top-of-book market data to predict on three proxies: the aggressor side of trading(as given by buy-sell indicator flags in the data), an estimate of spreads, and the permanent price effect of trades. Further analysis will apply those methods and test for the accuracy of our model against those. On the other hand, after news release market movements significantly different from non-news release days express the intentions of the public, or retail investors. To address the difference on before/after news release sessions, we develop an index which is 0 at the moment of the news release, or the closest on distance bar.

Another key method implemented involves the development of a function which calculates the normalized (to deal with outliers) absolute (to deal with negative values) percentage change on any specified parameter. These procedures allow us to evaluate index detailed behavior on catalyst days, therefore being able to identify the exact index location of the average wave start, end, and strength.

Results

The goal of the research is to find and to explain the market momentum generated by news release. This section statistically proves that the news catalyst generally generates an impact on the stock market spread, volume, sampling frequency, and price.

To begin with, the first section of the research was dedicated into finding statistical differences on the daily and overall aggregated parameters on the presence of news against non-news days. The following tables provide the p-values of the tests, where p-values smaller than or equal than 0.05 result on confidence to reject the null hypothesis of non-significant difference:

OVERALL AGGREGATED VOLATILITY (P-VALUES)

	Tick Bar	Volume Bar
CUMULATIVE SPREAD	6.944522e-12	8.025555e-10
VOLUME	7.223194e-148	2.472206e-220
VOLUME WEIGHTED AVERAGE PRICE	1e-06	0.020242

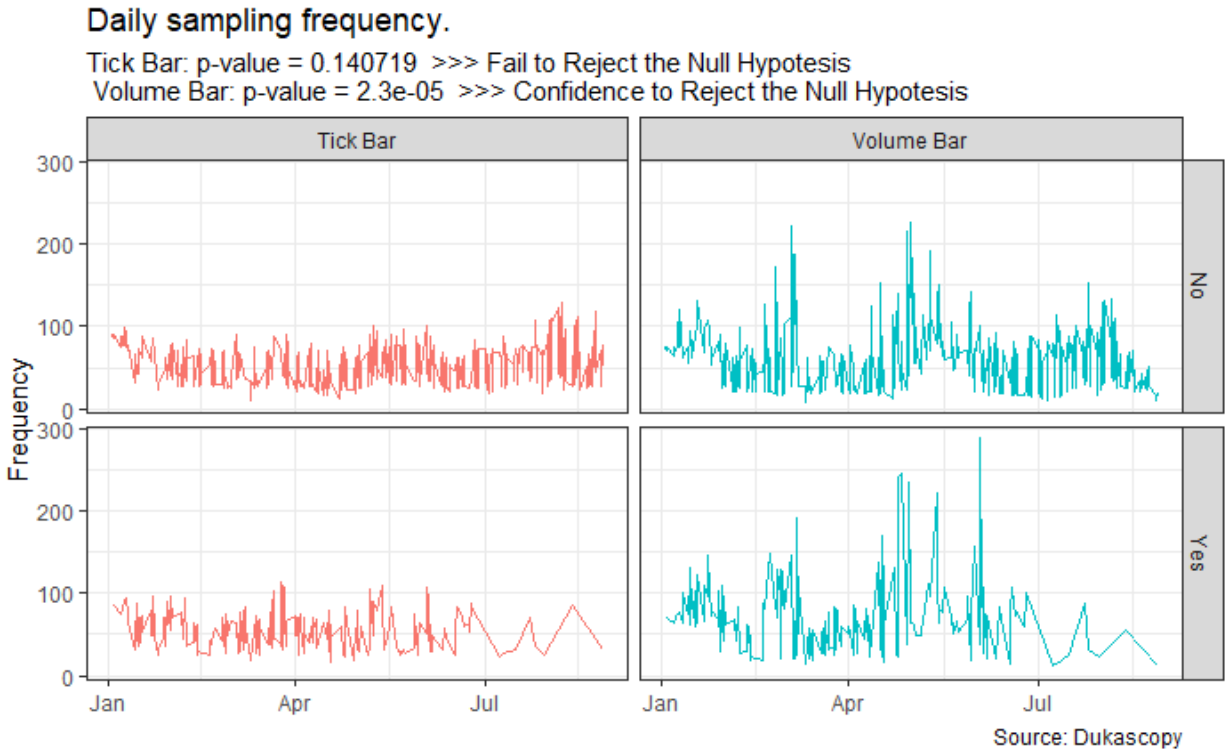
The table above confirm statistical differences among the means of the populations in both tick and volume bars. These results serve as an introductory view, significance differences on the presence of news invites us to deeply analyze these differences.

DAILY AGGREGATED VOLATILITY (P-VALUES)

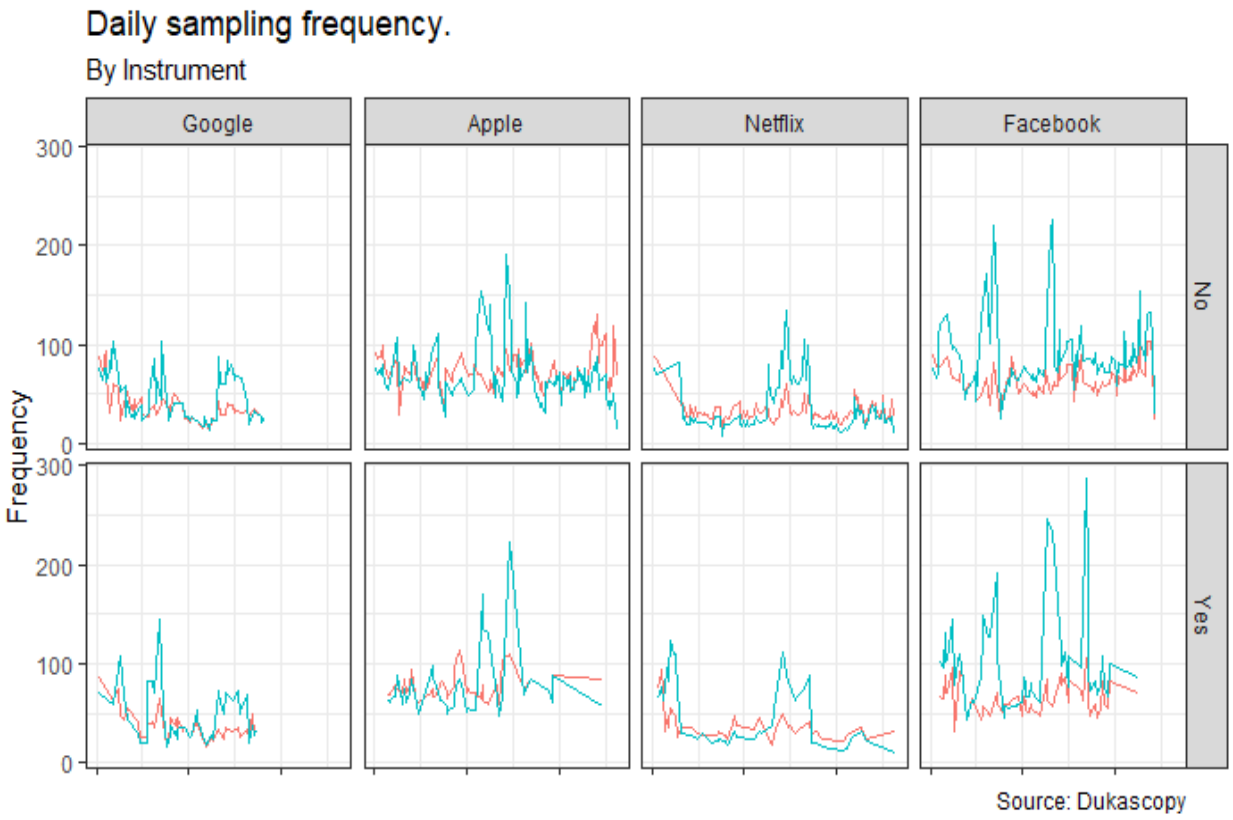
	Tick Bar	Volume Bar
CUMULATIVE SPREAD	0.034393	0.080944
VOLUME	0.636922	0.499073
VOLUME WEIGHTED AVERAGE PRICE	0.035065	0.056408

In contrast, this table aggregates the data by day and statistically compares the average standard deviations, of the respective parameters, on the presence of catalyst or not. In this case, only tick bars show significant differences, on the cumulative spread and the VWAP parameters.

Another key parameter to understand the difference on behavior, produced by a news release event, is the sampling frequency. Tick and Volume bars are not sampled at a fix time interval; however, they are developed when the market tick data meets certain threshold. Therefore, it would be reasonable to hypothesize that on days of news release, the market should have more movement, therefore, the frequency of bars sampling should be significantly different than those days of no news release. The following graphs strives to test the hypothesis:



Statistically, volume bars show significant difference on the number of sampled bars on the days when news catalyst arrive the market. These results set the base for further analysis, acknowledging that, generally, information will be gain as the market reacts. In contrast, time bars, as they sample on time values, do not conceive information about the market stress or depress, as sampling will be performed regardless of the investor's intentions shown by orders.



So far, the analysis has proven that there is a generalized significant difference on the behavior of the market on days when there is news to be released (informed investors), or already released (retail investors). Next, the research assigns indexes to the respective bars based on the distance from the news event. The purpose of these procedure is to analyze the index wise percentage change of the assigned parameters and test their significant differences in order to assign an average start-end to the wave.

The following graph aims to visualize “the wave” by plotting the index (distance from news release) versus the index-wise percent change, and to test for significant differences of the average percentage changes:

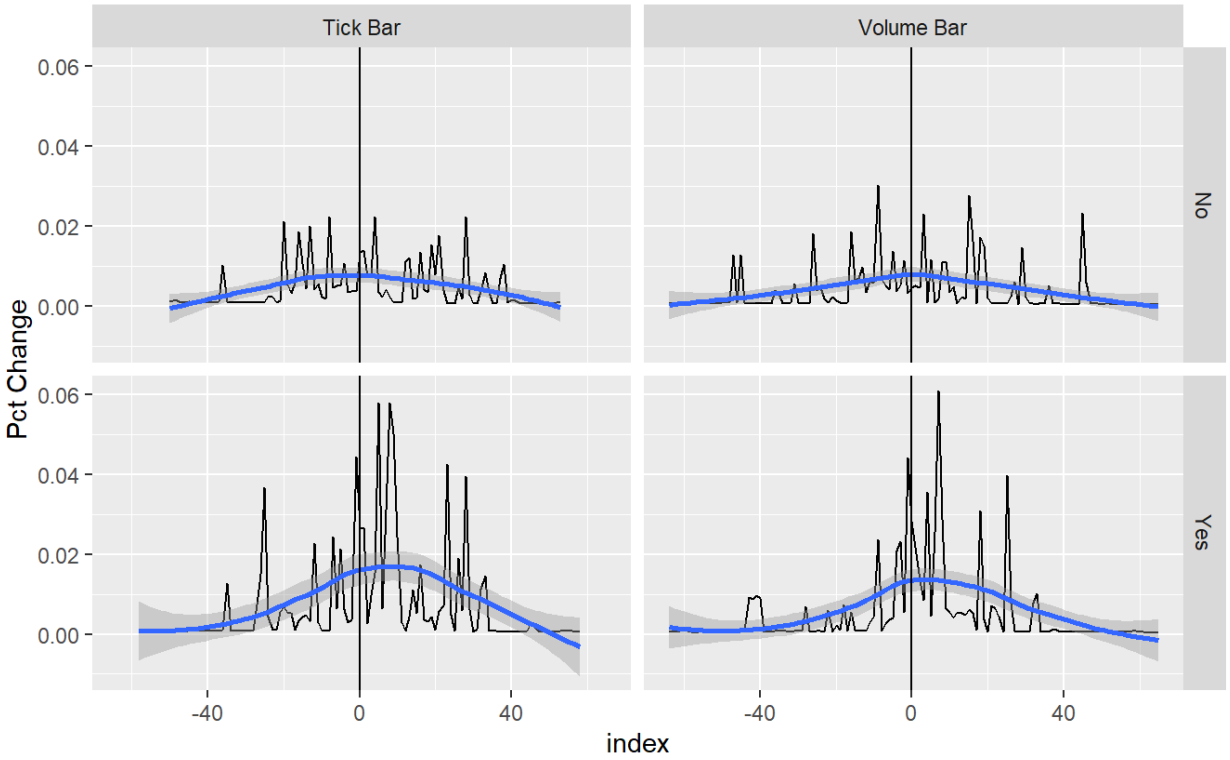
Index Series of % Change

Tick Bar Volume: p-value 0.001817 >>> Confidence to Reject the Null Hypotesis
Volume Bar Volume: p-value = 0.000157 >>> Confidence to Reject the Null Hypotesis
Tick Bar Spread: p-value 0.126011 >>> Fail to Reject the Null Hypotesis
Volume Bar Spread: p-value = 0.000176 >>> Confidence to Reject the Null Hypotesis.

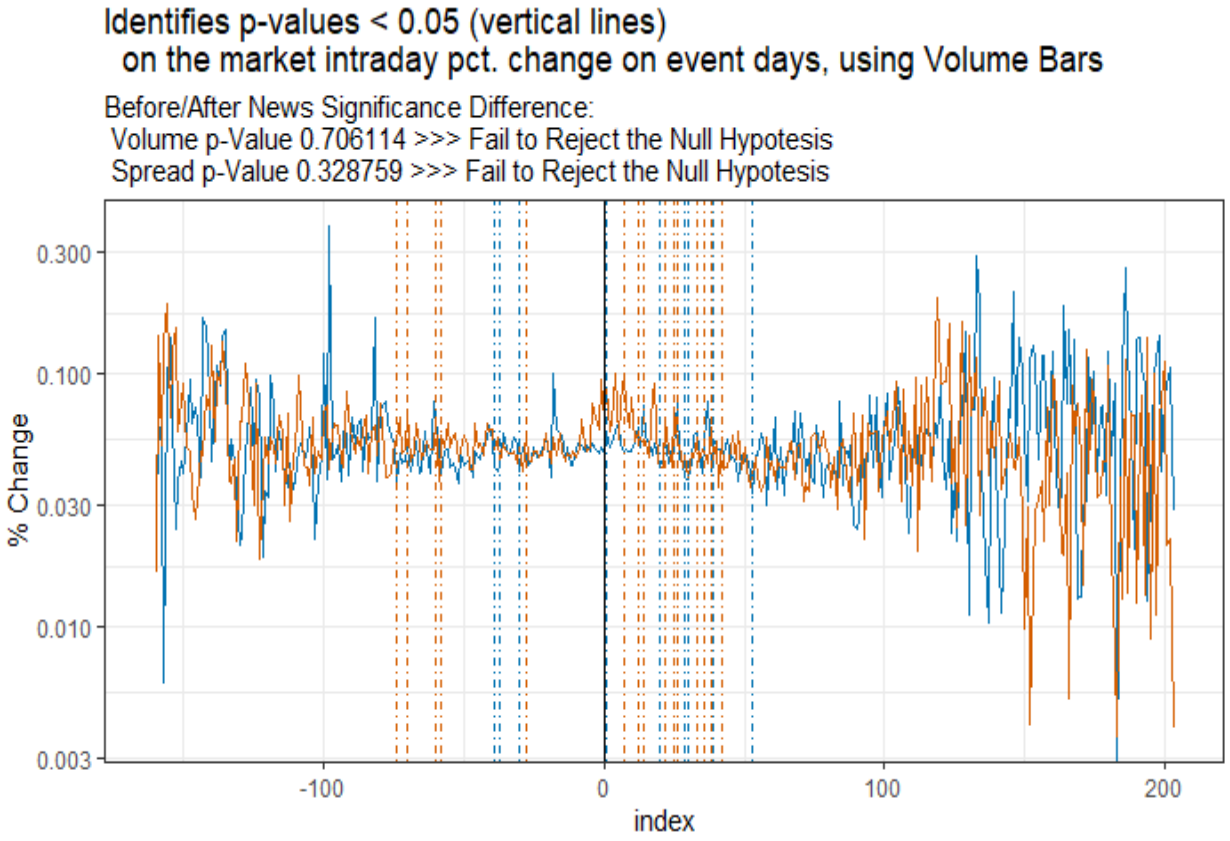
Source: Dukascopy

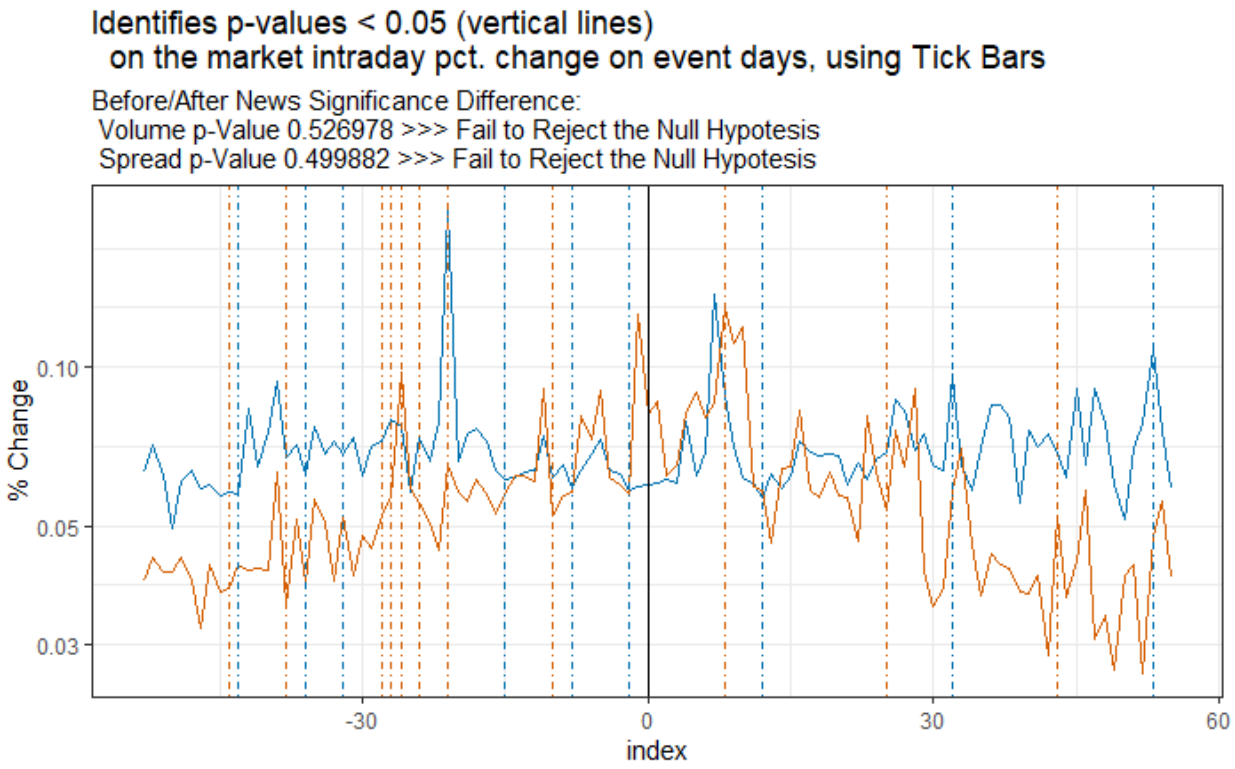
Index Series of % Change

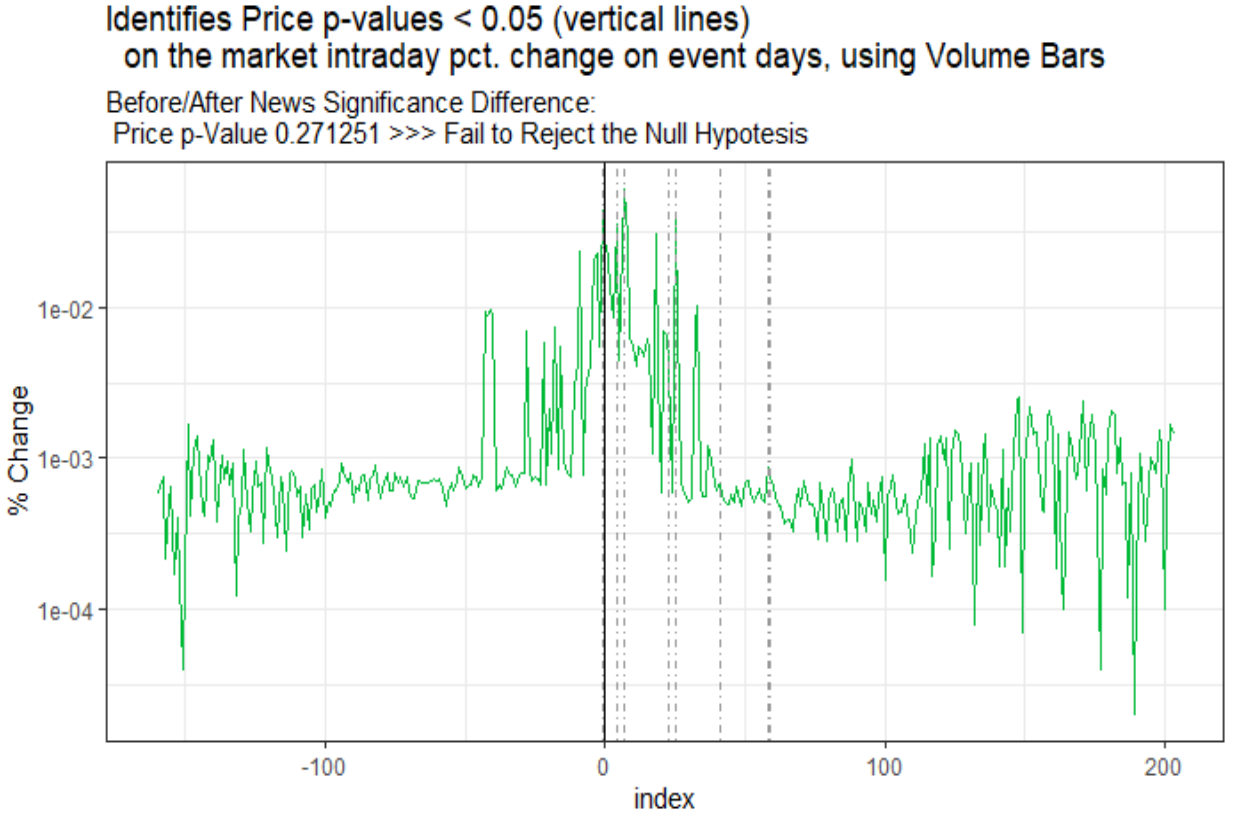
Tick Bar Price: p-value 0.040677 >>> Confidence to Reject the Null Hypotesis.
Volume Bar Price: p-value = 0.209389 >>> Fail to Reject the Null Hypotesis.

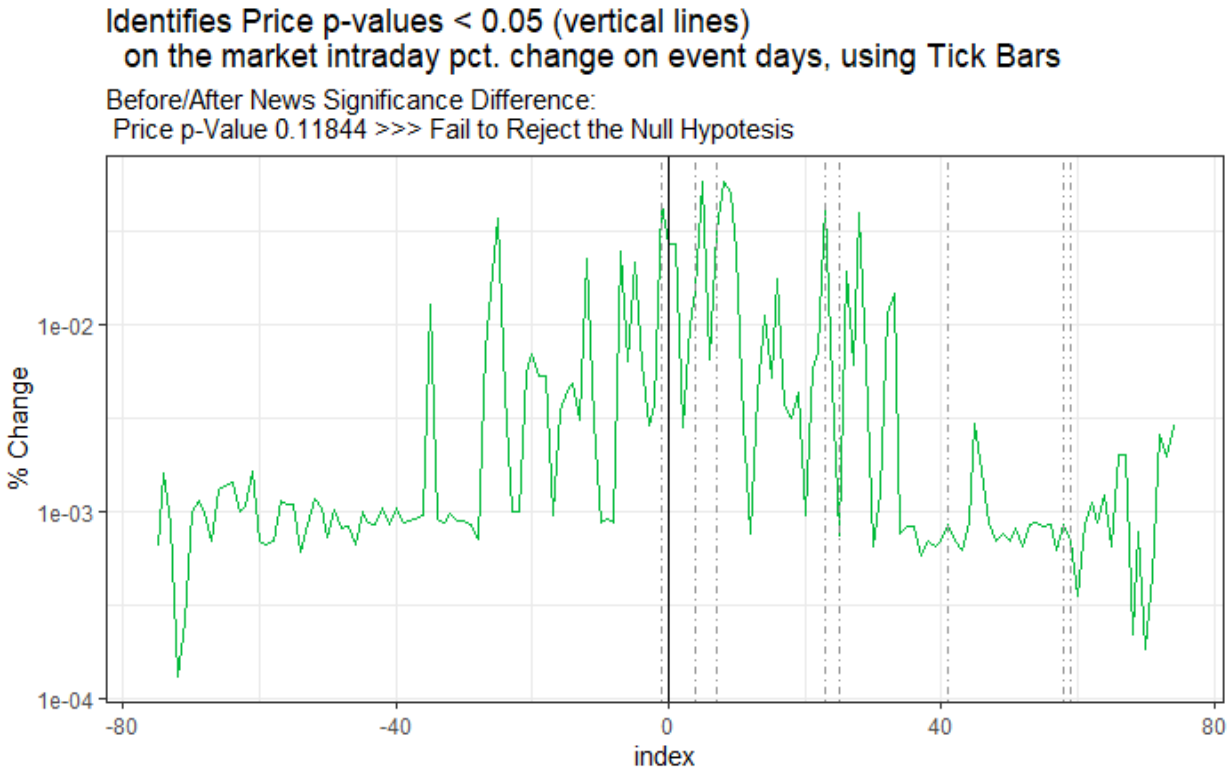


Finally, the research strives to find the average moments when these significant differences occur. For that, we test the statistical differences on days of events against days of no events, for the average percent change at every index. The following graphs show the tick-volume bar spread and volume representations, along with vertical lines which show the exact index locations at which statistical differences were confirmed:




Identifies p-values < 0.05 (vertical lines)
on the market intraday pct. change on event days, using Tick Bars
Before/After News Significance Difference:
Volume p-Value 0.526978 >>> Fail to Reject the Null Hypotesis
Spread p-Value 0.499882 >>> Fail to Reject the Null Hypotesis
% Change
0.10
0.05
0.03
-30
0
30
60
index


Identifies Price p-values < 0.05 (vertical lines)
on the market intraday pct. change on event days, using Volume Bars
Before/After News Significance Difference:
Price p-Value 0.271251 >>> Fail to Reject the Null Hypotesis
% Change
1e-02
1e-03
1e-04
-100
0
100
200
index



Conclusions

In conclusion, our research tested whether news catalyst had, in general, an impact on the market volatility. We have tested that both tick and volume bars accurately show these differences, but they also convey information more frequently when markets are moving more (more market orders and more volume being traded).

Wave (by index)

Bar	Measure	Start	End	Duration
Volume	Spread	-74	42	116
	Volume	-39	53	92
	Price	-1	59	60
Tick	Spread	-44	43	87
	Volume	-43	53	96
	Price	-46	35	81

Also, we have proven that, in general, before and after news sessions are not significantly different, which reflects that information influences trades (informed investors) before the news are released to the public.

The research serves as a ground to deep dive into the research of machine learning automated models which would analyze the patterns to signal significant probabilities that news release may arrive and its direction based on the influence on the trades, setting investors steps ahead of the public.

Future Research

Introduction

Some additional sections are provided here, which show the path of future research:
Section 2 would exploit clustering as means to develop stock profiles, referring to the traders that manipulate it and to the historical characteristics of the stock, while analyzing their behavior on the presence of information. Section 3 would utilize time series predictions to measure the inference of alternative data sources (offers the opportunity to work with truly unique, hard-to-process datasets) on the stock prices, as a means to determine the characteristics of the information-driven price waves. Section 4 develops an automated system to feed live news into the data warehouse and then use neural networks to confirm the wave characteristics. It would also compose a structure to take actions based on the research. Define the weight of the different models and add them to a stochastic strategy.

Background Analysis

Section 2 – Sentiment Analysis & Momentum

Lots of research has been done exploring the sentiment of news and their market reactions. This section utilizes time series predictions to measure the inference of alternative data sources (offers the opportunity to work with truly unique, hard-to-process datasets) on the stock prices as a means to determine the characteristics of the information-driven price waves.

An interesting approach aims to predict Chinese stock market movements, tendency [4]. To do that, they implement *support vector machines (SVM),* to classify the tendency, using *text Mining* on online news, their respective comments, and stock market data. Also, their findings show that

source's news quality and audience number can serve to calculate their influence by predicting results' difference from the normal prediction results.

A similar study examines the predictability of market reactions to bad news [5]. Their approach is innovative as they first use *time series clustering*, to generate news clusters based on subsequent stock returns, and then apply *SVM* to classify the features extracted from each cluster, by *natural language processing (NLP)* which is a method that interprets human language. Their research identifies four types of market reactions after the news becomes public: downward drift, short-term reversal, medium-term reversal, upward drift.

In their article [6], the authors explore the impact of *Seeking Alpha* (crowd-sourced content service for financial markets) on the stock market's movements after SA research articles are published. They use classification and regression methods and find incremental in order imbalance, significantly related to the sentiment of research articles and comments, which begins within half-hour after SA publications.

Lopez de Prado and Rebonato [7] examines the use of *Kinetic Component Analysis (KCA* - a state-space application that extracts the signal from a series of noisy measurements by applying a Kalman Filter on a Taylor expansion of a stochastic process - on tick data to determine what they refer as "financial inertia". They compare the use of KCA to other signal processing tools as *Fast Fourier Transform (FFT)* or *Locally Weighted Scatterplot Smoothing (LOWESS)*, and find that their method: provides confidence intervals estimates of the signal's position, define the wave's characteristics (velocity and acceleration of the series), does not exhibit Gibbs phenomenon, and it can be updated online to be forward-looking and resilient to structural changes.

Section 3 – Network Analysis & Instrument Profiling

Finally, we develop clusters of instruments and their respective networks to identify significant information transfers in the different stock networks. The aim is to profile the different stocks to develop behavioral classification models to predict the price wave.

Stephen Taylor [8] fundaments the use of Fisher Information Metric, which is relevant in the fields of information geometry and computing geodesics, to generate clusters of stocks based on their distribution. They contribute valuable theory examples: nearest neighbor comparison using

the generalized Pareto distribution, and the maximum daily loss over and annual period distribution hierarchical clustering techniques using the generalized extreme value distribution. Even though his findings do not consider instruments networks, it provides behavioral features that contribute to the stocks profiling.

Pawel Fiedor deeply explores the development of financial networks on the New York Stock Exchange. He analysis the causal relationships within the markets by using *partial mutual information (*a generalization of *partial correlations)*, which is sensitive to non-linear dependencies. Fiedor compares partial mutual information methods against those networks methods based on correlation, partial correlation, and mutual information; he proves that the use of his network implementation provides different and more accurate networks [9]. Then, the author extends the analysis using transfer entropy, which is a measure that uses time lags to quantify causal information transfer between systems evolving in time, based on conditional transition probabilities. Pawel Fiedor confirms that asynchronous links are rarely based on the sector of economic activity of the connected stocks [10]. We will extend his studies by identifying the behavior of the stocks in each network when information from one of the stocks in the cluster has news release.

References

[1] G. Dionne and X. Zhou and H. Montréal, “High Price Impact Trades in Different Information Environments: Implication for Volatility and Price Efficiency”. [Online] Available: https://www.researchgate.net/publication/340606392 [April 13, 2020]

[2] E. Boehmer and C. Jones and X. Zhang and X. Zhang, “Tracking Retail Investor Activity”. [Online] Available: https://ssrn.com/abstract=2822105 [April 8, 2020]

[3] D. Easley and M. López de Prado and M. O'Hara, “Discerning Information from Trade Data”. *Journal of Financial Economics,* 120(2), pp. 269-286. May 2016; Johnson School Research Paper Series No. 8-2012. [Online] Available: https://ssrn.com/abstract=1989555 [March 1, 2015]

[4] Y. Xie and H. Jiang, “Stock Market Forecasting Based on Text Mining Technology: A Support Vector Machine Method.”. [Online] Available: https://arxiv.org/ftp/arxiv/papers/1909/1909.12789.pdf [July 28, 2016]

[5] Y. Xiaowen and X. Xin and C. Liangliang and K. Hang Sun, “Predicting Market Reactions to Bad News”. [Online] Available: https://ssrn.com/abstract=3144041 [March 19, 2018]

[6] M. Farrell and T. Green and J. Russell and S. Marcov, “The Democratization of Investment Research and the Informativeness of Retail Investor Trading”. [Online] Available: https://ssrn.com/abstract=3222841 [October 15, 2019]

[7] M. López de Prado and R. Rebonato, “Kinetic Component Analysis”. *Journal of Investing,* Vol. 25, No. 3, 2016. [Online] Available: https://ssrn.com/abstract=2422183 [June 5, 2016]

[8] S. Taylor, “Clustering Financial Return Distributions Using the Fisher Information Metric”. [Online] Available: https://ssrn.com/abstract=3182914 [May 22, 2018]

[9] P. Fiedor, “Partial Mutual Information Analysis of Financial Networks” *Acta Physical Polonica A,* vol. 127, no. 3, pp. 863–867. [Online] Available: ResearchGate, https://www.researchgate.net/publication/260678996_Partial_Mutual_Information_Analysis_of_Financial_Networks [March 11, 2014]

[10] P. Fiedor, “Causal Non-Linear Financial Networks”. [Online] Available: ResearchGate, https://www.researchgate.net/publication/264084873_Causal_Non-Linear_Financial_Networks [August 8, 2018]