Empirical Study: Understanding The Relationship Between Financial News and Stock Prices

Omair Shafi Ahmed {shafiahmed.o@northeastern.edu}

Introduction

It is a well accepted fact that financial markets are influenced by news stories and vice versa. However, the relationship between the two has been hard to quantify, empirically. This paper attempts to qualify the relationship between the news articles and performance of the general stock market. We draw on and combine previous methodologies as described in 'News and narratives in Financial systems: exploiting big data for systemic risk assessment'[1] and ' Computer-assisted text analysis methodology in the social sciences'[2] to break down each word into it's associated categories and look for associations and correlations with the market. Further research could be carried out to the effect of quantifying simultaneous causation, if any.

Hypothesis

We aim to quantify the degree to which market prices react to the news. This, in a way, also allows us to quantify the efficiency of the market. According to economic theory, market efficiency is conducive to the optimum allocation of resources[3]. In such circumstances, the price mechanism is thought to ensure that the products and services produced will end up in the hands of those that value them most[4]. The efficient market mechanism is thought to ensure that those who value the products and services are the ones to receive it.

As the Efficient Market Theory[5] hypothesizes that all publicly available information about the market at any point has already been priced into the market value of an asset, any statistically significant association between the news and the market trend would be an evidence for the efficiency of the markets. To this end, we leverage the news dataset and quantify them using the categories defined in the Harvard General Inquirer[6]. We then attempt to measure the response of the dependent variable (the market) to the primary independent variables of interest (the news). Breaking down the news headlines into categories should allow us to capture the underlying sentiment of the headlines without taking into consideration the entire vocabulary of the corpus. Statistical significance in categories should provide evidence for the efficiency of the markets by implying that the presence of words belonging to certain categories is statistically significant to the performance of the market.

Data Sources and Preprocessing

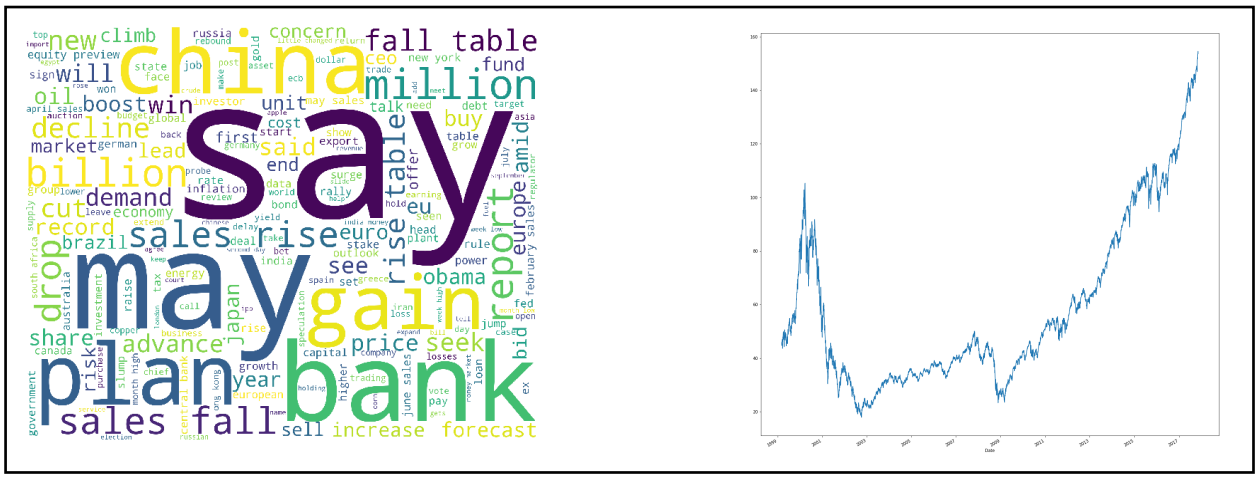

Fig 3: Raw insight into the two datasets. (Left) Word count of the most commonly occurring words denoted by size. (Right) QQQ ETF price from 2006 to 2014.

This research attempts to use financial news from Reuters and Bloomberg from October 2006 to November 2013, released by Ding et al. [2014][7], as well as market data from Kaggle[8] to unearth relationships between the two. The data released by Ding et al., contains 450,341 news from Bloomberg and 109,110 news from Reuters from 2006 to 2013, enough a sample to create a general market sentiment over a reasonable timeframe. The measures of sentiment have been derived from Harvard General Inquirer, which maps each text file with counts on dictionary-supplied 182 categories. The categories on here are specified from four sources: the Harvard IV-4 dictionary, the Lasswell value dictionary, several categories recently constructed, and "marker" categories. The words from the Bloomberg headlines are then mapped to these categories and subsequent statistics are generated and analyzed. The generated statistics are mapped to the Nasdaq 100, via the QQQ ETF to generate correlations and regressions.

Variables and Their Analysis

News Variables:

The news articles and headlines obtained from the kaggle dataset are mapped to the Harvard General Inquired Dictionary based on their semantic categories as shown in Fig 1:

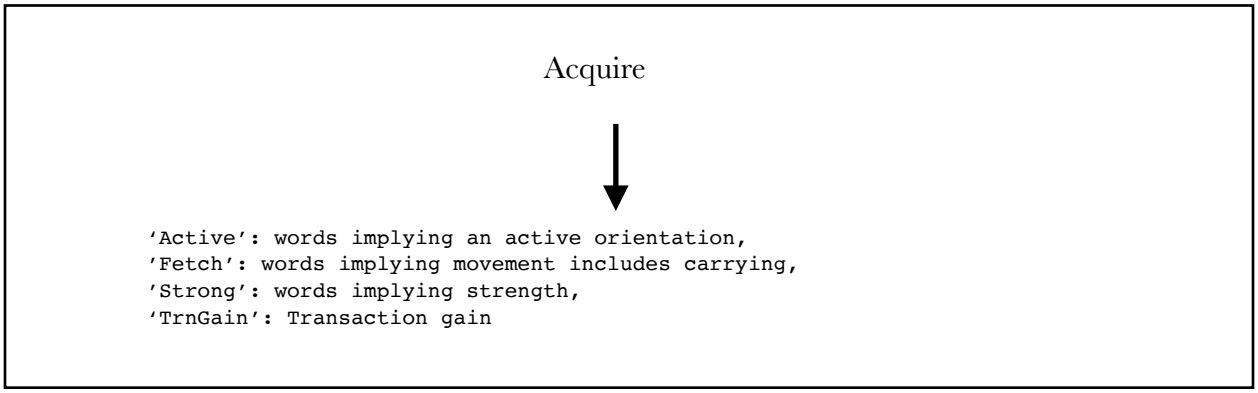

Fig 1: Word converted into Harvard Inquirer Categories and it's Meanings

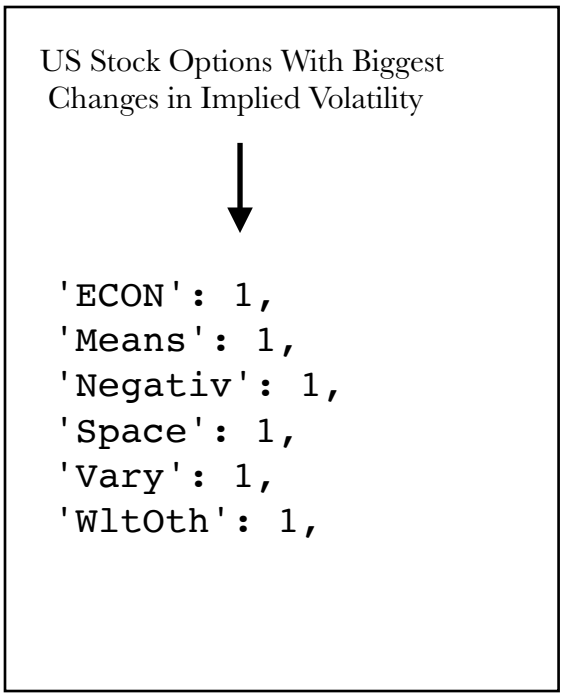

Fig 2: One Headline from 2011-10-06 converted into counts of categories

The category mappings for each word in the headlines are counted. These counts are then used to calculate the mean number of times a category appears in the headlines, per day as shown in Fig 2. The mean values of the category counts per day, we hypothesize should be a reasonable measure of the presence of semantic categories for that day.

PowerShares QQQ Trust:

The PowerShares QQQ, previously known as the QQQQ, is a widely held and traded exchange-traded fund (ETF) that tracks the Nasdaq 100 Index[7]. The Nasdaq 100 Index is composed of 100 of the largest international and domestic companies, excluding financial companies, that are listed on the Nasdaq stock exchange, based on market capitalization. The data for this find included the Open, High, Low and Close and Volume since the beginning of the fund. However, for the purposes of our analysis only data from 2006 with 2013 is used.

As the value of the QQQ ETF on any given day does not hold any latent information about the sentiment or the direction of the market, which is of our primary interest, we need to transform the price data into something that does. To that end, we adapt a function to transform the data of from T-5 days to T days and extract it's slope coefficient. The slope coefficient would be indicative of the positivity or negativity of the trend in the past 5 trading days. The procedure is repeated for High, Open and Low variables.

Preliminary Analysis of the Data:

After converting both the datasets from it's raw values to something that could be ingested by our model, we perform some preliminary analysis of the variables in order to get some insight as well as perform sanity checks.

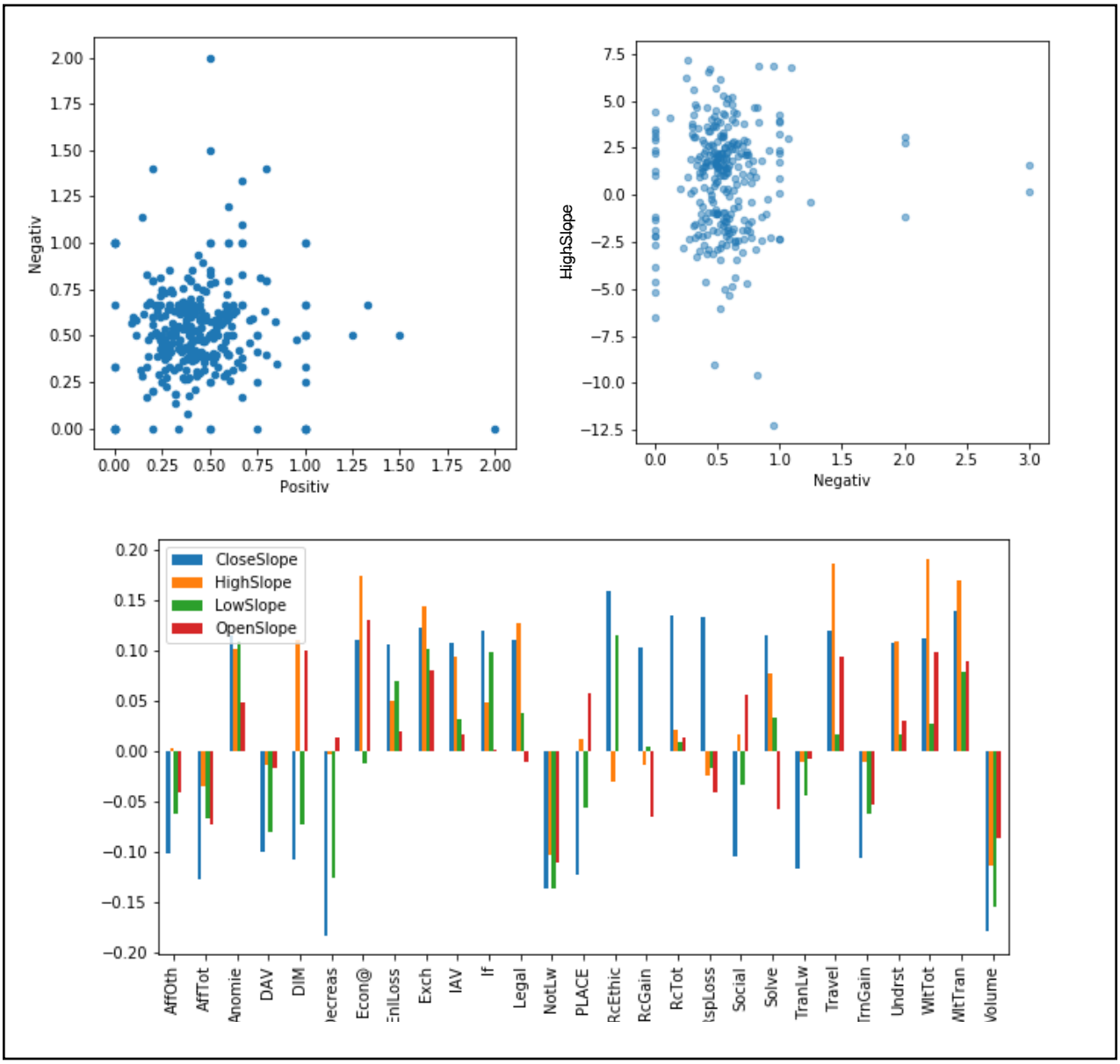

Fig 4: (Top left) Figure is a scatterplot of mean 'positiv' and 'negativ' values in the headlines, for each day. (Top Right) Figure is a scatterplot of mean 'positiv' values and their corresponding slopes for the 'High' variable - HighSlope. (Bottom) Figure is a bar chart of correlations of various categories with slopes for the 'High', 'Low', 'Open' and 'Close' variables

.

The preliminary analysis of the data throws some light into the dynamics of the interactions between news and the performance of the market. While overall correlations are weak, their directions turn out to be as expected. Eyeballing the scatterplots of 'Postiv' vs 'Negativ' values, one can observe a higher negative correlation at extreme values, but almost negligible correlations at lower values. A similar scatterplot of 'HighSlope' vs 'Positiv' seems to show a positive correlation between the two values. However, is is worth noting that the correlations are extremely small.

	Positiv	Negativ	HighSlope	LowSlope
Positiv	1.000000	-0.042278	-0.086336	-0.064992
Negativ	-0.042278	1.000000	0.029277	0.114718
HighSlope	-0.086336	0.029277	1.000000	0.680456
LowSlope	-0.064992	0.114718	0.680456	1.000000

Fig 5: Correlation matrix for the 'Positiv' vs 'Negativ' vs 'High' vs 'Low'.

The difference in the correlations between 'Open', 'High', 'Low', and 'Close' variables also paint an interesting picture as observed in 'Decreas', which pertains to words like 'depreciate' and 'dwindle', has the highest correlation with 'Close' followed by 'Low'. Categories like RspLoss that consists of words that imply 'loss of respect, have positive correlation with slopes of 'Close' and negative correlations with slopes of 'High', 'Low' and 'Open', possibly indicating intra-day volatility.

Model and Interpretations

A least square, multiple regression model was estimated for this analysis. We assume that events as reported by the news, and thereby the mean counts of Harvard Inquirer categories for every headline, per day, to be exogenous variables. The slope of the closing price of the QQQ ETF is assumed to be the endogenous, dependent variable. The final specification resulted after experimentation with various The specification is as follows:

$$\text{CloseSlope} = B0(\text{AffTot}) + B1(\text{PLACE}) + B2(\text{TranLw}) + B3(\text{Social}) + B4(\text{Try}) + B5(\text{WltTran}) + B6(\text{Solve}) + B7(\text{Exch}) + B8(\text{Travel}) + e$$

The regression output generated the coefficients as follows:

$$\text{CloseSlope} = +0.6597^{**} - 2.1658\,(\text{AffTot})^{**} - 1.8595\,(\text{PLACE})^{***} - 1.6358\,(\text{TranLw})^{**} + 2.1652(\text{Social})^{**} - 0.7949\,(\text{Try}) + 5.3562(\text{WltTran})^{***} + 2.5855(\text{Solve})^{***} - 4.4440(\text{Exch})^{***} + 2.2601(\text{Travel})^{***}$$

$R^2 = 0.044$, Adj. $R^2 = 0.034$ and F-Statistic = 4.464,

*** indicates $p < 0.01$ and ** indicates $p < 0.10$

The R^2 being low indicates that only 4.4% of the total variation in the daily price of the QQQ ETF can be explained by the presence of certain words in the headlines of the newspapers. The model holds true for the assumptions of OLS linear regression, while the residuals exhibited are normal and homoskedastic. The input variables do not exhibit multicollinearity, or other regression errors. Additionally, this model was selected after testing various categories for statistical significance. Linear-log, log-log and log-linear models were also tested to improve the R^2 however, they did not seem to reflect the relationship between the underlying data.

As for the hypothesis, there is a statistically significant relationship between some categories and the 5 day slope of the closing prices of the QQQ ETF. Specifically,

- A unit increase in the mean value of AffTot category, which consists of words valuing love and friendship, decreases the slope by 2.16 units.
- A unit increase in the mean value of PLACE category, which consists of words referring to places, locations and routes between them, decreases the slope by 1.85 units.
- A unit increase in the mean value of TranLw category, which consists of words of transaction or exchange in a broad sense, but not necessarily of gain or loss, decreases the slope by 1.63 units.
- A unit increase in the mean value of Social category, which consists of words for created locations that typically provide for social interaction and occupy limited space, increases the slope by 2.16 units.
- A unit increase in the mean value of WltTran category, which consists of words words for pursuit of wealth, such as buying and selling, increases the slope by 5.35 units.
- A unit increase in the mean value of Solve category, which consists of words referring to the mental processes associated with problem solving, increases the slope by 2.58 units.
- A unit increase in the mean value of Exch category, which consists of words Words concerned with buying, selling and trading, decreases the slope by 4.44 units.
- A unit increase in the mean value of Travel category, which consists of words Words for all physical movement and travel from one place to another in a horizontal plane, increases the slope by 2.26 units.
- The 0.65 positive intercept suggests a positive slope, which is in line with the positive returns the ETF generated over the time period.

Summary

While the R^2 of our model has been disappointingly small, the large vocabulary of the English language along with the possibility of semantic variation in any given set of words means that the predictive power of our model couldn't have been high without using more advanced NLP techniques. Another shortcoming of this study is the fact that the study only considered headlines, without considering the body of the news. Including the body of the news for our analysis would have given us enough data for the mean values of the categories to be more representative of the underlying sentiment and therefore more robust, statistically. Moreover, we only considered the QQQ ETF which excludes the financial sector from the stocks it tracks. Using a more inclusive ETF might have been a more accurate indicator of financial markets.

Having said that, this study did unravel statistically significant categories of words that appear to have an impact on the financial markets. While a lot of the categories and their effect on the markets make intuitive sense, there are some that don't and warrant investigation into linguistics and behavioral psychology. Categories that relate to love and friendship, locations places and routes along with buying, selling, trading, transaction or exchange all affect the dependent variable - the market - negatively, which defies the general intuitive understanding that everything positive is positive for the market.

References

1. Rickard Nyman, David Gregory, Sujit Kapadia, Paul Ormerod , David Tuckett & Robert Smith. News and narratives in financial systems: Exploiting big data for systemic risk assessment1. September 2016
2. Melina Alexa, Computer-assisted text analysis methodology in the social sciences, October 1997
3. Giannoni, Marc P. and Michael Woodford, "Optimal Interest-Rate Rules: I. General Theory," NBER working paper no. 9419, December 2002.
4. Daniel Levy, "Price rigidity and flexibility: new empirical evidence", 10 October 2007
5. MALKIEL, B. G., The Efficient Market Hypothesis and Its Critics, 2003
6. 4-5. Philip J. Stone, Dexter C. Dunphy, Marshall S. Smith, Daniel M. Ogilvie, and associates. The General Inquirer: A Computer Approach to Content Analysis, MIT Press, 1966.
7. [Ding et al., 2014] Xiao Ding, Yue Zhang, Ting Liu, and Junwen Duan. Using structured events to predict stock price movement: An empirical investigation. In Proc. of EMNLP, pages 1415–1425, Doha, Qatar, October 2014. Association for Computational Linguistics.
8. Boris Marjanovic. https://www.kaggle.com/borismarjanovic/price-volume-data-for-all-us-stocks-etfs. Nov. 2017

Appendix A

OLS Regression Results

Dep. Variable:	CloseSlope	R-squared:	0.070
Model:	OLS	Adj. R-squared:	0.046
Method:	Least Squares	F-statistic:	2.916
Date:	Tue, 17 Apr 2018	Prob (F-statistic):	9.22e-06
Time:	17:22:19	Log-Likelihood:	-2280.8
No. Observations:	880	AIC:	4608.
Df Residuals:	857	BIC:	4718.
Df Model:	22		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	0.3997	0.373	1.072	0.284	-0.332	1.131
Decreas	0.8626	1.219	0.708	0.479	-1.530	3.255
NotLw	-4.3538	4.181	-1.041	0.298	-12.560	3.852
AffTot	-2.7944	1.083	-2.580	0.010	-4.920	-0.668
PLACE	-2.3445	0.753	-3.113	0.002	-3.823	-0.867
TranLw	-2.1199	0.886	-2.392	0.017	-3.859	-0.381
DIM	-4.7590	2.117	-2.247	0.025	-8.915	-0.603
POS	1.9173	1.153	1.662	0.097	-0.346	4.181
TrnGain	1.8230	1.115	1.635	0.102	-0.365	4.011
Social	2.7795	1.083	2.566	0.010	0.653	4.905
Try	-1.1001	1.022	-1.076	0.282	-3.106	0.906
PowCon	1.2319	1.072	1.149	0.251	-0.872	3.335
RcEthic	-3.0342	2.629	-1.154	0.249	-8.193	2.125
WltTran	5.3482	1.585	3.373	0.001	2.237	8.460
Solve	2.1993	0.848	2.593	0.010	0.534	3.864
RcGain	5.8412	5.553	1.052	0.293	-5.059	16.741
RcTot	3.5260	2.046	1.724	0.085	-0.489	7.541
RspLoss	1.4761	4.400	0.335	0.737	-7.160	10.112
If	1.6732	0.919	1.820	0.069	-0.131	3.477
Exch	-4.9213	1.557	-3.160	0.002	-7.978	-1.865
Travel	2.2211	0.791	2.807	0.005	0.668	3.774
Anomie	-5.5097	17.646	-0.312	0.755	-40.143	29.124
Begin	-1.4760	1.639	-0.901	0.368	-4.692	1.740

Omnibus:	59.709	Durbin-Watson:	0.936
Prob(Omnibus):	0.000	Jarque-Bera (JB):	206.628
Skew:	-0.229	Prob(JB):	1.35e-45
Kurtosis:	5.329	Cond. No.	187.