Using Machine Learning and Alternative Data to Predict Movements in Market Risk

Thomas Dierckx
KU Leuven

Using Machine Learning and Alternative Data to Predict Movements in Market Risk

Thomas Dierckx
Department of Statistics
Department of Computer Science
KU Leuven, Belgium
thomas.dierckx@cs.kuleuven.be

Jesse Davis
Department of Computer Science
KU Leuven, Belgium
jesse.davis@cs.kuleuven.be

Wim Schoutens
Department of Statistics
KU Leuven, Belgium
wim.schoutens@kuleuven.be

September 18, 2020

Abstract

Using machine learning and alternative data for the prediction of financial markets has been a popular topic in recent years. Many financial variables such as stock price, historical volatility and trade volume have already been through extensive investigation. Remarkably, we found no existing research on the prediction of an asset's market implied volatility within this context. This forward-looking measure gauges the sentiment on the future volatility of an asset, and is deemed one of the most important parameters in the world of derivatives. The ability to predict this statistic may therefore provide a competitive edge to practitioners of market making and asset management alike. Consequently, in this paper we investigate Google News statistics and Wikipedia site traffic as alternative data sources to quantitative market data and consider Logistic Regression, Support Vector Machines and AdaBoost as machine learning models. We show that movements in market implied volatility can indeed be predicted through the help of machine learning techniques. Although the employed alternative data appears to not enhance predictive accuracy, we reveal preliminary evidence of non-linear relationships between features obtained from Wikipedia page traffic and movements in market implied volatility.

***K*eywords** Implied Volatility · Machine Learning · Alternative Data

1 Introduction

To predict the stock market, statisticians generally relied on econometric methods. However, recent literature [1] shows that machine learning models typically outperform statistical and econometric models ([2, 3, 4, 5, 6]). Machine learning models typically provide more flexibility by requiring less distributional assumptions about the data [6]. Moreover, they are able to recognize patterns in time series data more easily [4] and can even be combined to reduce over-fitting and further improve performance [5]. In addition, with the recent paradigm shift in machine learning towards deep learning, researchers can extract features and model non-linear correlations without relying on econometric assumptions and human expertise [7]. Selecting an appropriate model alone is never sufficient since predictive performance will significantly be determined by the employed data. Most studies have traditionally focused on quantitative market data (e.g. [8, 9, 10, 11, 12, 13, 14]). Interestingly, since the inception of Web 2.0, research has been exploring numerous alternative data sources to improve predictive accuracy. These novel sources, such as social media and online news, exert a positive impact on information diffusion and thus affect the market and its investors in various ways. Although it is not clear whether we are rather emotional than rational beings, it is not inconceivable that investors can act emotionally and are influenced by these alternative data sources. This is in line with proponents of behavioural science, where investment decisions are influenced by investor sentiment ([15, 16, 17]). Moreover, studies have already shown that investors often exhibit herd behaviour ([18, 19, 20]). Consequently, fueled by advancements in Natural Language Processing (NLP), the last few years have seen an explosion in research on quantifying the synergy between online media and stock markets. On top of that, due to the surge in popularity of certain data analysis techniques and increasing computational power, it has never been easier to examine vast amounts of features for more accurate predictions.

A wealth of financial market variables have already been subject to prediction attempts by research using both machine learning and alternative data. Most research focuses on stock price prediction, seen as the *holy grail* within finance (e.g. [21, 21, 22, 23, 24, 2]). However, other popular market variables such as historical volatility ([25, 26, 27, 28] and trade volume [26] have also been considered. Remarkably, we found no research that explored the use of machine learning, not to mention alternative data sources, to predict the market implied volatility of assets. Derived from option prices, this variable is deemed to be one of the more important parameters in the world of derivatives. In contrast to historical volatility, this forward-looking measure indicates how much risk the market *expects* a certain asset to exhibit in the coming period. As this variable is paramount for the pricing of options, the ability to predict its movements would be advantageous for the practice of asset management and market making alike. Consequently, the research focus in this paper is twofold. First we investigate whether we can use machine learning algorithms to predict if an asset's market implied volatility will have moved up or down by the end of the next trading day. Second, we examine whether the introduction of alternative data sources has any positive impact on predictive accuracy.

2 Preliminaries

Using historical quantitative market data for stock prediction has been around for decades. Interestingly, since the inception of Web 2.0, the search for predictive information from alternative data sources has expanded notably. A tremendous amount of research has been published on quantifying the interplay between online media and stock markets. As a short literature study, Section 2.1 rehashes the most important recent findings based on recent surveys ([29, 30, 31, 32]). Subsequently, Section 2.2 elaborates on the concept of market implied volatility.

2.1 The interplay between the stock market and alternative data

A variety of alternative sources have been considered to aid in stock market prediction thus far [31, 32]. Most studies have focused on Twitter because of its popularity and easy accessibility (e.g. [33, 34]), but also discussion boards (e.g. [35, 36]) and news sources such as The Wall Street Journal, Financial Times and Thomson Reuters (e.g. [37, 38]) have been considered. Corporate disclosures and financial reports are published only a few times a year and have received somewhat less attention, but are gaining momentum (e.g. [39, 40]). Moreover, surveys [29] and [32] note that Google Trends and Wikipedia page views are also being considered as predictive sources in recent years (e.g. [2, 1]]). Most of the aforementioned sources are mostly comprised of textual data rich in information. In order for this type of content to be useful for machine learning algorithms, transformation into a machine interpretable form, such as a scalar or tensor, is necessary. However, extracting all relevant information contained in text remains a challenging research topic to this day [29].

Over a decade worth of research is evidently accompanied by a wealth of findings. Many data sources have been investigated for their predictive power, but one that stands out in particular is Twitter. This platform, due to its popularity, accessibility and compact content is significantly more investigated than any other alternative data source. Table 1 presents the main findings on Twitter and shows empirical evidence for its predictive power. However, not all research was in favor for the predictive power of tweets. For example, [41] only found a weak statistical correlation between Twitter sentiment and the S&P500's closing price. Overall, [30] found 28 papers supporting the predictive power of Twitter sentiment, and 5 papers with mixed or non-supportive results. This suggests that there is indeed at least some supportive evidence for the predictive power of investor sentiment in the context of capital markets. Moreover, [42] discovered that social media had a stronger relationship with stock performance than any other alternative data source, and that the impact of different types of sources varied widely. Table 2 presents the most important findings from studies using other types of alternative data such as discussion boards, news articles, Wikipedia or Google Trends as predictors. Although discussion boards and news articles should also contain an abundance of information, there is significantly less research published.

Once interesting data sources are identified and processed into (hopefully predictive) features, a predictive model needs to be fitted to the data. [31] notes that Regression and Support Vector Machines have been popular methods in the past decade. However, with the recent paradigm shift in machine learning towards deep learning, researchers are starting to favor deep learning methods. Although these methods are prone to overfitting, they can extract features and model non-linear correlations without relying on econometric assumptions and human expertise [7]. Note that deep learning methods generally work best when there is an abundance of training examples available, a characteristic often not present in the financial domain. Which category of algorithms is ultimately most appropriate is still considered an open question [31].

Table 1: Key Twitter findings throughout the years retrieved from surveys [29] and [30].

Finding	Year	Reference
The degree to which companies are jointly mentioned is correlated to the co-movement of these stocks.	2011	[43]
Emotions such as "hope", "fear" and "worry" in tweets are negatively related to stock market indexes and positively correlated to the volatility index .	2011	[44]
Twitter sentiment can be predictive for intraday exchange rates.	2013	[45]
Sentiment polarity extracted from users with many followers is associated with abnormal returns on the same day for S&P 500 stocks.	2014	[46]
Twitter sentiment affects abnormal returns during peaks of Twitter volume.	2015	[47]
Existence of a nonlinear causal relationship between Twitter investment on stock returns on DJIA.	2015	[48]
Only negative emotion has significant impact on stock returns.	2015	[49]
Similar firms, clustered based on firm-specific microblogging metrics, have higher co-movements than those in the same industry.	2015	[50]
Direct relationship between an IPO its Twitter sentiment and first trading day returns.	2016	[51]
Twitter sentiment can be used to predict the price trend of GOOGL, AAPL and FB. In addition, tweet volume has a strong impact on both price and trend.	2016	[52]
Daily bullish percentage extracted from Twitter helps explain excess returns even when the traditional factors used in asset pricing models are considered.	2017	[53]
Volatility sentiment on social media contains information regarding future stock volatility.	2017	[54]

Table 2: Key findings from using discussion boards, news articles, Wikipedia and Google Trends based on [29] and [32].

Finding	Year	Reference
Fraction of negative words in firm-specific news predicts low firm earnings	2008	[55]
Stocks with no media coverage earned higher returns	2009	[56]
Connection between sentiment of forum posts and stock indices, volumes and volatility.	2007	[36]
Page views of articles on Wikipedia relating to companies or financial topics increases before stock market declines.	2013	[57]
Google search volumes for keywords related to financial market increases before stock market declines.	2013	[58]
Movements in financial markets and movements in financial news are intrinsically interlinked	2013	[59]
Opinions expressed on Seeking Alpha have significant stock returns forecasting ability	2014	[60]

2.2 Market Implied Volatility

In the world of derivatives, options are one of the most prominent types of financial instruments available. As sellers of options are exposed to risk for the duration of the contract, they want to be properly compensated. To measure this risk, the expected price fluctuations of the underlying asset are considered over the course of the option contract. This measure is better known as implied volatility and varies with the strike price and duration of the option contract.

Interestingly, when considering the selection of implied volatilities for different strike prices and fixed duration, a more general measure called market implied volatility can be extracted. A famous example of this generalized measure is the CBOE Volatility Index, better known as the VIX.

VIX is a measure of expected price fluctuations in the S&P 500 Index options over the next 30 days. It's famously known as the *fear index* and is considered a reflection of investor sentiment on the market. Interestingly, the calculation for the VIX for term T (Equation 1 - taken from the VIX white paper [61]) can be applied to any asset with available options. Although this measure can be calculated for any arbitrary term, the duration of the option contracts will seldom match. To overcome this obstacle, the market implied volatility is calculated for the option contracts expiring right before and after the desired target date. We then linearly interpolate between these two measures to correspond with the desired term, as outlined in [61]. Formally, VIX is defined as:

$$VIX = 100 \times \sqrt{\frac{2}{T}\sum_i \frac{\Delta K_i}{K_i^2} e^{RT} Q(K_i) - \frac{1}{T}\left[\frac{F}{K_0} - 1\right]^2} \tag{1}$$

where:

T is time to expiration

F is the forward index level derived from the index option prices

K_0 is the first strike below the forward index level F

K_i is the first strike price of the i^{th} out-of-the-money option; a call if $K_i > K_0$ and a put if $K_i < K_0$; both put and call if $K_i = K_0$

ΔK_i is the interval between strike prices

R is the risk-free rate to expiration

$Q(K_i)$ is the midpoint of the bid-ask spread for each option with strike K_i

3 Methods

Recent research focusing on stock price prediction using machine learning and alternative data sources has reported some considerable successes ([1, 2, 62, 21, 22, 23, 24, 63]). Motivated by the novelty of our research, and the promising nature of alternative data, we investigated whether we can predict if an asset's market implied volatility will have moved up or down by the end of the next trading day. In addition to market data, we employed Wikipedia page traffic and Google News statistics as alternative data. Section 3.1 describes in detail how we performed our data acquisition and constructed additional features. Section 3.2 then details for which machine learning algorithms we opted, and how we preprocessed our data in order to maximize performance.

3.1 Data acquisition and feature generation

In this paper, we focus on predicting whether AAPL's market implied volatility at the end of the next day will have moved up or down. We considered a 24 month trading period from January 1, 2016 till December 31, 2017 for which three different data sources were utilized. First, we collected publicly available market data from Yahoo Finance including the stock's daily trading volume and daily opening, high, low and closing prices. We obtained the stock's historical market implied volatility by applying the VIX formula (Equation 1) on personal end-of-day option data. The second data source is Google News, a source that aggregates news and blog posts relevant to provided keywords. Google News allows for queries on how many publications containing certain keywords were made on any given day. We scraped this information and built a time series of daily news counts for AAPL. The third and last data source is Wikipedia, which provides daily visitor statistics per page. We retrieved this data via the Wikimedia API, and built a time series of daily page views for the Apple Inc page. In total, eight different features are obtained for each trading day from three different types of data sources (Table 4).

Table 3: Summary of considered data sources and corresponding original features (daily granularity).

Market Data	Option Data	Google News	Wikipedia
OHCL Volume	Market Implied Volatility	News Counts	Page Views

Table 4: Summary of considered data sources and corresponding original features (daily granularity).

Market Data	Google News	Wikipedia
OHCL Volume Market Implied Volatility	News Counts	Page Views

Influenced by the work done in [1] and [2], we employed a set of technical analysis tools with varying parameters to generate additional predictors on the aforementioned features. Table 5 summarizes our process, where each part of the table is used on a different subset of the eight original features. The techniques listed in the top part (1) are applied only on *market implied volatility*, *news counts* and *page views* yielding 60 new features. The techniques in the middle (2) are applied to the previous features and *close price* yielding eight new features. The last part is solely applied on *close price*, yielding two new features. Note that we replicate the feature generation strategy from [2], and that daily volume, daily high and daily low are not used for additional feature generation. In total 78 features are obtained for each trading day, yielding 486 feature vectors in total.

Table 5: Summary of selected feature generation techniques with corresponding parameters. The top portion is used on market implied volatility, page views and news count features. The middle is used on the previous features and close price. The last portion is solely used on close price [2].

	Technique	Parameters
(1)	Moving Average (MA)	$n \in \{3, 5, 10\}$
	Moving Average Move (MA_Move)	$n \in \{3, 5, 10\}$
	Exponential Moving Average (EMA)	$n \in \{3, 5, 10\}$
	Exponential Moving Average Move (EMA_Move)	$n \in \{3, 5, 10\}$
	Rate Of Change (ROC)	$n \in \{5\}$
	Rate Of Change Move (ROC_Move)	$n \in \{5\}$
	Disparity Index	$n \in \{3, 5\}$
	Disparity Index Move	$n \in \{3, 5\}$
	Momentum1 [1]	$n \in \{5\}$
	Momentum2 [2]	$n \in \{5\}$
(2)	Relative Strength Index (RSI)	$n \in \{14\}$
	Relative Strength Index Move (RSI_Move)	$n \in \{14\}$
(3)	Williams %R	$n \in \{14\}$
	Stochastic Oscillator [3]	$n \in \{14\}$

To construct the target feature, we define the next day's difference in market implied volatility on day i as $y_i^* = (ivolatility_{i+1} - ivolatility_i)$ where $ivolatility_i$ denotes the end of day market implied volatility on day i. We consider a move to be upwards whenever $y_i^* > 0$ and downwards whenever $y_i^* \leq 0$. The final target feature is therefore a binary feature obtained by applying Case Equation 2.

$$y_i = \begin{cases} 1, & \text{if } y_i^* > 0. \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

3.2 Machine Learning

We examined the effectiveness of Logistic Regression, Support Vector Machines (SVM) and Boosting Machines for predicting movements in market implied volatility on the AAPL stock. Although Logistic Regression is a linear model, its speed and interpretability makes it warranted to try [64]. Support Vector Machines have been a popular technique for stock market prediction using alternative data sources (e.g. [2, 1, 62, 65]). Trained with the non-linear Radial Basis Function, SVMs are able to fit to both linear and non-linear separable data. Lastly, Boosting Machines such

[1] See [2] for details on the momentum1 and momentum2 calculations.

[2] See footnote 1.

[3] Stochastic Oscillator by George Lane

as AdaBoost and XGBoost have surged in popularity last few years and have also been employed in stock market prediction using alternative data sources [1]. Note that when no parameters are mentioned in this or subsequent sections, the standard Python Sklearn library parameter configurations apply.

Feeding raw data to machine learning algorithms is seldom a good idea. One of the main problems when dealing with time series data is the presence of seasonality and trends. We therefore identify non-stationary features with the Augmented Dickey-Fuller test and transform said features into features representing their first difference. Some machine learning algorithms, such as Logistic Regression and Support Vector Machines, also require features to be standardized. We therefore standardized all features to have zero mean and unit variance before using said models.

To combat the abundance of features and the associated curse of dimensionality, we used a smaller subset of 29 features for training the Logistic Regression and SVM models. We obtained this smaller subset by fitting a standard AdaBoost model to the data. Tree-based methods like AdaBoost produce a list of relative variable importances, from which we extracted features with a higher than average importance. Aside from using AdaBoost as feature selection tool, we also used the model for prediction. Because forest classifiers such as AdaBoost are robust to large feature spaces and scaling issues, we do not perform standardization or feature selection prior to using this model for classification.

Note that we consciously refrained from using deep learning techniques due to the lack of data points. Random Forests were also not considered because the algorithm's random sampling clashes with the sequential nature of the given data.

3.3 Evaluation

The built models are evaluated using cross validation, where data is repeatedly split into non-overlapping train and test sets. This way models are trained on one set, and afterwards tested on a test set comprised of unseen data to give a more robust estimate of the achieved generalization. However, special care needs to be taken when dealing with time series data. Classical cross validation methods assume observations to be independent. This assumption does not hold for time series data, which inherently contains temporal dependencies among observations. We therefore split the data into training and test sets taking temporal order into account to avoid data-leakage. More concrete, we employ Walk Forward Validation (or Rolling Window Analysis) where a sliding window of t previous trading days is used to train the models, and where t_{t+1} is used for the out-of-sample test prediction. Table 6 shows an example of this method where t_i denotes the feature vector corresponding to trading day i. Note that in this scenario, when given a total of n observations and a sliding window of length t, you can construct a maximum of $n - t$ different train-test splits.

Table 6: Example of Walk Forward Validation where t_i represents the feature vector of trading day i. In this example, a sliding window of size three is taken. We therefore consistently use the feature vectors of the previous three trading days to train a model (underlined) and subsequently test said model on the fourth day (bold).

Iteration	Variable roles
1	$\underline{t_1\ t_2\ t_3}\ \mathbf{t_4}\ t_5\ \cdots\ t_n$
2	$t_1\ \underline{t_2\ t_3\ t_4}\ \mathbf{t_5}\ \cdots\ t_n$
⋮	⋮
m	$t_1\ \cdots\ \underline{t_{n-3}\ t_{n-2}\ t_{n-1}}\ \mathbf{t_n}$

As with any cross validation method, models have to be retrained during each iteration of the evaluation process. Note that in this case, we also have to standardize and perform feature selection again each iteration as to not introduce look-ahead bias.

4 Experimental results and discussion

In this section we present our experiment methodology and findings from our study. First we investigated whether we can use machine learning algorithms to predict if an asset's market implied volatility will have moved up or down by the end of the next trading day. Second, we examined whether the introduction of alternative data sources has any positive impact on predictive accuracy. To this end, we conducted an ablation study where the effectiveness of each of the different sources is investigated. Five different scenarios were considered in total, shown in Table 7. In each scenario, only original and generated features of the listed data sources were considered for prediction.

Evaluation is done on a temporal ordered dataset of 486 feature vectors each corresponding to a different trading day. For each of the different scenarios (Table 7), we evaluate the models on 106 different train-test splits obtained using Walk Forward Validation (Section 3.3). We employed a sliding window of 379 trading days (78% of total available

Table 7: Different scenarios in which only features of listed sources are considered. This includes original and generated features.

Scenario	Description
1	Market data
2	Google News counts, Wikipedia traffic
3	Market data, Wikipedia traffic
4	Market data, Google News counts
5	Market data, Wikipedia traffic, Google News counts

data), where the next day was used as the out-of-sample test case. We therefore made 106 consecutive out-of-sample movement predictions for AAPL's market implied volatility. Table 8 shows the class distribution for up and down movements during this period.

Table 8: Class distribution of down and up movements over the testing period of 106 days.

Dependent Class Distribution	
Down Movement	Up Movement
59%	41%

The results from our ablation study using Logistic Regression, a rbf-kernel SVM and AdaBoost are presented in Table 9. Note that Table 8 suggests we are dealing with a relatively imbalanced class distribution. We therefore considered balanced accuracy, which is defined as the average of recall obtained on each class, as evaluation metric instead of raw predictive accuracy.

Table 9: Summary of obtained balanced accuracy for different models and data source scenarios. Note that the models used were standard configurations from Sklearn's Python library.

Scenario	Logistic Regression	SVM	AdaBoost
1	**63.3%**	**64.2%**	53.7%
2	52.3%	50.5%	55.0%
3	52.8%	55.6%	**63.0%**
4	51.8%	53.3%	49.0%
5	61.3%	59.1%	57.3%

Remarkably, using only features from market data (Table 3.1) yielded the best results for both Logistic Regression and the SVM. This is in contrast with work in [1] and [2], where alternative data was found to have a positive impact on stock price movement prediction. However, concluding that alternative data sources contain no useful information for implied volatility movement prediction is not entirely accurate. For example, the Adaboost classifier performed best using market and Wikipedia features. Logistic Regression performed poorly for this scenario, suggesting there is at least some non-linear relationship between Wikipedia features and movements in implied volatility that the AdaBoost classifier was able to find. Google News count appears to be the least effective data source in our experiments. Although combined with market and Wikipedia features, all models achieved their second best score.

Despite the fact that we have shown AAPL's implied volatility movements can be predicted to a certain extent, the results are somewhat disappointing for the case of alternative data sources. It would be interesting to verify this behaviour on bigger datasets comprised of more trading days, as the vast amount of features might introduce too much noise for the models to effectively generalize on the relatively small training sets.

5 Conclusion and Future Research

We have shown that we can predict, to a certain extent, whether AAPL's market implied volatility at the end of next day will have moved up or down. The best results were obtained by only using market data, excluding Google News counts and Wikipedia page traffic. However, the results in Table 7 suggests there are predictive relationships to be found in these alternative data sources. How to effectively exploit these relationships is an open question.

One of the biggest hurdles for this research was the availability of options data. We only had two years of end-of-day options data, which greatly constrained the time span for our research. Given the abundance of available features, it would be interesting to further investigate this topic on larger datasets. It also constrained us in our potential prediction targets. As our options data only consisted out of end-of-day statistics, we could only investigate movements from day end to day end. Lastly, there's a wealth of other alternative data sources left untapped (see Section 2.1) that might further increase predictive accuracy. However, exploiting these alternative data sources is mostly a labour intensive task.

Aside from increasing efforts on feature engineering by generating smarter and more features, target construction can be improved. We only considered binary prediction where any small increment or decrement represented a move. The introduction of another label representing neutral moves is therefore definitely something worth investigating. Lastly, we didn't fully exploit the capabilities of the employed models. As we didn't perform parameter optimization for the employed machine learning algorithms, there might be room for improvement. In addition, all three models assign probabilities to the labels they predict. It would therefore be interesting to investigate whether higher probabilities are correlated with precision and movement magnitude.

References

[1] Bin Weng, Lin Lu, Xing Wang, Fadel M. Megahed, and Waldyn Martinez. Predicting short-term stock prices using ensemble methods and online data sources. *Expert Systems With Applications*, 112:258–273, 2018.

[2] Bin Weng, Mohamed A. Ahmed, and Fadel M. Megahed. Stock market one-day ahead movement prediction using disparate data sources. *Expert Systems With Applications*, 79:153–163, 2017.

[3] Ming-Wei Hsu, Stefan Lessmann, Ming-Chien Sung, Tiejun Ma, and Johnnie E.V. Johnson. Bridging the divide in financial market forecasting: machine learners vs. financial economists. *Expert Systems with Applications*, 61:215 – 234, 2016.

[4] P. Meesad and R. I. Rasel. Predicting stock market price using support vector regression. In *2013 International Conference on Informatics, Electronics and Vision (ICIEV)*, pages 1–6, 2013.

[5] Jigar Patel, Sahil Shah, Priyank Thakkar, and K Kotecha. Predicting stock market index using fusion of machine learning techniques. *Expert Systems with Applications*, 42(4):2162 – 2172, 2015.

[6] Yudong Zhang and Lenan Wu. Stock market prediction of s&p 500 via combination of improved bco approach and bp neural network. *Expert Systems with Applications*, 36(5):8849 – 8854, 2009.

[7] Xinxin Jiang, Shirui Pan, Jing Jiang, and Guodong Long. Cross-domain deep learning approach for multiple financial market prediction. *International Joint Conference on Neural Networks (IJCNN)*, 2018.

[8] J. C. Patra, N. C. Thanh, and P. K. Meher. Computationally efficient flann-based intelligent stock price prediction system. In *2009 International Joint Conference on Neural Networks*, pages 2431–2438, June 2009.

[9] E. W. Saad, D. V. Prokhorov, and D. C. Wunsch. Comparative study of stock trend prediction using time delay, recurrent and probabilistic neural networks. *IEEE Transactions on Neural Networks*, 9(6):1456–1470, Nov 1998.

[10] J. Roman and A. Jameel. Backpropagation and recurrent neural networks in financial analysis of multiple stock market returns. In *Proceedings of HICSS-29: 29th Hawaii International Conference on System Sciences*, volume 2, pages 454–460 vol.2, Jan 1996.

[11] Hengjian Jia. Investigation into the effectiveness of long short term memory networks for stock price prediction. *CoRR*, abs/1603.07893, 2016.

[12] Wen-Chyuan Chiang, David Enke, Tong Wu, and Renzhong Wang. An adaptive stock index trading decision support system. *Expert Systems with Applications*, 59:195–207, 2016.

[13] Eunsuk Chong, Chulwoo Han, and Frank Chongwoo Park. Deep learning networks for stock market analysis and prediction: Methodology, data representations, and case studies. *Expert Systems with Applications*, 83:187–205, 2017.

[14] Nikitas Goumatianos, Ioannis T. Christou, Peter Lindgren, and Ramjee Prasad. An algorithmic framework for frequent intraday pattern recognition and exploitation in forex market. *Knowledge and Information Systems*, 53:767–804, 2017.

[15] Malcolm Baker and Jeffrey Wurgler. Investor sentiment and the cross-section of stock returns. *The Journal of Finance*, 61(4):1645–1680, 2006.

[16] Marilyn Clark-Murphy and Geoffrey Soutar. Individual investor preferences: A segmentation analysis. *Journal of Behavioral Finance*, 6(1):6–14, 2005.

[17] Muhammad Zubair Tauni, Hong Xing Fang, and Amjad Iqbal. The role of financial advice and word-of-mouth communication on the association between investor personality and stock trading behavior: Evidence from chinese stock market. *Personality and Individual Differences*, 108:55 – 65, 2017.

[18] David S. Scharfstein and Jeremy C. Stein. Herd behavior and investment. volume 80, pages 465–479, 1990.

[19] Eric C Chang, Joseph W Cheng, and Ajay Khorana. An examination of herd behavior in equity markets: An international perspective. *Journal of Banking & Finance*, 24(10):1651 – 1679, 2000.

[20] Zoran Ivković and Scott Weisbenner. Information diffusion effects in individual investors' common stock purchases: Covet thy neighbors' investment choices. *The Review of Financial Studies*, 20(4):1327–1357, 2007.

[21] Andrius Mudinas, Dell Zhang, and Mark Levene. Market trend prediction using sentiment analysis: Lessons learned and paths forward. *WSDM*, 2018.

[22] Yefeng Ruan, Arjan Durresi, and Lina Alfantoukh. Using twitter trust network for stock market analysis. *Knowledge-Based Systems*, 145:207–218, 2018.

[23] Qili Wang, Wei Xu, and Han Zheng. Combining the wisdom of crowds and technical analysis for financial market prediction using deep random subspace ensembles. *Neurocomputing*, 299:51–61, 2018.

[24] Ziniu Hu, Weiqing Liu, Jiang Bian, Xuanzhe Liu, and Tie-Yan Liu. Listening to chaotic whispers: A deep learning framework for news-oriented stock trend prediction. *WSDM*, 2018.

[25] Adam Atkins, Mahesan Niranjan, and Enrico Gerding. Financial news predicts stock market volatility better than close price. *The Journal of Finance and Data Science*, 4(2):120 – 137, 2018.

[26] Nuno Oliveira, Paulo Cortez, and Nelson Areal. The impact of microblogging data for stock market prediction: using twitter to predict returns, volatility, trading volume and survey sentiment indices. *Expert systems with applications*, 73:125–144, 2017.

[27] Navid Rekabsaz, Mihai Lupu, Artem Baklanov, Alexander Dür, Linda Andersson, and Allan Hanbury. Volatility prediction using financial disclosures sentiments with word embedding-based ir models. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1712–1721. Association for Computational Linguistics, 2017.

[28] Ha Young Kim and Chang Hyun Won. Forecasting the volatility of stock price index: A hybrid model integrating lstm with multiple garch-type models. *Expert Systems with Applications*, 103:25 – 37, 2018.

[29] Qing Li, Yan Chen, Jun Wang, Yuanzhu Chen, and Hsinchun Chen. Web media and stock markets: A survey and future directions from a big data perspective. *IEEE Transaction On Knowledge And Data Engineering*, 30(2), 2017.

[30] Heba Ali. Twitter, investor sentiment and capital markets: What do we know? *International Journnal of Economics and Finance*, 10(8), 2018.

[31] Frank Z. Xing, Erik Cambria, and Roy E. Welsch. Natural language based financial forecasting: a survey. *Artificial Intelligence Review*, 50(1):49–73, 2018.

[32] Shweta Agarwal, Shailendra Kumar, and Utkarsh Goel. Stock market response to information diffusion through internet sources: A literature review. *International Journal of Information Management*, 45:118 – 131, 2019.

[33] Johan Bollen, Huina Mao, and Xiaojun Zeng. The impact of social and conventional media on firm equity value: A sentiment analysis approach. *Journal of Computational Science*, 2(1):1–8, 2011.

[34] Jianfeng Si, Arjun Mukherjee, Bing Liu, Qing Li, Huayi Li, and Xiaotie Deng. Exploiting topic based twitter sentiment for stock prediction. In *ACL*, 2013.

[35] Julien Velcin Thien Hai Nguyen, Kiyoaki Shirai. Sentiment analysis on social media for stock movement prediction. *Big Expert Systems with Applications*, 42(24):9603–9611, 2015.

[36] Sanjiv R. Das and Mike Y. Chen. Yahoo! for amazon: Sentiment extraction from small talk on the web. *Management Science*, 53(9):1375–1388, 2007.

[37] Akira Yoshihara, Kazuhiro Seki, and Kuniaki Uehara. Leveraging temporal properties of news events for stock market prediction. *Artificial Intelligence Research*, 5:103–110, 2016.

[38] Xiao Ding, Yue Zhang, Ting Liu, and Junwen Duan. Deep learning for event-driven stock prediction. In *IJCAI*, 2015.

[39] Samuel W. K. Chan and James Franklin. A text-based decision support system for financial sequence prediction. *Decision Support Systems*, 52:189–198, 2011.

[40] Sven S. Groth and Jan Muntermann. An intraday market risk management approach based on textual analysis. *Decision Support Systems*, 50:680–691, 2011.

[41] Brown E. Bulls, bear and birds? studying the correlation between twitter sentiment and the s&p500. *Proceedings of the Thirty Third International Conference on Information Systems*, 30:1–14, 2012.

[42] Yang Yu, Wenjing Duan, and Qing Cao. The impact of social and conventional media on firm equity value: A sentiment analysis approach. *Decision Support Systems*, 55(4):919–926, 2013.

[43] Timm O. Sprenger and Isabell M. Welpe. Tweets and peers: Defining industry groups and strategic peers based on investor perceptions of stocks on twitter. *Algorithmic Finance*, 1(1):57–76, 2011.

[44] Xue Zhang, Hauke Fuehres, and Peter A. Gloor. Predicting stock market indicators through twitter "i hope it is not as bad as i fear". *Procedia - Social and Behavioral Sciences*, 26:55 – 62, 2011. The 2nd Collaborative Innovation Networks Conference - COINs2010.

[45] Panagiotis Papaioannou, Lucia Russo, George Papaioannou, and Constantinos I. Siettos. Can social microblogging be used to forecast intraday exchange rates? *NETNOMICS: Economic Research and Electronic Networking*, 14(1-2):47–68, 2013.

[46] Hongkee Sul, Alan R. Dennis, and Lingyao Ivy Yuan. Trading on twitter: The financial information content of emotion in social media. *Hawaii International Conference on System Sciences*, 14, 2014.

[47] Gabriele Ranco, Darko Aleksovski, Guido Caldarelli, Miha Grčar, and Igor Mozetič. The effects of twitter sentiment on stock price returns. *PLoS One*, 10:1–21, 2015.

[48] Olga Kolchyna, Thársis Tuani Pinto Souza, Philip Treleaven, and Tomaso Aste. Twitter sentiment analysis: Lexicon method, machine learning method and their combination. 2015.

[49] Marten Risius, Fabian Akolk, and Roman Beck. Differential emotions and the stock market - the case of company-specific trading. *Proceedings of the European Conference on Information Systems (ECIS)*, 2015.

[50] Ling Liu, Jing Wu, Ping Li, and Qing Li. A social-media-based approach to predicting stock comovement. *Expert Systems with Applications*, 2(8):3893–3901, 2015.

[51] Jim Kyung-Soo Liew and Garrett Zhengyuan Wang. Twitter sentiment and ipo performance: A cross-sectional examination. *Journal of Portfolio Management*, 42(4):129–135, 2016.

[52] Francesco Corea and Enrico Maria Cervellati. The power of micro-blogging: How to use twitter for predicting the stock market. *Eurasian Journal of Economics and Finance*, 3:1–7, 2015.

[53] Jim Liew and Tamas Budavari. The 'sixth' factor - social media factor derived directly from tweet sentiments. *Journal of Portfolio Management*, 43:102–111, 2017.

[54] Ahmet K. Karagozoglu and Frank J. Fabozzi. Volatility wisdom of social media crowds. *Journal of Portfolio Management*, 43:136–151, 2017.

[55] Paul C. Tetlock, Mayta Saar-Tsechansky, and Sofus MacSkassy. More than words: Quantifying language to measure firms' fundamentals. *The Journal of Finance*, 63(3):1437–1467, 2008.

[56] Lily Fang and Joel Peress. Media coverage and the cross-section of stock returns. *The Journal of Finance*, 64(5):2023–2052, 2009.

[57] Helen Susannah Moat, Chester Curme, Adam Avakian, Dror Y. Kenett, H. Eugene Stanley, and Tobias Preis. Quantifying wikipedia usage patterns before stock market moves. *Scientific Reports*, 3:1–5, 2013.

[58] Tobias Preis, Helen Susannah Moat, and H. Eugene Stanley. Quantifying trading behavior in financial markets using google trends. *Scientific Reports*, 3, 2013.

[59] Alanyali Mrve, Helen Susannah Moat, and Tobias Preis. Quantifying the relation- ship between financial news and the stock market. *Scientific Reports*, 3, 2013.

[60] Hailiang Chen, Prabuddha De, Yu (Jeffrey) Hu, and Byoung-Hyoun Hwang. Wisdom of crowds: The value of stock opinions transmitted through social media. *The Review of Financial Studies*, 27(5):1367–1403, 2014.

[61] Cboe Exchange. Vix white paper. 2018.

[62] Xi Zhang, Siyu Qu, Jieyun Huang, Binxing Fang, and Philip S. Yu. Stock market prediction via multi-source multiple instance learning. *IEEE Access*, 6:50720 – 50728, 2018.

[63] Ehsan Hoseinzade and Saman Haratizadeh. Cnnpred: Cnn-based stock market prediction using several data sources. 2018.

[64] Bin Weng, Mohamed A. Ahmed, and Fadel M. Megahed. Aggregating multiple types of complex data in stock market prediction: A model-independent framework. *Knowledge-Based Systems*, 164:193–204, 2019.

[65] Xi Zhang, Yixuan Li, Senzhang Wang, Binxing Fang, and Philip S. Yu. Enhancing stock market prediction with extended coupled hidden markov model over multi-sourced data. *Knowledge and Information Systems*, 2018.