Wikipedia and Stock Return

Wikipedia Usage Pattern Helps to Predict the Individual Stock Movement

Pengyu Wei
Mathematical Institute
Oxford-Man Institute of Quantitative Finance
Oxford-Nie Financial Big Data Lab
University of Oxford
Oxford OX2 6GG, UK
weip@maths.ox.ac.uk

Ning Wang
Oxford NIE Financial Big Data Lab
Mathematical Institute
Oxford Internet Institute
University of Oxford
Oxford OX2 6GG, UK
wangn@maths.ox.ac.uk

ABSTRACT

Vast volumes of online information related to news stories, blogs and online social media have an observable effect on investor's opinions towards financial markets. But do these particular information reflect or impact people's decision-making in investment? This paper investigates whether data generated from Internet usage can be used to predict the movements in the financial market. We provide evidence that data on how often a company's Wikipedia page is being viewed is linked to its subsequent performance in the stock market. We then develop a portfolio in line with the Wikipedia usages and demonstrate that our investment strategy based on Wikipedia views is profitable both financially and statistically. Our finding implies that online web data such as Wikipedia presents an alternative insights on collecting and quantifying investor's sentiments towards financial markets, which can be further employed as a timely approximation of investor's behaviours in decision-making.

Keywords

Wikipedia, behavioural analytics, stock market prediction

1. INTRODUCTION

With Internet provision becoming so widespread, online resources have become the most important source of information for many people, making online users' activity a valuable dataset for understanding and measuring people's behaviour. Vast volumes of online information related to news stories, blogs and online social media have an observable effect on investor's opinions towards financial markets. However, do these particular information reflect or impact people's decision-making in investment? Some debate that it is extremely hard to "defeat the market" due to the fact that market efficiency triggers existing prices to continually incorporate and reflect all relevant information. The others claim that the online information is endogenous, and tend to be biased by nature. In spite of this there are more and more empirical evidence indicating that information requires time to diffuse and that the price effect of news events is sluggish over time [4].

Previous studies have demonstrated the predictive power of online users' activity. [11] discovered the correlation between search volumes for certain terms on Google and the breakeout of flu infections in US. [6] provided evidence that search data can be used to predict consumer behaviour and tourism. [13] employed Wikipedia activity data to predict movie box office. This resulted in a new research direction by investigating online usage data for financial markets. [7, 8] observed that Google search volume generally is a direct measure of investor attention and sentiment. [1] applied Twitter data to predict the Dow Jones Industrial Average (DJIA) index. [16, 14] attempted to predict stock indices based on Google trend and Wikipedia activity data respectively.

In this paper we analyze the relationship between the Wikipedia activity and subsequent stock returns of individual stocks. We find the Wikipedia homepage activities of particular publicly-listed companies are directly connected with its stock's subsequent performance. We develop classifiers in accordance with each company's past returns, prices, market values, book values and Wikipedia activity. We then use these classifiers to select stocks in an effort to predict the direction of markets, i.e. going up or going down, and design portfolios accordingly. The research exhibits that we can profit 57.46% per annum from the long position, and 7.81% from the short position, while the average return of all stocks is 15.37% during the whole period. Incorporating the long and the short position, we are able to construct a zero-investment long-short strategy, with an annual return of 65.27%. We additionally provide evidence that Wikipedia activity data plays an important role in the prediction, i.e., if we exclude Wikipedia data from the classifiers, the return of the long position shrinks to 54.18% per annum, both economically and statistically significant. Therefore our finding reveals that Wikipedia generally is a timely measure for investors attention.

Our paper is closely related to but distinct from [14] which presents the correlation between the usage of financially related Wikipedia pages and future stock market movement. Their research is concentrated on the DJIA index (which consists of only 30 stocks) while ours is a comprehensive evaluation of all stocks listed on the NYSE and NASDAQ. More-

WWW'16 Companion, April 11–15, 2016, Montréal, Québec, Canada.
ACM 978-1-4503-4144-8/16/04.
http://dx.doi.org/10.1145/2872518.2890089.

over, they only exhibit the predictive power of Wikipedia activity for the market index, but they do not explain that this predictive power cannot be absorbed by traditional risk factors, such as past returns, prices, market values and book values. We provide evidence that Wikipedia can be used to study the pricing impact of investor attention, even after controlling for the above risk factors. Additionally we observe that [5, 10] examines the impact of media on the stock market. As illustrated in subsequent sections, Wikipedia coverage is quite different from that of media.

The remaining paper is organised as follows: section 2 describes data used in this paper and relevant summary statistics; section 3 illustrates methods used in this paper; section 4 presents the result; section 5 concludes the paper.

2. DATA

2.1 Data Description

Stocks considered in this paper are restricted to firms listed on the NYSE and NASDAQ between 2011 and 2014, excluding closed-end funds. To remove listing and de-listing biases, we only preserve firms that are listed through the entire time period. The stock data is obtained from CRSP and the accounting data is from Compustat[1]. We keep firms that both have return and accounting data through the time period, ending up with 2756 firms.

We use a web crawler to identify each company's Wikipedia page[2] and download its traffic statistics (if the page ever exists[3]) from http://stats.grok.se. Table 1 describes all data mentioned above.

Table 1: Data de nition

Variable	Definition
wiki_view	daily access to a company's Wikipedia page
return	daily return
ret_month	return in the last month
bkvf	last available book value
b2m	last available book-to-market value
ask	ask price in the last month end
bid	bid price in the last month end
price	close price in the last month end

2.2 Summary Statistics

By the end of Dec 2014, 1596 out of 2756 firms have a Wikipedia article page. We define,

$$wiki = \log_{10}(1 + wiki\_view).$$

Table 2 reports summary statistics for the Wikipedia coverage. The first column specifies different values of $wiki$, and 0 1 stands for 0 $wiki < 1$, 1 2 stands for 1 $wiki < 2$, etc. Columns 2-5 show the (average) numbers of firms in each category of the $wiki$, as described above. The last row reports the average number across all years. Figure 1 plots the daily number of firms with different values of $wiki$.

Wikipedia coverage is relatively low (50%), compared to news media (around 70% as in [10]. There are more than one

[1]Wharton Research Data Services (WRDS) was used to access the data.

[2]We consider only the view data of the target pages and do not include that of the redirect pages. However, as noted in [12], redirects can have important effects. We left this to our future research.

[3]We take this value as 0 if the company does not have a Wikipedia page.

Table 2: Summary statistics of Wikipedia coverage

	2011	2012	2013	2014	Average
0-1	1490	1390	1365	1292	1384
1-2	752	767	765	799	771
2-3	433	505	536	581	514
3-4	79	93	88	82	85
4-5	2	1	2	2	2
5-6	0	0	0	0	0

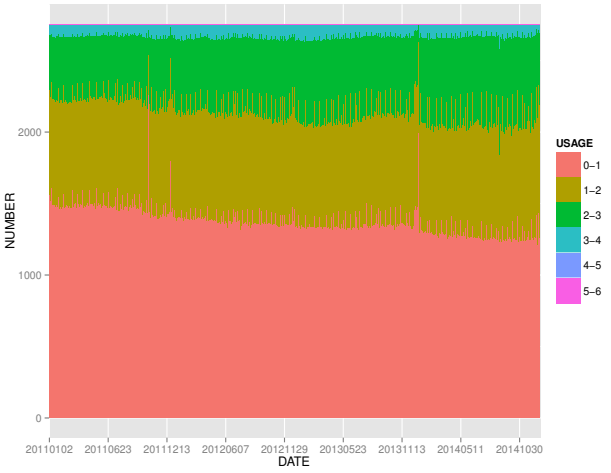

Figure 1: Wikipedia usage pattern

half of the firms considered in our sample who do not have Wikipedia articles or any traffic. This is in sharp contrast to news media. In addition, only a small fraction (less than 4%) of firms have active Wikipedia traffic (1,000 or more views in a single day).

3. METHODOLOGY

In this section we construct classifiers based on each company's past returns, prices, market values, book values and Wikipedia activity. We then use these classifiers to select stocks predicted to go up or go down, and construct portfolios accordingly.

First, we define the state of a stock on day t as UP, if the daily return of that stock on day t is no less than 1%, and as DOWN, if the daily return of that stock on day t is no more than -1%, and NEUTRAL otherwise. We then try to classify the state of each stock based on the past data. The set of predictors used to predict the state on day t+1 is described in Table 3.

We use the book value, book-to-market, past monthly return, bid/ask and price of a firm to control the pricing effect of its size, growth, momentum and liquidity, which are well-known risk factors in the finance literature [9, 3, 15]. We use wikidiff to measure the changes in Wikipedia activity.

For every t, we try to predict the state of each stock on t+1, based on data described in Table 3. More precisely we use data through t-39 to t as the training set, resulting

Table 3: Predictor de nition

Variable	Definition
return_1	daily return on day t
return_2	daily return on day t-1
return_3	daily return on day t-2
return_4	daily return on day t-3
return_5	daily return on day t-4
ret_month	last available monthly return before t
bkvf	last available book value before t
b2m	last available book-to-market value before t
ask	ask price in the last month end before t
bid	bid price in the last month end before t
price	close price in the last month end before t
wiki_1	wiki on day t
wiki_2	wiki on day t-1
wiki_3	wiki on day t-2
wiki_4	wiki on day t-3
wiki_5	wiki on day t-4
wikidiff_1	wiki_1 - wiki_2
wikidiff_2	wiki_2 - wiki_3
wikidiff_3	wiki_3 - wiki_4
wikidiff_4	wiki_4 - wiki_5

in 35 training samples, and train two classifiers based the Random Forest [2] with 500 trees each. Classifier I uses all predictors but variables related to Wikipedia and represents the classifier based on the traditional data. Classifier II uses all available predictors. We then construct 5 portfolios on t+1, based on the two classifiers, as reported in Table 4. We use the portfolio A to represent the market portfolio. Equally-weighted returns are calculated for each portfolio.

Table 4: Portfolio

Portfolio	Definition
U	stocks predicted to be UP by Classifier I
D	stocks predicted to be DOWN by Classifier I
U_W	stocks predicted to be UP by Classifier II
D_W	stocks predicted to be DOWN by Classifier II
A	all stocks

As we proceed in time, we train the classifiers adaptively and rebalance the aforementioned portfolios everyday.

4. RESULTS

Equally-weighted returns for each portfolio are reported in Table 5. The second column reports the average daily return, the third reports the t-statistics (p-values) and the fourth reports the (annualized) Sharpe ratio. We find that U, U_W and A perform well, both economically and statistically significant. Whereas D and D_W are indifferent from 0.

Table 5: Portfolio performance

Portfolio	Mean	T-test	Sharpe ratio
U	0.215%	5.24(<0.001)	2.677
D	-0.031%	-0.73(0.232)	-0.375
U_W	0.228%	5.49(<0.001)	2.805
D_W	-0.031%	-0.75(0.227)	-0.383
A	0.061%	1.66(0.048)	0.849

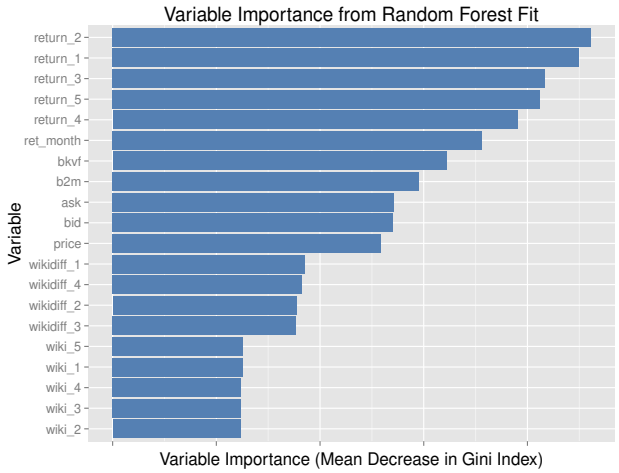

Figure 2: Variable importance

For a better illustration, we further construct 6 more portfolios, whose performance are described in Table 6. Portfolio U_W - U means taking a long position in U_W and a short position in U. The second column states the average daily return and the third reports the t-statistics (p-values). For robustness, we also perform the Wilcoxon signed-rank test[17] and the corresponding statistics and associated p-values are explained in the fourth column.

Table 6: Portfolio performance

Portfolio	Mean	T-test	Wilcoxon Signed-Rank Test
U_W - U	0.013%	2.12(0.017)	246724(0.064)
U - A	0.154%	11.81(<0.001)	344961(<0.001)
U_W - A	0.167%	11.57(<0.001)	348747(<0.001)
D_W - D	-0.001%	-0.15(0.442)	226514(0.209)
D - A	-0.091%	-7.18(<0.001)	163742(<0.001)
D_W - A	-0.092%	-6.94(<0.001)	167900(<0.001)

We uncover that U_W outperforms U 0.013% on average, statistical significant under both the t-test and the Wilcoxon signed-rank test[4]. This implies that Wikipedia usage pattern is helpful for predicting the individual stock movement. However, it is less useful for predicting which stock is going down, as the difference between D and D_W is neglectable. We also observe that all portfolios are distinctive from the market portfolio A. If we long the U_W and short the D_W, we can easily construct a zero-investment long-short strategy with a daily return around 0.259%, approximately 65.27% per annum. Therefore, we believe that our classifiers are effective and Wikipedia data is beneficial in the prediction.

This result reveals different perspectives of Wikipedia data. [16, 14] demonstrates that online data can be employed to predict financial markets. However, their strategy is solely based on the online data, without any reference to the traditional data. It is unclear whether the predictive power from the online data can be absorbed by the traditional data. In contrast, our research presents evidence that Wikipedia usage pattern can be used to predict financial market move-

[4]at 10% confidence level.

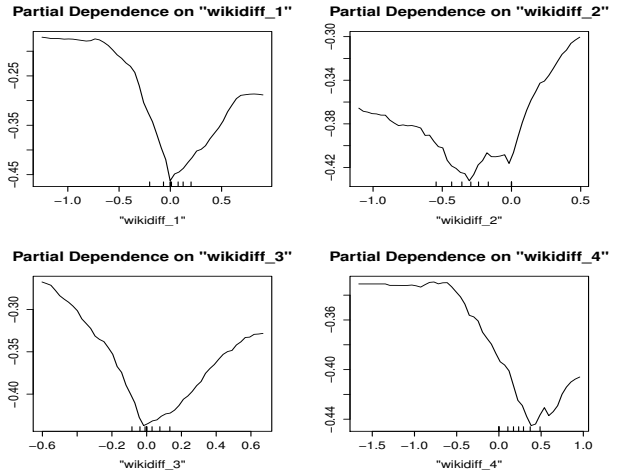

Figure 3: Partial dependence plot for class UP

ment, even after controlling for traditional data. In addition we notice that the traditional data contributes to a large proportion of the predictive power, as the difference between U and U_W is somewhat small compare to the difference between U_W and A.

Figure 2 plots the importance of variables from the last random forest fit. The importance of variables is measured by the mean decrease in Gini index if we exclude a certain variable. Consistent with our previous analysis, traditional data is the most important one, while the changes in Wikipedia activity also improve the performance.

Furthermore we depict the partial dependence plot on the changes of Wikipedia activity in Figure 3. The x axis is the variable for which partial dependence is sought and the y axis is the logits (i.e., log of fraction of votes) for the class UP. The results indicate that both the increase and decrease in the Wikipedia traffic can increase the probability for a stock to go up in the next day. However, since Wikipedia data is less useful when predicting which stock is going down, we are unable to make any solid conclusion on the pattern of DOWN.

5. CONCLUSIONS

In summary, we present evidence that data on how often a company's Wikipedia page was viewed is related to the subsequent performance of the company's stock. Our findings demonstrate that online data provides new insight in the process of investor's information gathering. By combining traditional data and large online data, we are able to gain new insight on investor's behaviour. However, despite the predictive power of Wikipedia data, it is still rather challenging to establish causal relationships between data and investor's trading choices for the reason that the views of Wikipedia page purely represent unobserved attributes of the firms that influence information diffusion and investor's choices simultaneously.

6. ACKNOWLEDGMENTS

The authors acknowledge financial supports from the Oxford-Man Institute of Quantitative Finance and the Oxford-Nie Financial Big Data Laboratory. Wharton Research Data Services (WRDS) was used in preparing this paper. This service and the data available thereon constitute valuable intellectual property and trade secrets of WRDS and/or its third-party suppliers.

7. REFERENCES

[1] J. Bollen, H. Mao, and X. Zeng. Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1):1–8, 2011.

[2] L. Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[3] M. M. Carhart. On persistence in mutual fund performance. *Journal of Finance*, 52(1):57–82, 1997.

[4] A. Carretta, V. Farina, D. Martelli, F. Fiordelisi, and P. Schwizer. The impact of corporate governance press news on stock market returns. *European Financial Management*, 17(1):100?119, 2010.

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

[6] H. Choi and H. Varian. Predicting the present with google trends. *Economic Record*, 88(s1):2–9, 2012.

[7] Z. Da, J. Engelberg, and P. Gao. In search of attention. *Journal of Finance*, 66(5):1461–1499, 2011.

[8] Z. Da, J. Engelberg, and P. Gao. The sum of all fears investor sentiment and asset prices. *Review of Financial Studies*, 28(1):1–32, 2015.

[9] E. F. Fama and K. R. French. The cross-section of expected stock returns. *Journal of Finance*, 47(2):427–465, 1992.

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

[11] J. Ginsberg, M. H. Mohebbi, R. S. Patel, L. Brammer, M. S. Smolinski, and L. Brilliant. Detecting influenza epidemics using search engine query data. *Nature*, 457(7232):1012–1014, 2009.

[12] B. M. Hill and A. Shaw. Consider the redirect: A missing dimension of wikipedia research. In *Proceedings of The International Symposium on Open Collaboration*, page 28. ACM, 2014.

[13] M. Mestyán, T. Yasseri, and J. Kertész. Early prediction of movie box office success based on wikipedia activity big data. *PLoS ONE*, 8(8):e71226, 2013.

[14] H. S. Moat, C. Curme, A. Avakian, D. Y. Kenett, H. E. Stanley, and T. Preis. Quantifying wikipedia usage patterns before stock market moves. *Scientific Reports*, 3, 2013.

[15] L. Pástor and R. F. Stambaugh. Liquidity risk and expected stock returns. *Journal of Political Economy*, 111(3):642–685, 2003.

[16] T. Preis, H. S. Moat, and H. E. Stanley. Quantifying trading behavior in financial markets using google trends. *Scientific Reports*, 3, 2013.

[17] F. Wilcoxon. Individual comparisons by ranking methods. *Biometrics Bulletin*, pages 80–83, 1945.