Can Word Usage and Sentiment on Twitter Predict Stock Market Prices?

*An analysis in covariance between keyword polarity and stock prices.*

**M.A.C. van Dijk (U1266928/ ANR: 724171)**

**B. Gerla (U1264461/ ANR: 263212)**

**B. G. M. Pinheiro (U2009518/ ANR: 866719)**

**S. Cece (U2009468/ ANR: 657912)**

**Tilburg University: Social Media Analytics (BDM/HAIT)**

**Words: 4908**

**Abstract**

During recent years, machine learning techniques have found their way into a multitude of (economic) research. This paper attempts to utilize these techniques by analyzing Twitter sentiment surrounding shoe industry specific keywords such as “shoes”, “sneakers” and “boots”. Our analysis concerns four companies that are active within the shoe industry, namely, Nike, Sketchers, Steven Madden and Wolverine Worldwide. Sentiment will be extracted from tweets through the use of a machine learning algorithm, the Naïve Bayes Classifier. Our measure of sentiment will be aggregated and used as a proxy for industry confidence. The paper therefore attempts to encapsulate whether a link exists between this paper’s proxy of market confidence and shoe industry stocks. It is found that weak links might exist between our measure of market confidence and shoe industry specific stocks. More research is needed to thoroughly substantiate our claims.

*Table of Contents*

1. Introduction……………………………………………….
2. Method……………………………………………………
3. Results…………………………………………………….
4. Discussion………………………………………………....
5. Conclusion………………………………………………...
6. References………………………………………………...
7. Introduction

In the past couple of years, the presence of big-data and machine learning techniques have been rising in popularity in business as well as academia. The growth of data is expected to continue, following a hyperbolic trend and will have increased 300-fold through the years 2005-2020 (Gantz & Reinzel, 2012). With these amounts of data and the uprising of social networking sites (hereinafter referred to as SNSs) such as Twitter, Facebook and Instagram, it is highly likely that valuable data is present within these sites. Since SNSs provide such vast amounts of data, it seems only logical that they might yield valuable insights into various fields of research such as sentiment analysis (Pak & Paroubek, 2010), political science (Sandner, Sprenger, Tumasjan & Welpe, 2010), disaster detection (Matsuo, Okazaki & Sakaki, 2010) and finance (Bollen, Mao & Zeng, 2011). This paper will attempt to use one of these SNSs (Twitter), and combine it with the field of financial research. Specifically, we will lay focus on stock market prediction.

Earlier research into the topic, such as conducted by Fama (1965), relied heavily upon the Efficient Market Hypothesis (EMH). The EMH states that stock prices are purely driven by new information, implicitly stating that past information should not yield any predictive power over future stock prices. Moreover, since news is future information and cannot be predicted, the accuracy of predicting stock price movements cannot exceed 50 percent. The EMH is further supported with theoretical literature by Samuelson (1965) and proven to be true in the empirics by many papers, as summarized by Jensen (1978). However, more recently, the academic literature on the topic has been steered into the opposite direction. Fama & French (1992) for example show that the EMH need not hold true in all cases, and that some past indicators hold statistically significant predictive power over future stock price movements. Furthermore, a case can be made against the preconceived notion, laid upon by the EMH, that future stock price returns cannot be predicted by past information. Several papers (such as e.g., Basu, 1977; Lo & MacKinlay, 1999) have found outright empirical evidence in contradiction to the EMH. Moreover, looking more closely into the field of research that this paper will delve in, several papers (such as e.g., Bartov, Faurel, & Mohanram, 2016; Bollen, Mao, & Zeng, 2011) have found statistically significant predictive power over future stock returns with the help of sentiment analysis using Twitter. Textual analysis and machine-learning techniques can be applied to several other fields of financial research as well. Kearney and Liu (2014) comprehensibly summarize how textual analysis has predictive power over various financial measures, such as revenues, trading volumes of stocks and stock returns. Their summary coherently portrays that media, such as earnings press releases, 10-Ks and a variety of other corporate news-outlets, can yield statistically significant predictive power over various important financial measures of companies and aggregated markets alike.

This paper will expand on the existing literature by analyzing what effect specific keyword sentiment on Twitter might have on stock prices. The paper at hand will attempt to find a statistically significant correlation between our measure of market confidence in the shoe manufacturing industry (derived from sentiment on Twitter), and shoe industry-specific stocks. Our analysis will lay focus on four particular stocks, namely: Nike, Sketchers, Steven Madden and Wolverine Worldwide. We have opted to select for these four companies as they represent the four largest companies within the shoe manufacturing industry, as measured by their market capitalization. Furthermore, we use several keywords to extract sentiment regarding the shoe market industry, specifically: “shoe(s)”, “boot(s)”, “sneaker(s)”. Past research, such as conducted by Bartov, Faurel and Mohanram (2016), specifically laid focus on keywords that are directly linked to the respective companies (i.e. using the specific ticker that the company is traded under on the stock markets). Our research investigates whether more general keywords, concerning a specific industry, might also yield predictive power over stock returns. Since no prior research has been done into this topic, it would be of great interest to academics.

In order to analyze the sentiment of tweets, this paper will make use of the Twitter Streaming API[[1]](#footnote-1) to load in tweets at real-time speed. Afterwards, a machine-learning algorithm will analyze tweets and classify them as either negative or positive. The polarity measure induced from said tweets will then be put next to the stock returns of the companies we intend to analyze. As such, we will attempt to find a statistically significant relationship between our proxy of shoe industry confidence and our selected companies’ stock returns. The results indicate that there might be links between our measure of keyword polarity and stock market prices, albeit that these links are likely to be weak. Further research is needed in order to convincingly substantiate our results.

The paper is structured as follows. Section 2 discusses our method of obtaining and analyzing the data. Section 3 explains the results derived from the data. Section 4 discusses the shortcomings of our research and gives recommendations for future research. Section 5 concludes.

1. Method

Tweets are obtained in real-time via the Twitter Streaming API, with the help of the Python (3.5) programming language. Within this environment, good use is made of the Tweepy module, in order to keep the program as accessible as possible. The program is expected to run every trading day from one hour before market open in the U.S. to one hour after market close. Starting the program means manually inserting the run-time and interval to save the data. This has to be done manually for every trading day, if not there will be no data for a given day *k*. When being streamed in via connecting to the Twitter Streaming API, a tweet looks like the following.



Figure 1. An example of how a tweet looks like when it is loaded in through the Streaming API. (1), (2) and (3) are what is used in the process of cleaning and analyzing the tweet. Redundant (meta)data is not shown on this figure.

While this paper will make most use of the “text” portion of a tweet to analyze sentiment, the metadata as provided by Twitter makes it considerably easier to clean and filter tweets. Hence, before tweets are analyzed they are run through a similar filtering process, based on Bollen, Mao and Zeng (2011). In addition to the filtering techniques applied by said authors, this paper will also exclude any tweets that are not written in English, tweets that mention a YouTube link or tweets that originate from users that have less than 50 followers. These additions help to make sure that the sentiment analysis is done only on those tweets that show significant value to the overall sentiment within a given time-period *t*. Moreover, they make sure that the sentiment of tweets is not biased towards any direction due to an overflow of, for example, retweets or website links.

Tweets are classified using machine-learning algorithms that have been previously trained on a subset (corpus) of tweets. One of these algorithms is provided for free by the TextBlob[[2]](#footnote-2) module, this algorithm is partially based on the one as described by Narayanan, Arora and Bhatia (2013). It is pre-trained on an IMDB movie review corpus and makes use of Naïve Bayes Classification[[3]](#footnote-3) for sentiment extraction. In the case of this paper, the classifier is trained to recognize positive and negative sentiment. Furthermore, we opted to test another classifier, using the Naïve Bayes Classification algorithm as well, only now training it ourselves on a corpus of tweets as provided by the NLTK library for Python[[4]](#footnote-4). We initially suspected that the classifier as trained by ourselves would yield higher accuracy rates because its training data were tweets instead of movie reviews. In the process of training our classifier, it had been fed 5000 tweets, of which 2500 were positively labelled and 2500 of them were negatively labelled. Moreover, in order to confirm or refute our suspicion that our own classifier would achieve higher accuracy, we have had both algorithms subject to testing. Testing data was taken from the same corpus as mentioned above, we made sure there was no overlap. We used 2000 tweets in total for testing purposes, of which 1000 were positively labeled and 1000 of them were negatively labelled. Our testing procedures resulted in an accuracy of 77.65% for our own classifier, while the pre-trained TextBlob classifier resulted in an accuracy score of 87.35%. Our initial expectations were that our own classifier would be better at predicting sentiment stemming from tweets, as this is also the data it had been trained on. However, since we came to the conclusion that the pre-trained classifier yielded better results, we decided to focus purely on this classifier and disregard our own.

Since the program is expected to run in real-time during every trading day, sentiment is being measured as tweets are being streamed in. The program uses a ternary distinction in mood, measuring the amount of positive, negative or neutral tweets that are being streamed in. An example of tweets and classification can be seen below.

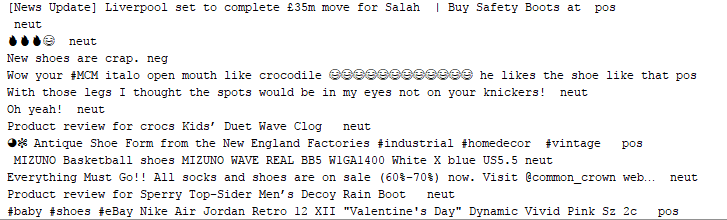


Figure 2. On the left: Text as being loaded in by the Twitter Streaming API. On the right: algorithms’ classification of mood.

Whenever a tweet is labelled as positive, the program will increase the variable ‘*pos\_count’* to *pos\_count + 1.* The same goes for the variable ‘*neg\_count’* that keeps track of all negative tweets. Neutral tweets however will not be saved into a variable, as they do not provide any contribution to the measure of overall sentiment (Cao, Duan & Yu, 2013). Every *t* minutes, these sentiment variables are reset and our polarity measure is deducted from them (note that *t* = 5 in this paper). Polarity here is simply defined as:

*Polarityt =*

When *t* minutes have past, the program will append the polarity measure to another variable that keeps track of all past polarity measures for a given day *k*. Furthermore, every *t* minutes the real-time stock price data will be pulled in from the Yahoo Finance API. This data will also be saved to a variable for each day *k.* Additionally, a time-key will also be added to the data such that specific events can be linked to the polarity and/or stock price data within a given time period. At the ending of each trading day, all of the measures that have been saved during the day will be put together into one (Excel) file for further analysis. Hence, at the ending of each day the data is structured as follows.



Figure 3. Structure of our data after it has been cleaned, processed and saved to Excel by our Python file. Analysis on the data is made easy by storing it as such. Alphabetically ordered letters in the beginning of each column are added such that our program sorts the data conveniently for further analysis.

1. Results

Our program recorded a total of 30 trading days, which equates to one and a half months of data, since markets are not open during weekends. Furthermore, we analyzed 513,880 tweets over the span of these 30 days, which equates to an average of 17.129 tweets per day. Our greatest recorded jump in stock price happened on 21-04-2017, where the stock price of Steven Madden surged by 1.9 percent at markets open. The greatest recorded decrease in stock price stems from Nike, their stock price fell by 2.9 percent when markets opened on 16-05-2017. While this information is interesting on its own, the main goal of this paper is to see whether a trading strategy could be profitable, using real-time Twitter data as a trading mechanism. In order to achieve this, we analyze whether our measure of polarity positively or negatively covaries with the stock prices for different lags of time.

To start our analysis procedure, we start with normalization of the data, in an attempt to ensure that our data remains comparable. We go about doing this by simply adding extra data-rows to our Excel files that contain relative changes instead of absolute changes by applying the following formula.

*=*

Where indicates either stock prices, our polarity measure or the total number of tweets.

Since we are interested in co-movements between stock prices and our other variables we put focus on calculating Pearson correlation coefficients, and see whether the correlations found are statistically significant from zero. Additionally, we were not only interested in seeing whether our measure of polarity had any relationship between stock prices, but also the total number of tweets sent in. Furthermore, we are most interested in seeing a statistically significant effect of a/ past change in polarity affecting a future change of stock prices, as that would enable us to trade on it profitably. Hence, we introduced lagged variables of our polarity measure and the total number of tweets, with a maximum lag of 240 minutes (4 hours). We first analyzed the correlation between the number of tweets sent per time interval *t*, and the stock price movement. We show how the correlation measures changes over time per lag, treating every day *k* as a separate observation (N = 30). The results of this analysis are as follows.

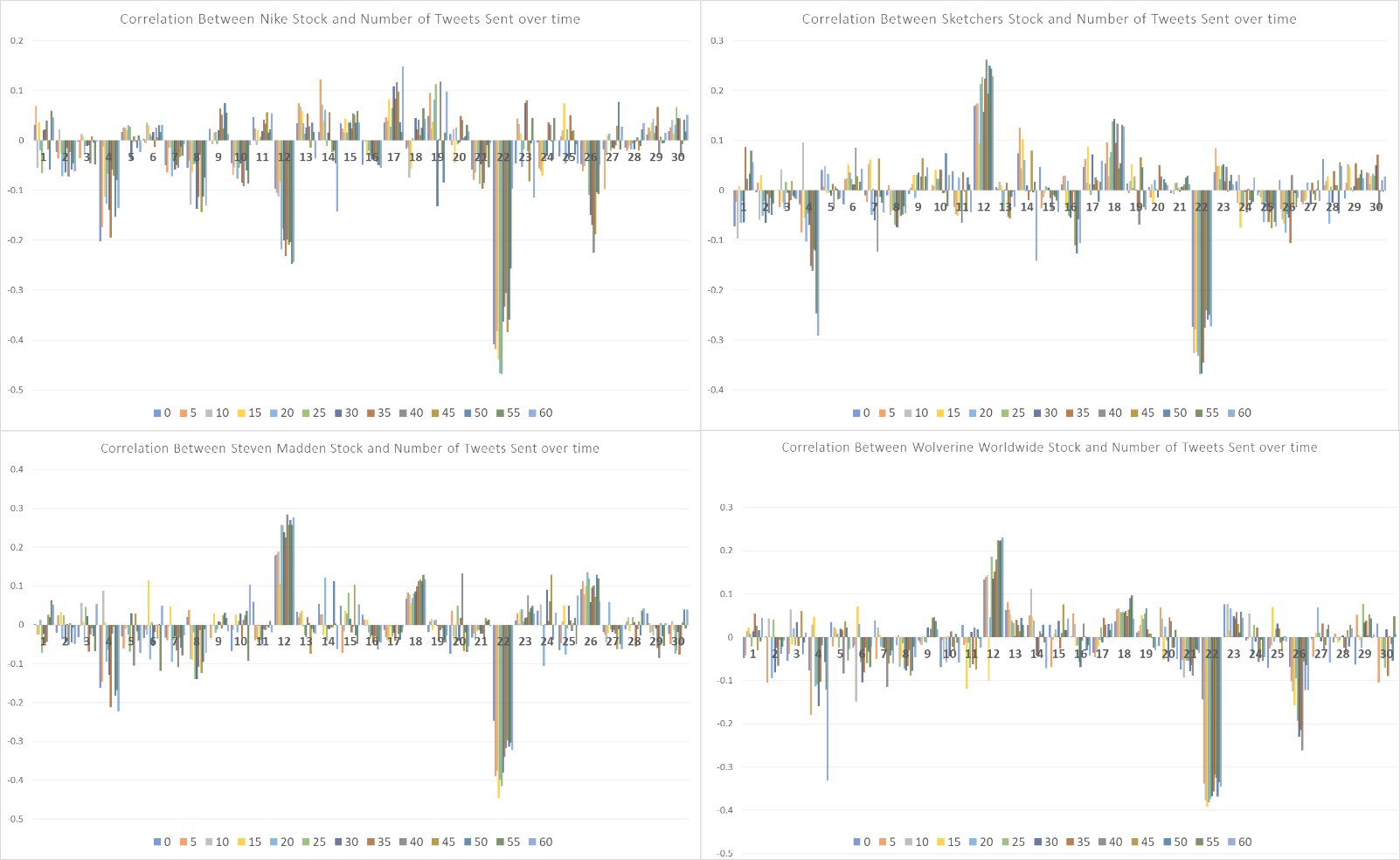


Figure 4. Pearson correlation coefficients for each stock over time at a given lag. Files are treated independently but are chronologically ordered. The legend on the bottom displays how much lag (in minutes) corresponds which color in our chart. Important events during our recording period have been marked on the top-right graph. (1) French Elections. (2) Fed policy meeting. (3) Trump-Russia scandal aftermath.

Deducing from the graphs we can see that the correlation coefficients seem to vary quite substantially on a day-to-day basis. Furthermore, the lagged correlation coefficients seem to not show great differences. If a day *k* has an overall strong negative correlation coefficient, this usually counts for every lagged correlation coefficient on that particular day, and vice-versa. Moreover, we see that political turmoil and other stock-market sensitive events affect the correlation coefficients quite significantly. We attribute most of these extreme changes to heavy overall stock-market volatility, thus causing for great increase or decrease of the total correlation for each lag.

However, our research is not focused on investigating empirically whether a link exists between the number of tweets sent and stock prices, but on the covariance between stock prices and keyword polarity. Hence, we analyze the correlation coefficients between keyword polarity and stock prices as well. We treat every day *k* as a separate observation, just as it has been done in the previous graph.

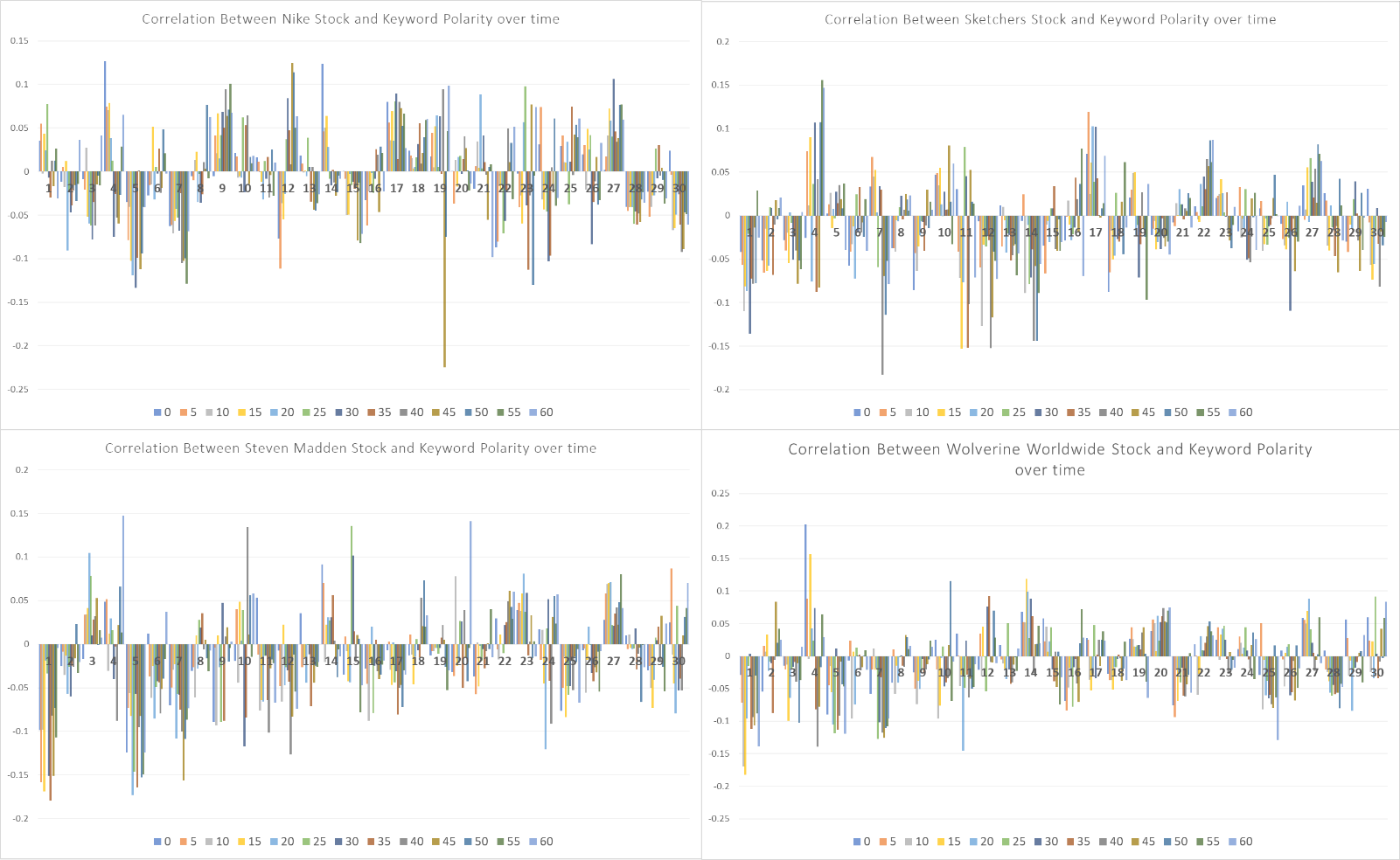


Figure 5. Pearson correlation coefficients for each stock over time at a given lag. Files are treated as independent observations, but are chronologically ordered. The legend on the bottom of each graph displays how much lag (in minutes) corresponds to which color in our chart. Important events are left out of these graphs as they are identical to figure 4.

Comparing figure 4 to figure 5, it becomes immediately clear that the graphs differ quite substantially. The correlation coefficients regarding number of tweets sent and stock prices seem to yield substantially higher absolute values. Furthermore, we see that the correlation coefficients on keyword polarity and stock prices tend to go below zero more often, especially in the case of Steven Madden. While taking a glance at the data as presented might yield for interesting insights, it does however not provide definitive and reliable ways to implement a trading strategy on. Hence, we further investigate whether our measure of correlation is provably and statistically different from zero.

In order to analyze whether our correlation coefficients are statistically different from zero, we need to impose the correlation coefficients for each company over all of our observations to statistical testing. Since we assume that our correlation coefficients are not distributed normally, we are not able to utilize t-tests, as they explicitly require normally distributed sample means. Considering the fact that t-tests are not suitable for our data, we will make use of Wilcoxon signed-rank tests (Wilcoxon, 1945) to investigate statistical significance. We test the hypothesis that our sample means for our Pearson correlation coefficients are symmetrically distributed around zero (H0), against the hypothesis that our sample means do not follow a symmetrical distribution around zero (H1). We use significance tests for the correlation coefficients between stock prices and number of tweets sent, as well as the correlation coefficients between stock prices and keyword polarity. The results are as follows.

Average Pearson Correlation Coefficients Between Stock Price and Number of Tweets Sent

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Lag/Company | Nike | Sketchers | Steven Madden | Wolverine Worldwide |
| 0 | -0.028  (163) | 0.007  (165) | -0.007  (219) | -0.010  (186) |
| 5 | -0.023  (185) | 0.001  (220) | -0.013  (199) | -0.025  (166) |
| 10 | -0.021  (203) | 0.005  (175) | -0.003  (232) | -0.004  (201) |
| 15 | -0.016  (218) | 0.002  (184) | -0.004  (201) | -0.019  (214) |
| 20 | -0.032  (161) | -0.011  (198) | -0.016  (166) | -0.018  (202) |
| 25 | -0.031  (149)\* | -0.006  (208) | -0.012  (193) | -0.021  (183) |
| 30 | -0.036  (150)\* | -0.014  (182) | -0.013  (190) | -0.028  (183) |
| 35 | -0.022  (225) | -0.005  (226) | -0.027  (120)\*\* | -0.021  (190) |
| 40 | -0.023  (190) | -0.008  (197) | -0.007  (166) | -0.021  (192) |
| 45 | -0.035  (142)\* | -0.006  (225) | 0.002  (221) | -0.009  (201) |
| 50 | -0.034  (147)\* | -0.006  (215) | -0.007  (196) | -0.011  (201) |
| 55 | -0.022  (186) | -0.005  (214) | -0.020  (152)\* | -0.010  (199) |
| 60 | -0.018  (191) | -0.015  (206) | -0.003  (229) | -0.030  (137)\*\* |

Table 1. Average Pearson correlation coefficients for a company’s stock price and number of tweets sent at a given lag in minutes. Correlation coefficients are rounded to three decimal places. Wilcoxon’s W-statistic in parentheses. \*\*\* p < .01; \*\* p < .05; \* p < .10

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Lag/Company | Nike | Sketchers | Steven Madden | Wolverine Worldwide |
| 0 | 0.004  (221) | -0.008  (186) | -0.012  (174) | 0.007  (213) |
| 5 | -0.003  (228) | -0.006  (182) | -0.010  (177) | 0.004  (187) |
| 10 | -0.006  (184) | -0.020  (143)\* | -0.025  (97)\*\*\* | -0.026  (111)\*\* |
| 15 | -0.001  (230) | -0.012  (161) | -0.020  (130)\*\* | -0.009  (176) |
| 20 | -0.004  (215) | -0.016  (125)\*\* | -0.023  (131)\*\* | -0.019  (146)\* |
| 25 | 0.006  (200) | -0.001  (228) | -0.006  (204) | -0.002  (222) |
| 30 | -0.013  (164) | -0.004  (216) | -0.013  (169) | -0.022  (210) |
| 35 | -0.009  (194) | -0.017  (143)\* | -0.032  (91)\*\*\* | -0.029  (111)\*\* |
| 40 | 0.001  (231) | -0.027  (124)\*\* | -0.024  (130)\*\* | -0.016  (77)\*\*\* |
| 45 | -0.018  (170) | -0.015  (158) | -0.015  (187) | -0.001  (159) |
| 50 | 0.002  (215) | -0.006  (205) | -0.014  (165) | -0.001  (219) |
| 55 | -0.001  (224) | 0.001  (221) | -0.020  (127)\*\* | -0.006  (204) |
| 60 | 0.010  (179) | -0.005  (184) | 0.001  (213) | -0.013  (200) |

Table 2. Average Pearson correlation coefficients for a company’s stock price and keyword polarity at a given lag in minutes. Correlation coefficients are rounded to three decimal places. Wilcoxon’s W-statistic in parentheses. \*\*\* p < .01; \*\* p < .05; \* p < .10

Average Pearson Correlation Coefficients Between Stock Price and Number of Keyword Polarity

Deducing from these tables we can clearly see how little effect our independent variables seem to have on the stock price. Our maximum correlation measured in absolute terms is 0.036, while our minimum is a mere 0.001. This means that, on average, our maximum expected co-movement between the independent variable (number of tweets in this case) and the stock price is 3.6 percent. Furthermore, we see that almost all correlation coefficients in Table 1 are not statistically significant. Besides this, statistical significance of the coefficients is spread out over all of the lagged intervals, which might indicate that the statistical significance is more accredited to variance than the correlation coefficients being actually significantly different from zero. However, when looking at Table 2, our focal point, we see more promising results. We see more correlation coefficients being statistically different from zero, albeit that the effect in absolute terms remains small.

When we focus more on the specifics of the results, two particular lags look the most promising, namely 35 and 40 minute lags. In these two lags, we see that for most companies (with the exception of Nike), the results are highly statistically significant. While the average correlation remains miniscule, we see opportunities for profitable trading strategies here. Intervals of 10, 15 and 20 minutes also provide us with statistically significant correlation coefficients. These effects are however less consistent across companies, and we therefore deem them to be of lower interest for our research. Moreover, our variables covary negatively, indicating that a rise in keyword polarity leads to a decrease in stock price and vice-versa. Using this knowledge, implementing trading strategies that utilize these results would mean placing opposite bets on stocks, relative to how keyword polarity is moving.

1. Discussion

Once a social media cycle is finished, there is an opportunity to analyze the results and look back at certain aspects of the analysis which have fallen short of expectations. By doing so, we create opportunity for improvement. If some aspects of how the data was collected, organized and analyzed were different, we may have ended up with different results. It is important to highlight these aspects, not only for academics, but it also enables for reflection of our own work.

In the correlation analysis, we established lagged comparisons. We wanted to check how much a change in the polarity or in the number of tweets in (*t* – *y)* would be correlated with the change in price of stock *i* in a given time *t*, with *y* being the amount of lag in minutes. In our analysis, we included *y*=0 (meaning instant changes, which were unlikely to occur but important for the study nevertheless) and added lagged variables for each 5-minute time interval, between 5 minutes and 240 minutes. If we increased the maximum lag, we might have achieved new insights and results with greater statistic relevance. Furthermore, adopting a different interval to calculate polarity variation (i.e. changes in *t*) might positively affect relevance of our results as well.

Additionally, if we had access to a bigger dataset, our results would gain in relevance, ensuring that our results are not biased into any particular direction. This could be achieved by either collecting data for a longer period of time, or by accessing a historical tweet archive such as GNIP[[5]](#footnote-5). The latter definitely grants more information in less time, but would unfortunately require monetary funds that were not compatible with our budget. Nevertheless, a significant increase in the sample-size of our data could lead to more substantiated and conclusive evidence.

In this work, we extracted tweets with the keywords “shoe(s)”, “boot(s)” and “sneaker(s)” and analyzed the stock returns from Nike, Sketchers, Steven Madden and Wolverine Worldwide. If we were to analyze more stocks, this could lead to more thorough research into the generalizability of our claims. Moreover, if we added the name of the companies as keywords for the specific correlation to each stock, we would change the results in a more precise and accurate way. Selecting different keywords and analyzing the results thereof might also yield intriguing insights.

As this study revolves around the stock market, we are interested in seeing if the insights obtained in the results could lead to profitable returns. Despite the volatility of our correlation measures, applying what we did find to be statistically significant to simulate a trading strategy can give interesting insights that might inspire further research. While the chances of this strategy being highly profitable are slim, the academic relevance of such an action is still pertinent.

Adopting a different classification method for the sentiment of the tweets could also be an interesting change. Bollen, Mao and Zeng (2011) used a mood-tracking tool called GPOMS (Google-Profile of Mood States) in their analysis. This tool classifies texts in six moods: Alert, Calm, Happy, Kind, Sure and Vital. Using such classification instead of simply positive and negative can yield for more meaningful results since it recognizes that human sentiment is multidimensional. While in theory this sounds interesting, the application is more difficult. Instead of using a single parameter of polarity, the model would now analyze six different polarity measures related to stock prices. Since the implementation of more variables usually leads to a more predictive model, it would hence be of interest for future research to include this.

5. Conclusion

In this paper, we investigated whether a link exists between industry sentiment surrounding specific keywords on Twitter, and stock returns. This paper builds on previous research as it attempts to analyze whether general (instead of stock-specific) sentiment on Twitter provides predictive value for industry stock prices. We used Twitter’s Streaming API to stream in over 500,000 tweets over the span of 30 days, analyzing whether our proxy of shoe industry confidence has an effect on shoe industry specific stocks. Moreover, we made use of a machine-learning algorithm in the shape of a Naïve Bayes Classifier to extract sentiment from the tweets. Our algorithm was trained on a corpus of movie-reviews, taken from IMDB’s website.

Additionally, we defined our own measure of polarity in order to investigate whether this measure covaried negatively or positively with stock prices. Our results show that positive changes in keyword polarity of 35 to 40 minutes in the past could lead to negative changes in stock prices now and vice-versa. We employ Wilcoxon signed-rank tests to test for statistical significance of our Pearson correlation coefficients, as t-tests require normally distributed sample means which we assume not to have. The effects are found to be statistically significant over most of our investigated stocks, at acceptable significance levels. However, we do not draw conclusive evidence from these results as our sample size is relatively small (N = 30). Effects regarding the total number of tweets sent, related to the shoe market industry, did not yield results that were of great interest (i.e. not statistically significant) to our research (i.e. not statistically significant in most cases).

Further research could focus on investigating whether adding additional stocks, lags or independent variables to the dataset could positively affect the results. Moreover, implementing a simple trading strategy that trades on our results would seem of interest to business and academics.

Reference List

* Asur, S., & Huberman, B. A. (2010). Predicting the future with social media. In Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference (Vol. 1, pp. 492-499). IEEE.
* Bartov, E., Faurel, L., & Mohanram, P. S. (2016). Can Twitter Help Predict Firm-Level Earnings and Stock Returns?. *Rotman School of Management Working Paper No. 2631421.*
* Basu, S. (1977). Investment Performance of Common Stocks in Relation to Their Price-Earnings Ratios: A Test of the Efficient Market Hypothesis. *The Journal of Finance,* 32(3), 663-682.
* Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of Computational Science, 2(1), 1-8.*
* Cao, Q., Duan, W., Yu, Y. (2013). The impact of social and conventional media on firm equity value: a sentiment analysis approach. *Decision Support Systems,* 55(4), 919-926.
* Fama, E. F. (1965). The behavior of stock-market prices. *The journal of Business*, *38*(1), 34-105.
* Fama, E. F., & French, K. R. (1992). The cross‐section of expected stock returns. *The Journal of Finance*, *47*(2), 427-465.
* Gantz, J., & Reinsel, D. (2012). The digital universe in 2020: Big data, bigger digital shadows, and biggest growth in the far east. *IDC iView: IDC Analyze the future*, *2007(2012), 1-16.*
* Jensen, M. C. (1978). Some Anomalous Evidence Regarding Market Efficiency. *Journal of Financial Economics,* 6(2-3), 95-101.
* Lo, A., & MacKinlay, C. A. (1999). A Non-Random Walk Down Wall Street. *Princeton:* *Princeton University Press*.
* Matsuo, Y., Okazaki, M., & Sakaki, T. (2010). Earthquake shakes Twitter users: real-time event detection by social sensors. *Proceedings of the 19th international conference on World Wide Web,*. pp. 851-860.
* Pak, A., & Paroubek, P. (2010). Twitter as a corpus for Sentiment Analysis and Opinion Mining. *LREc, 10*(2010).
* Samuelson, P. A. (1965). Proof That Properly Anticipated Prices Fluctuate Randomly. *International Management Review,* 6, 41-49.
* Sandner, P. G., Sprenger, T. O., Tumasjan, A., & Welpe, I. M. (2010). Predicting elections with twitter: What 140 characters reveal about political sentiment. *ICWSM*, *10*(1), 178-185.
* Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics bulletin,* 1(6), 80-83.

1. https://dev.twitter.com/streaming/overview [↑](#footnote-ref-1)
2. https://textblob.readthedocs.io/en/dev/ [↑](#footnote-ref-2)
3. A probabilistic classifier that is based on Bayes’ theorem with naïve independence assumptions between features. For more information: http://www.ic.unicamp.br/~rocha/teaching/2011s1/mc906/aulas/naive-bayes.pdf [↑](#footnote-ref-3)
4. http://www.nltk.org/api/nltk.corpus.reader.html#module-nltk.corpus.reader.twitter [↑](#footnote-ref-4)
5. https://gnip.com/ [↑](#footnote-ref-5)