CashTagNN: Using Sentiment of Tweets with CashTags to Predict Stock Market Prices

Neeraj Rajesh, Lisa Gandy
Department of Computer Science
Central Michigan University
Mt Pleasant, MI 48858
Email: rajesh1n, gandy1l@cmich.edu

***Abstract*—In this paper we discuss a system, CashTagNN, which uses the sentiment and subjectivity scores of tweets that include cashtags of two companies, Apple and Johnson and Johnson, to model stock market movement, and in particular predict opening and closing stock market prices. We demonstrate that by using only sentiment and subjectivity along with a neural network machine learning model we can predict the opening and closing prices of the two companies with high accuracy.**

I. Introduction

The large amount of data via twitter has allowed researchers unprecedented access to real time conversations about a wide array of topics. This data has allowed researchers to predict a wide range of behaviors and occurrences such as flu trends [1], postpartum behavior [2], voting intentions [3] etc using tweets. Not surprisingly researchers have also began to wonder if there is a way to use twitter data in regards to products and the consumer industry. Chamlertwat et al [4] have investigated how to predict product demands using twitter, and Jaring et al [5] among many others [6] [7] [8], have investigated how to market products on twitter effectively.

Moving to the stock market, researchers such as Bollen et al [9], Souza [10] and Zheludev [11] have shown that sentiment of tweets predicts the movement of the stock market with high accuracy. This work is exciting as it suggests that tweets could possibly as a variable in regards to stock market trading and investment.

Though a significant amount of work has been initiated about stock market price prediction, the work has not not utilized the presence of cashtags to focus on specific stocks. Cashtags are stock market symbols that can be included in tweets and when preceded with a dollars sign (for example $JNJ in regards to Johnson and Johnson) become clickable. By clicking on the cashtag the user can also view other current tweets with the same cashtag. Cashtags are a valuable resource for researchers and traders as they allow both to easily focus on financial news or discussion regarding specific companies. This specific information is also useful because at times a companies stock market movement might not mirror that of the stock market as a whole. We demonstrate that by using tweets with cashtags along with measures of text sentiment we can accurately model the stock movement of two highly traded stocks: Apple and Johnson and Johnson. We present our system, CashTagNN, which collects tweets with the appropriate cashtags, finds the polarity and subjectivity scores of these tweets using TextBlob, and then uses a neural network to predict opening and closing prices of the two stocks. We discuss our findings, and then conclude with future work.

II. Prior Work

As noted in the introduction, Twitter data has been used to study a wide range of topics such as flu trends, postpartum behavior, voting intentions, among many others. In regards to consume industry Chamerlat et al. have investigated how to predict product demands using twitter, and Jaring et al, among others [6] [7] [8], have all investigated how to market products on twitter effectively.

In regards to stock market prediction via tweets, Bollen et al. have shown that sentiment analysis can be used to predict stock market movement as a whole. They did not focus on tweets that specifically referenced the stock market, but rather they gathered a wide variety of tweets (approximately 9 million) which contained the phrases ‘I feel’, ‘I am feeling’, ‘I am feeling’,‘I don’t feel”, ‘I‘m’, ‘Im’, ‘I am’, and ‘makes me’. They then tested whether the mood of the public measured via these tweets would accurately predict stock market trading volumes on the Dow Jones Industrial Index. Their research concluded that 86.7% of the time there was a positive correlation between public sentiment and the upward/downward movement of the Dow Jones.

Souza et al find that the overall Twitter sentiment in regards to five retail companies (Abercrombie and Fitch, Nike, Home Depot, Mattel and Gamestop) has statistically significant relation with stock returns and volatility. Zheludev et al. also contrast the information content of both the sentiment of tweets and volume of tweets in terms of their influence on stock prices. They find that the sentiment of tweets contains significantly more lead-time information about the prices than the volume of tweets by itself, they look at a large array of financial instruments and find that only twelve are statistically significant.

Our work serves as a complement to Ranco et al. [12] who look at tweets with cashtags on the Dow Jones Index. Differences between our studies and theres include 1) they build an hourly times series whereas we look at opening and closing prices 2) sentiment is hand annotated and then used

978-1-5090-5781-8/16$31.00 ©2016 IEEE

Authorized licensed use limited to: Illinois Institute of Technology. Downloaded on October 24,2025 at 22:36:43 UTC from IEEE Xplore. Restrictions apply.

for broader classification, whereas we use automated sentiment analysis throughout 3) sentiment is negative, positive or neutral whereas we use a continuous sentiment designation 4) Pearson correlation and Granger causality tests is used for prediction whereas this work makes use of a neural network.

III. CashTagNN System

The CashTagNN System works in three stages: data collection and storage, sentiment assignment, and open and closing price prediction. In this section we will discuss each stage in detail. For further reference please view Figure 1 which highlights the system in graphical format.

A. Data Collection and Storage

The number of tweets collected per month is given below in Tables 1 and 2. The tweets specifically were collected between February 8 and April 12, 2016. In these tables you will notice that tweets have been divided into two categories: open and closing. Tweets created or shared before the stock market openings are marked as 'opening' by our system and will be used to predict opening prices. Tweets created between the opening and closing of the stock market are marked as 'closing' and will be used to predict the closing prices of that particular stock for the day.

Date	AAPL Closing	AAPL Opening
February 2016	4,846,727	2,657,627
March 2016	5,063,529	3,287,634
April 2016	1,222,219	919,472
Total	11,132,475	6,864,733

TABLE I
NUMBER OF AAPL TWEETS COLLECTED PER MONTH

Date	JNJ Opening	JNJ Closing
February 2016	1,045,021	2,669,606
March 2016	1,634,942	4,239,085
April 2016	460,006	1,052,608
Total	3,139,969	7,961,299

TABLE II
NUMBER OF JNJ TWEETS COLLECTED PER MONTH

In regards to data storage we used the NoSQL database MongoDB [13]. We chose MongoDB as we preferred to keep data in the json data format, and MongoDB works well with json. Before storing data in MongoDB we used twarc [14] to clean and filter out any information that was not needed. The database is indexed based on tweet date to increase lookup speed. We also collected and stored the opening and closing stock prices via Yahoo! Stocks for both companies.

B. Sentiment Analysis

The next step in the system is to find the sentiment of the tweets collected. TextBlob [15] was used to this end. TextBlob is a natural language processing package built on top of the NLTK (Natural Language Toolkit) framework. With regards to sentiment, we use two measures: subjectivity and polarity. Subjectivity is the measure of how subjective or conversely, objective, a text is. Polarity is the measure of how positive or negative the sentiment in the tweet is. In regards to the Textblob scoring system, a text's subjectivity ranges from [0, 1] with 0 being completely objective and 1 being completely subjective. The Polarity ranges from [-1, +1], with -1 being very negative and +1 being very positive. The data used to train TextBlob's sentiment and subjectivity analysis is 2888 words scored for polarity, subjectivity, intensity and reliability (this work uses only the polarity and subjectivity scores). The words are in general adjectives with no nouns present.

In regards to the stock market, an example of a negative tweet (with score of -0.025) would be: *@grayund The iphone was a decade long product. $AAPL cannot sustain 500 billion+ market cap over the next decade with iphone sales*. This tweet is also scored as slightly objective, with a score of 0.2. An example of a positive tweet, with a score of 0.29, would be *3,500% Gains On Stock Picks In 5 Months And An Amazing New Pick Is Coming Soon! Get It Early: https://t.co/Ikbnn3W31x $DO $ALL $AAPL*. This tweet was also scored as slightly subjective with a score of 0.55. In Table 3 and 4 below we give the numbers of positive/negative, subjective/objective tweets scored by the system.

Company / Event	# Positive	% Positive	# Negative	% Negative
AAPL / Opening	6,129,372	86%	603,481	14%
AAPL / Closing	10,009,752	91%	949,425	9%
JNJ / Opening	3,045,199	91%	124,585	9%
JNJ / Closing	7,787,813	96%	258,073	4%

TABLE III
NUMBER AND PERCENTAGE OF POSITIVE AND NEGATIVE TWEETS ASSIGNED.

Company / Event	# Subjective	% Subjective	# Objective	% Objective
AAPL / Opening	2,102,353	31%	4,630,500	69%
AAPL / Closing	3,523,977	32%	7,435,200	68%
JNJ / Opening	508,611	16%	2,661,173	86%
JNJ / Closing	1,033,408	12%	7,012,478	88%

TABLE IV
NUMBER AND PERCENTAGE OF SUBJECTIVE AND OBJECTIVE TWEETS ASSIGNED.

C. Neural Network Model

The last task for our system is to predict the opening and closing prices per company (Apple and Johnson and Johnson) based on the sentiment/subjectivity of tweets with the company's cashtags. The neural network is a feed forward neural net, made using the ffnet library [16]. As discussed

Authorized licensed use limited to: Illinois Institute of Technology. Downloaded on October 24,2025 at 22:36:43 UTC from IEEE Xplore. Restrictions apply.

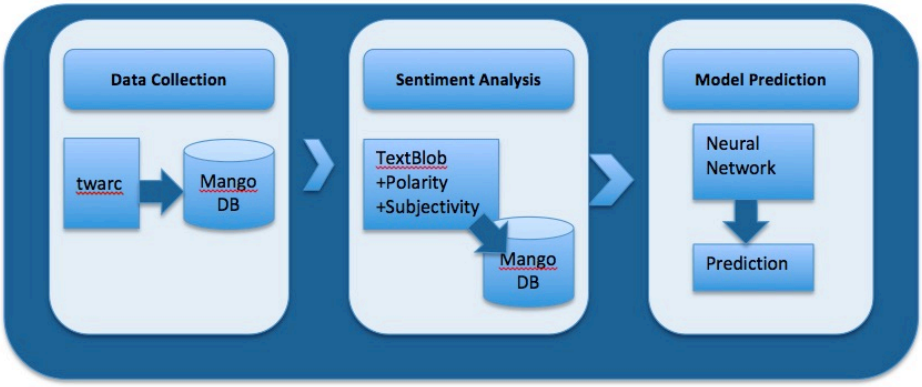


Fig. 1. CashTagNN System Figure

earlier, the database is parsed by date, segregating the tweets into two categories based on when the stock market open or closed. The sentiments scores from these two categories are averaged. The historical opening and closing stock quotes were manually downloaded from Yahoo! Stocks.

Therefore the training data for opening prices consists of the the averaged polarity and subjectivity score from 'opening' tweets. Likewise, the training data for closing prices consists of the averaged polarity score and the averaged subjectivity score from 'closing' tweets.

IV. Results

With the data given, we do have a regression model quite capable of predicting the general trend of a particular stock (see Table 5). The model in particular is good at predicting fluctuations (see Figures 2-5). Figures show the regression model's tight fit and its ability to use sentiment values to predict both opening and closing prices.

Event	R-squared	Max. Absolute Error
AAPL Opening	0.99	0.4
AAPL Closing	0.99	1.15
JNJ Opening	0.99	0.009
JNJ Closing	0.99	0.002

TABLE V
AAPL and JNJ Regression Results

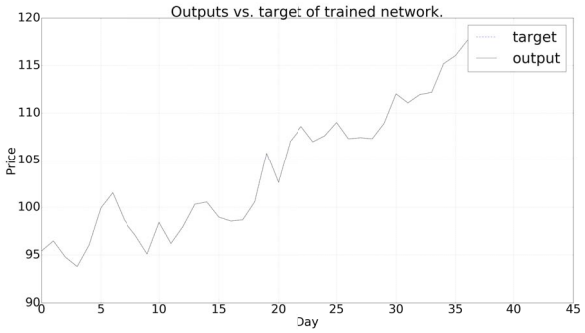


Fig. 2. AAPL Opening Results

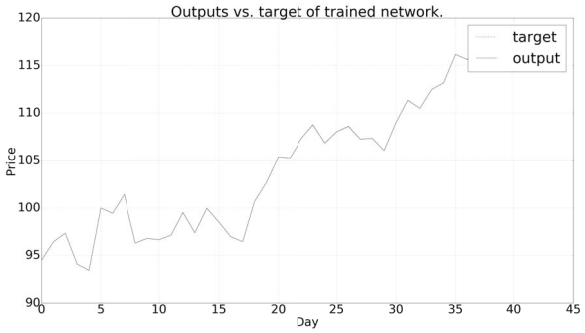


Fig. 3. AAPL Closing Results

V. Conclusion and Future Work

In this paper we present a system, CashTagNN, which uses the subjectivity and polarity of tweets with cashtags to predict opening and closing stock market prices. We focused on two companies: Apple and Johnson and Johnson and collected tweets containing their stock market cashtags between the dates of February 8 to April 15. We created a regression model between tweet sentiment and stock market prices which demonstrated high accuracy.

In the future, we plan to extend our system to all of the stocks on the NYSE and NASDAQ markets, to ensure that our model will be useful in regards to all publicly traded companies. We also plan to extend the system so that it will work in real time, and make it present on the web so that it can be used by the general public.

Authorized licensed use limited to: Illinois Institute of Technology. Downloaded on October 24,2025 at 22:36:43 UTC from IEEE Xplore. Restrictions apply.

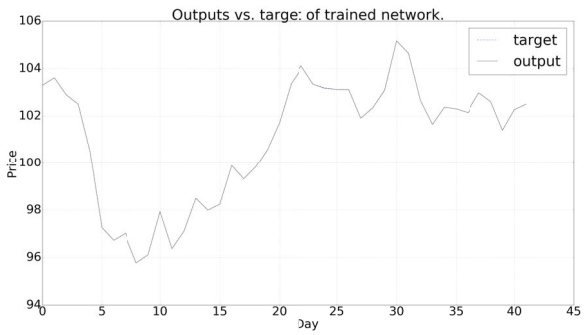


Fig. 4. JNJ Opening Results

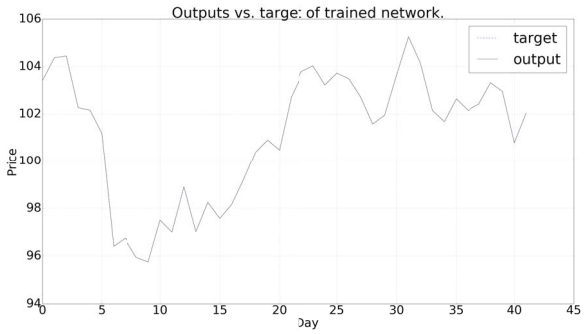


Fig. 5. JNJ Closing Results

REFERENCES

[1] A. Lamb, M. J. Paul, and M. Dredze, "Separating fact from fear: Tracking flu infections on twitter." in *HLT-NAACL*, 2013, pp. 789–795.

[2] M. De Choudhury, S. Counts, and E. Horvitz, "Predicting postpartum changes in emotion and behavior via social media," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2013, pp. 3267–3276.

[3] D. Gayo-Avello, "A meta-analysis of state-of-the-art electoral prediction from twitter data," *Social Science Computer Review*, p. 0894439313493979, 2013.

[4] W. Chamlertwat, P. Bhattarakosol, T. Rungkasiri, and C. Haruechaiyasak, "Discovering consumer insight from twitter via sentiment analysis." *J. UCS*, vol. 18, no. 8, pp. 973–992, 2012.

[5] P. Jaring, A. Bäck, M. Komssi, and J. Käki, "Using twitter in the acceleration of marketing new products and services," *Journal of Innovation Management*, vol. 3, no. 3, pp. 35–56, 2015.

[6] X. Y. Leung, B. Bai, and K. A. Stahura, "The marketing effectiveness of social media in the hotel industry a comparison of facebook and twitter," *Journal of Hospitality & Tourism Research*, vol. 39, no. 2, pp. 147–169, 2015.

[7] S. Goel, J. M. Hofman, S. Lahaie, D. M. Pennock, and D. J. Watts, "Predicting consumer behavior with web search," *Proceedings of the National academy of sciences*, vol. 107, no. 41, pp. 17 486–17 490, 2010.

[8] C. Dempster, D. S. Williams, and J. Lee, *The Rise of the Platform Marketer: Performance Marketing with Google, Facebook, and Twitter, Plus the Latest High-growth Digital Advertising Platforms*. John Wiley & Sons, 2015.

[9] J. Bollen, H. Mao, and X. Zeng, "Twitter mood predicts the stock market," *Journal of Computational Science*, vol. 2, no. 1, pp. 1–8, 2011.

[10] T. T. P. Souza, O. Kolchyna, P. C. Treleaven, and T. Aste, "Twitter sentiment analysis applied to finance: A case study in the retail industry," *arXiv preprint arXiv:1507.00784*, 2015.

[11] I. N. Zheludev, "When can social media lead financial markets?" Ph.D. dissertation, UCL (University College London), 2015.

[12] G. Ranco, D. Aleksovski, G. Caldarelli, M. Grčar, and I. Mozetič, "The effects of twitter sentiment on stock price returns," *PloS one*, vol. 10, no. 9, p. e0138441, 2015.

[13] K. Banker, *MongoDB in action*. Manning Publications Co., 2011.

[14]

[15] S. Loria, "Textblob: simplified text processing," *Secondary TextBlob: Simplified Text Processing*, 2014.

[16] M. Wojciechowski, "Ffnet: Feed-forward neural network for python,(2011)," *URL http://? ffnet.? sourceforge.? net/?, access date*, vol. 20, 2011.

Authorized licensed use limited to: Illinois Institute of Technology. Downloaded on October 24,2025 at 22:36:43 UTC from IEEE Xplore. Restrictions apply.