Can Twitter User's Moods Predict the Stock Market?

Aaron Gonzales and Adam Delora
Department of Computer Science, University of New Mexico

Background

- Twitter is a microblogging service with 284 million monthly active users who post 500 million updates ("tweets") per day.
- Latent Semantic Indexing[1] is a technique used to summarize words (documents) into representative ideas similar to principle component analysis.
- The AFINN semantic indexing database [2] assigns coded valence values to common words to allow quantification of a set of word's "mood".

Methods

We collected tweets using the public Twitter Streaming RESTful API over 2014-10-17 —2014-11-09 by tracking words that related to various tech stocks indexed by NASDAQ. Tweets were stored in a Mongo noSQL database.

- Maine nurse defies Ebola quarantine order by taking bike ride. http://t.co/eERINkm3AQ via @indystarJ
- Tim Cook: Apple CEO Says 'Being Gay Is One Of God's Greatest Gifts' Earlier today, the chief executive of Apple,... `http://t.co/zIXb2HbDmd`
- Meet the Swedish twin sisters who want to be 'identical artificial dolls' `http://t.co/77eA3esaQa`
- Last Christmas I got a black iPhone it was the worst Christmas present ever, I wanted a white one. I mean I'm not... `http://t.co/jhP9ODs3QK`

Fig. 1: Several examples of de-identified tweets from our dataset

nadella, twrt, amazon, amzn, prime, aws, fb, facebook, google, gmail,
ebola, aapl, apple, mac, tim cook, goog, youtube, microsoft, msft

Fig. 2: Kewords tracked

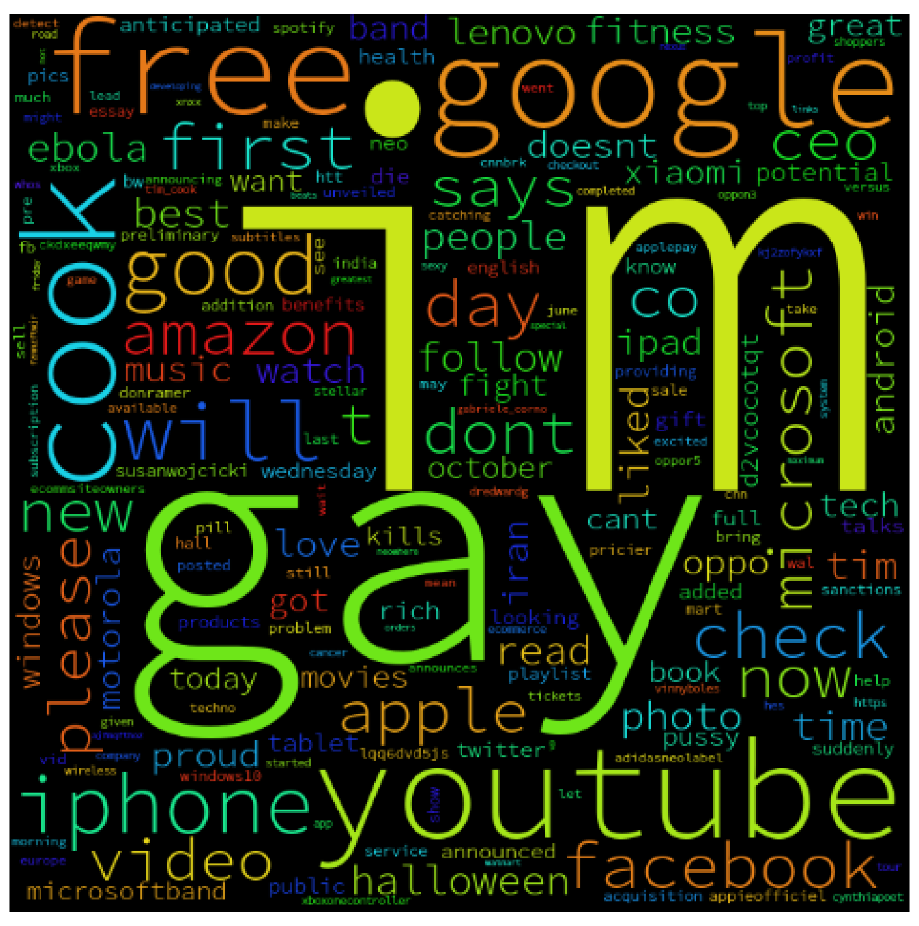
NASDAQ market data was collected over a slightly longer period, 2014-08-30 - 2014-11-09. Aaple (aapl), Amazon (amzn), Facebook (fb), Google (goog), Microsoft (msft), and Twitter (twtr) stocks were analyzed using their hourly closing price. Tweet text was preprocessed to remove common stopwords and punctuation and each hour bin was represented as a bag-of-words vector. Latent semantic indexing was performed on the one-hour bins of tweets, giving a total of 218 hours included in analysis. Each hour's LSI topics were scored using the AFINN database, resulting in a single number indicating semantic valence for each hour. The LSI score was smoothed using rolling means and assessed for periodicity. Semantic data was combined with the stock data and standardized for visualization (see Figure 5) and analysis. A model was created using vector autoregression (VAR) to assess the predictive power of the semantic data against the stock data. All work was done using Python: Tweets were harvested using Tweepy[3], text preprocessing was performed using Gensim [4], visualizations were made with Matplotlib[5], and statistical analysis was performed with Statsmodels [6].
Within an hour bin, LSI would provide a set of topics as noted in Figure 3. Representative Word clouds were generated for each hour as shown in Figure 4.

scientists	are	about	do	don't	over	say	climate	it	panic
0.416	0.219	0.217	0.216	0.214	0.211	0.209	0.209	0.209	0.208
on	is	get	from	ebola	i	google	it	really	liked
0.586	-0.315	-0.217	0.165	-0.140	-0.129	-0.126	-0.122	-0.117	0.208

Fig. 3: Two LSI topics from 2014-10-16, 14:00 - 15:00

Summary, Descriptive Statistics

- Total tweets collected: 84 million
- Total hours analyzed: 218
- Number of tweets per hour: 205,200; std 71,700
- Mean semantic score: 4.78; std 1.39



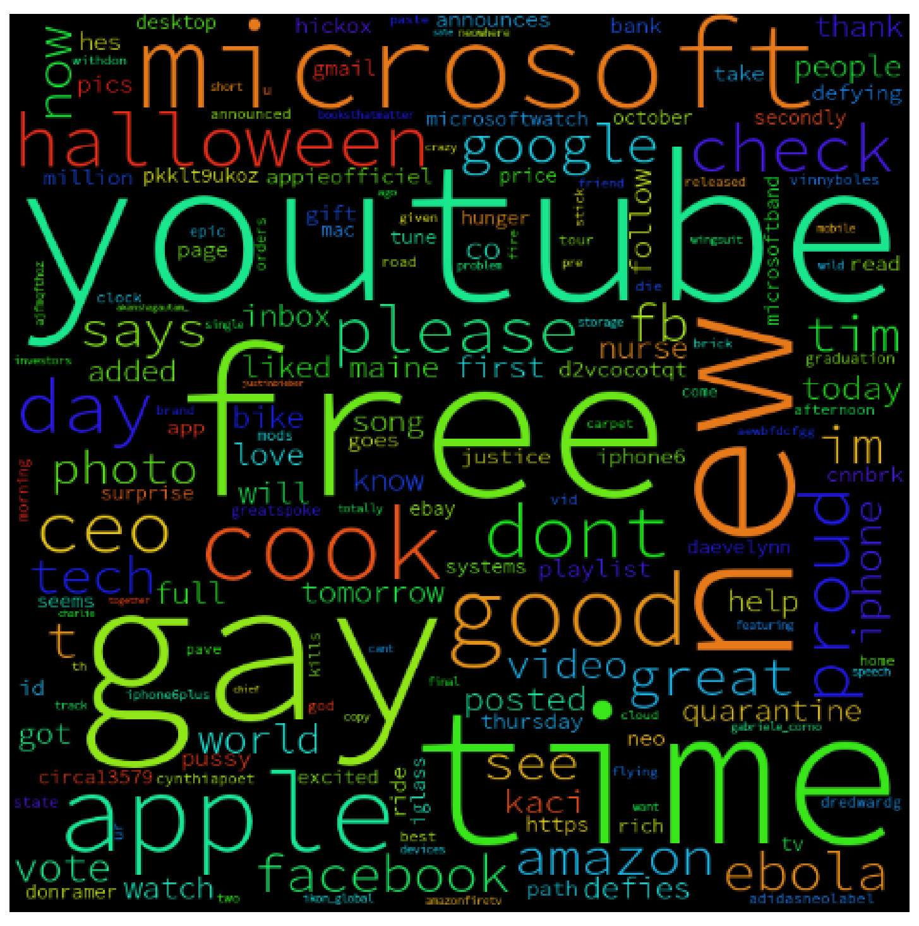


Fig. 4: Word clouds generated from topics on 2014-10-30, 11:00-12:00 and 12:00 - 13:00. The LSI model was able to capture information about Tim Cook (Apple's CEO) coming out as gay on that day.

Summary, Stocks and LSI score

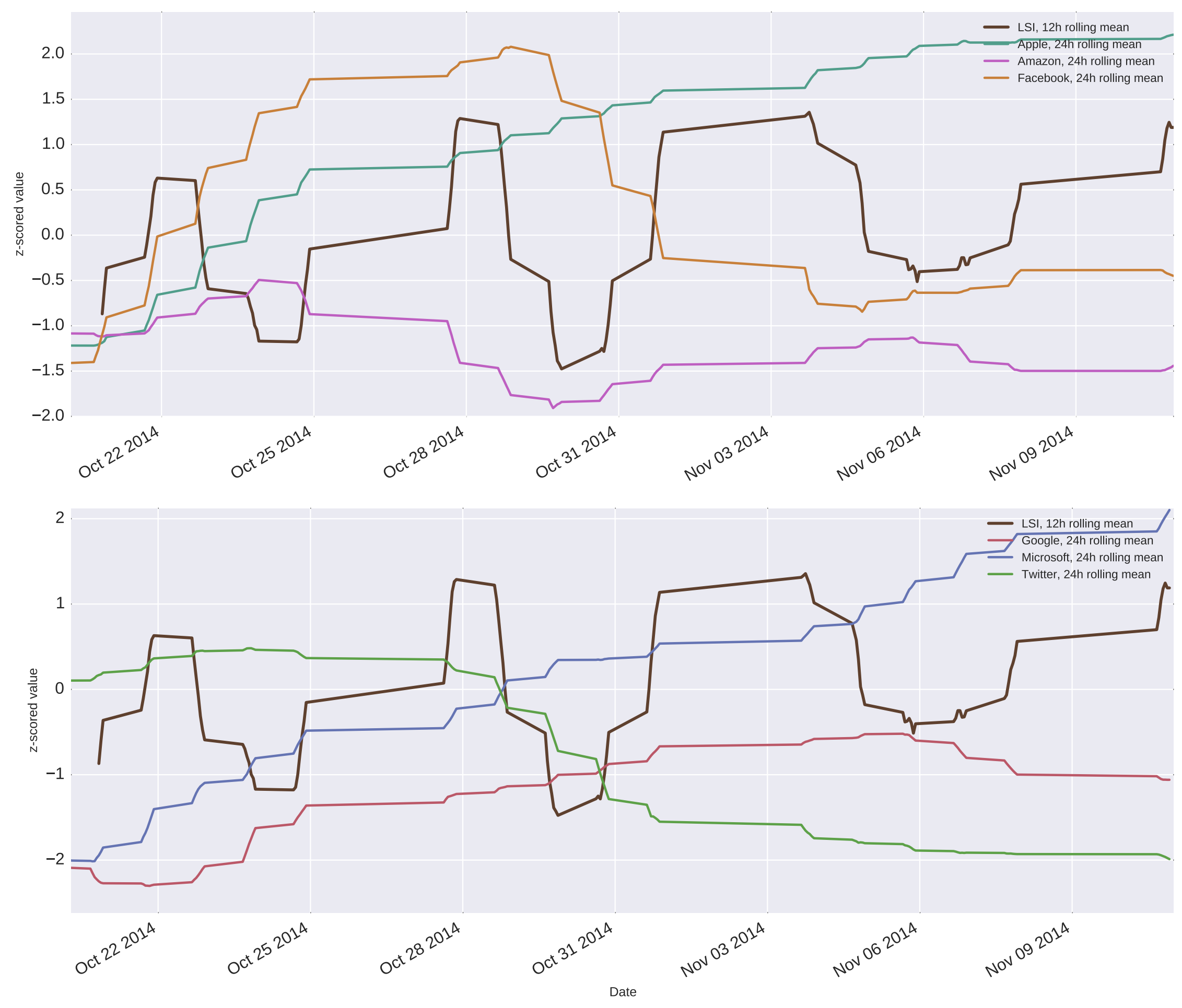


Fig. 5: Rolling means of standardized LSI score and closing stock prices per hour

Results and Conclusions

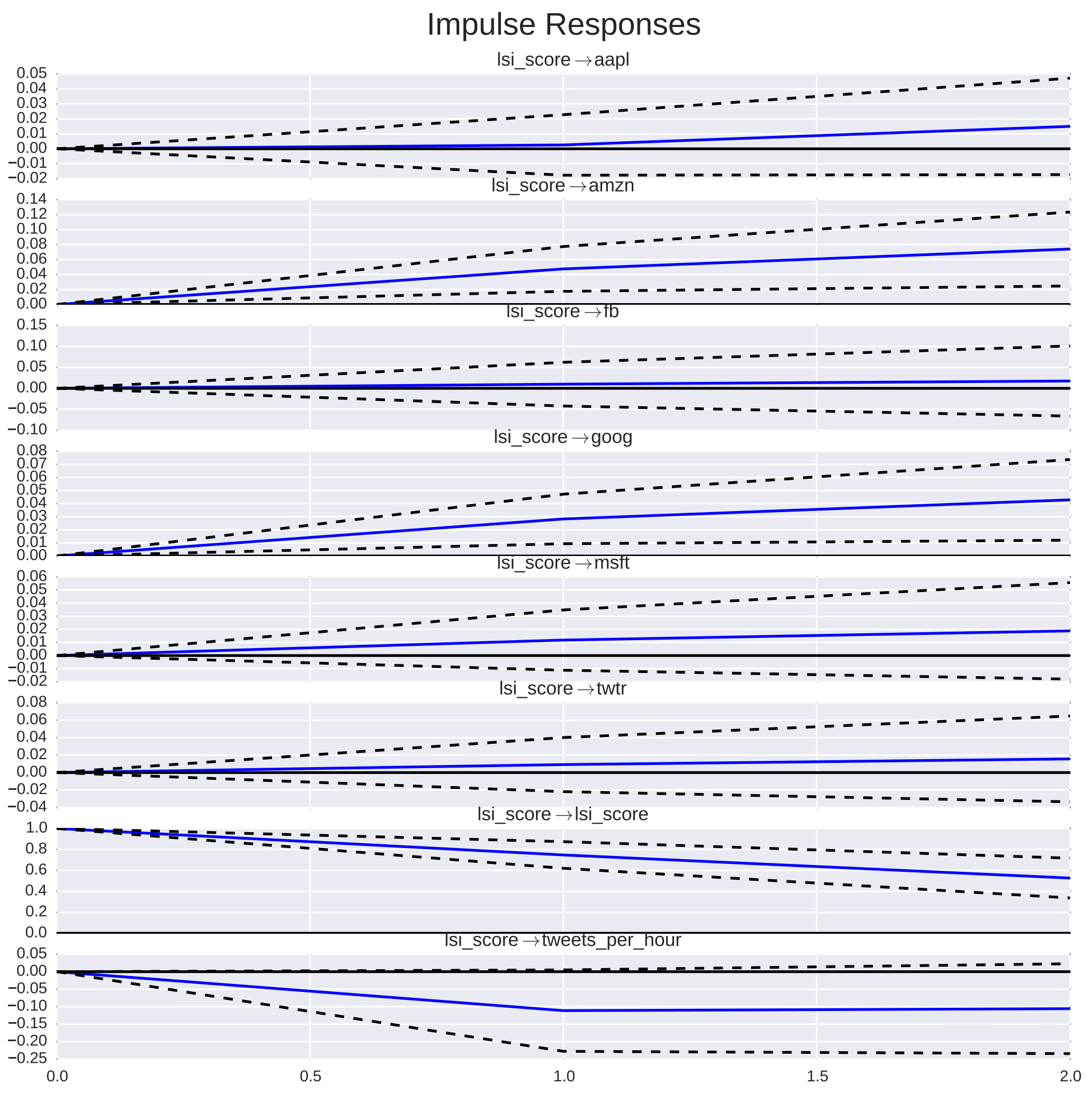


Fig. 6: Impulse Response analysis of LSI score change

VAR models describe a set of k variables as a linear function of their previous values. A p-th order VAR model is denoted by $y_t = c + A_1 y_{t-1} + A_2 y_{t-2} + \cdots + A_p y_{t-p} + e_t$. VAR models are often used with a lag parameter that operates on the elements of a time series to produce the previous element. We determined the lag order by employing an information criteria-based order selection which led us to use a lag of 1 hour in the model.
Impulse response analysis and a Granger's causality test were performed on the fitted VAR model. A causal effect of LSI score was found at the 90% level for the following variables in the model; impulse responses are plotted in Figure 6.

- Amazon stock price ($f = 9.68, p = 0.02$),
- Google stock price ($f = 8.584039, p = 0.003$),
- tweets per hour ($f = 3.51, p = 0.06$)

It is difficult to infer exactly why the semantic scores predicted some of these stocks, though there does seem to be a heavy overrepresentation of tweets that follow a form of "I liked a photo from Facebook" or "I liked a video on Youtube", which are automated tweets that some users have enabled in their accounts. Volume of these tweets may be a proxy for these service's usage rates, which could predict minor fluctuations in stock prices. It is useful to note that the fluctuations in prices are minor, 1-2% of total prices, but may not be so minor to investors. We will revisit this dataset after collection of broader data is complete in several months and will investigate other drawbacks of using the AFINN database and more effective filtering of the tweets.

References and Acknowledgements

We thank Dr. Abdullah Mueen for valuable discussions regarding our analyses.

References

[1] Thomas K Landauer, Peter W Foltz, and Darrell Laham. "An introduction to latent semantic analysis". In: *Discourse processes* 25.2-3 (1998), pp. 259–284.

[2] Finn Årup Nielsen. "A new ANEW: Evaluation of a word list for sentiment analysis in microblogs". In: *CoRR* abs/1103.2903 (2011). URL: `http://arxiv.org/abs/1103.2903`.

[3] Tweepy.com. *Tweepy*. 2014. URL: `https://github.com/tweepy/tweepy`.

[4] Radim ehek and Petr Sojka. "Software Framework for Topic Modelling with Large Corpora". English. In: *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. `http://is.muni.cz/publication/884893/en`. Valletta, Malta: ELRA, May 22, 2010, pp. 45–50.

[5] J. D. Hunter. "Matplotlib: A 2D graphics environment". In: *Computing In Science & Engineering* 9.3 (2007), pp. 90–95.

[6] J.S. Seabold and J. Perktold. "Statsmodels: Econometric and Statistical Modeling with Python". In: *Proceedings of the 9th Python in Science Conference*. 2010.