Learning Dow Jones From Twitter Sentiment

Au, Benjamin; Zhang, Qian; Zhang, Wanlu

Abstract

In 2010, Bollen used Twitter data to find high predictability of Twitter sentiment on the stock market. [1]. We hypothesized that while Bollen's results from analyzing the full breadth of the Twitter pipeline found significant results, fine-tuning the Twitter pipeline to only 'high-impact' financial tweets would improve the data signal and further improve results. As a result, we filtered a dataset of the Twitter pipeline for high impact tweets by user and financially-related 'short list' keywords and applied sentiment analysis on this filtered data and combined this in conjunction with DJIA stock market outcomes. Analyzing this data using logistic regression, SVM and time series analysis, we found modest outcomes, with predictability, peaking at 62.13%. While our filtered approach did not reach the levels claimed by Bollen, we showed substantial results in showing that applying appropriate *pre-filtering on Twitter data is necessary in running future analysis on the predictive power of the Twitter pipeline to maximize the sentiment signal of Twitter data.*

1. Introduction

In behavioral economics, market outcomes are affected by the sentiment of market agents themselves. Twitter, a platform publishing over 400 million tweets per day, seems to be a treasure trove of big data to mine to find an appropriate proxy for market sentiment. Bollen achieved impressive results from the sentiment of the Twitter pipeline on the Dow Jones Industrial Average in 6 dimensions. Motivated by these promising results, we suspected we could augment Bollen's analysis in a few ways. 1) We would filter the Twitter data to use only high-impact financially related tweets. Preliminary analysis of Tweet dataset quickly found that the vast majority of tweets were inane and utterly unrelated to the stock market. We suspected that proper filtering of Twitter data would reduce the risk of a 'garbage in-garbage out' low-signal-to-noise dataset. 2) We would apply SVM techniques to the filtered data to find an efficient decision boundary condition. 3) We would apply time series methods in our prediction. We were motivated by the potential of combining these tools to build on the results of Bollen.

2. Data

2.1 Source

Our dataset consisted of two sources. First, we have a slice of the full twitter pipeline, from June 11 to December 31, 2009, which consisted of about 476 million tweets [2]. Each tweets consisted of timestamp, username and tweet content. Our second source was daily closing prices of the Dow Jones Industrial Average for the same time duration as our Twitter dataset [3].

2.2 Preprocessing

Motivated by our desire to improve the signal of our dataset we pre-filtered our dataset for high-impact content as well as high-impact users. We generated a list of 131 high-impact finance-related Twitter users [4][5] and filtered twitter content for only those high-impact users. Individuals on this short list would likely tweet content finance-related and they would better approximate the sentiment of the stock market. Secondly, we filtered our dataset based on 20 high-impact finance-related keywords picked by ourselves. Filtering for these high-impact keywords would increase the signal of our dataset by obtaining only tweets related to the stock market in content, which in preliminary analysis, consisted of a small minority of overall tweets.

Further preprocessing techniques were performed to scrub tweet content, including making all content lowercase, removing all tweets that were not in English.

Using these preprocessing methods, we obtained a cleaner dataset with far less noise than the original. Given our original dataset of about 476 million tweets, our filtering did not pose a risk on overall sample size.

2.3. Sentiment Analysis

To obtain sentiment for each tweet in our filtered dataset, we used a pre-constructed Twitter Sentiment Analysis word list by Alex Davies to obtain dimensions of "happiness" and "sadness" of each tweet token. Overall sentiment for each tweet was taken based on averaging the sentiment for all applicable tokens in the sentiment word list. We would use these sentiment statistics as the basis of our sentiment analysis.

3. Machine Learning Models

Our goal was to use Machine Learning techniques to use sentiment data of a given day predict a binary change (positive or negative) on the DJIA closing price of the following day. Given our DJIA and tweet sentiment data, we performed several machine learning analyses, including logistic regression, SVM with linear, radial and sigmoid kernels, and applying time series analysis techniques in including previous day DJIA changes. We applied a few cross-validation techniques to train our algorithm, including 10-fold, 20-fold and Leave-one-out cross validation. Results are below. 4.1 shows results for machine learning analysis on tweets pre-filtered for high-impact Twitter users. 4.2 shows results for similar analysis for high-impact tweet content. 4.3 shows results for mixing both high-impact user and high-impact keyword content techniques.

4. Results and Discussion

Notations for this section:

- Time unit: day
- t – today, t+1 – tomorrow, t-1 – yesterday, etc.
- Happy_U, Sad_U represents sentiment value generated from high impact users
- Happy_W, Sad_W represents sentiment value generated from high impact tweet content
- LIBLINEAR and LIBSVM are SVM libraries
- For LIBSVM, I omitted the results of 10-fold and 20-fold Cross Validation and only kept the LOOCV results

Upon performing logistic analysis on a variety of flavors of sentiment and outcome-based models, we found that for high-impact user models, high-impact keyword content models and for combined models, the best sentiment model was Model 4.3d, which predicts DJIA(t) based on independent variables Happy_U(t), Sad_U(t), Happy_W(t), Sad_W(t), Happy_U(t-1), Sad_U(t-1), Happy_W(t-1), Sad_W(t-1), Happy_U(t-2), Sad_U(t-2), Happy_W(t-2), and Sad_W(t-2).

Using this model in conjunction with SVM and Leave-One-Out Cross Validation, we achieved our greatest predictive power: 62.32% for high-impact user models, for this combined model. This implies that applying time series instruments on previous day close DJIA close and sentiment is significant in boosting next day predictive power. See figures below for detailed results.

All models in each model type performed similar with ranges of no greater than 3% in performance. High-impact user model performed on the whole better than high-impact keyword models, with a mix of the two performing better than either or in isolation. In addition, because our data for the time period (140 trading days) is scarce, we just focused on the LOOCV result.

4.1 High-Impact User Results

The following tables show the results from using High-Impact User tweet filtering technique and applying logistic regression and SVM with linear, sigmoid and radial kernels on the resulting sentiment. We used a variety of flavors of happy/sad/DJIA-previous outcome to model DJIA outcome, and we collected results from 6 different models with different lags. The best model was Model 3:

DJIA(t+1) ~ Happy_U(t) + Sad_U(t) + Happy_U(t-1) + Sad_U(t-1) + DJIA(t),

using SVM with a linear kernel, which achieved predictive power of 59.7122%. The results of our basic model and best model are listed below:

(Basic) Model 1: DJIA (t+1) ~ Happy_U(t) + Sad_U(t)				
LIBLINEAR				
L2 Logistic	L1 Logistic	10-fold CV	20-fold CV	LOOCV
55.7143%	55.7143%	57.1429%	57.1429%	56.4286%
LIBSVM (LOOCV)				
Linear Kernel	Radial Kernel	Sigmoid Kernel		
55.7143%	53.5714%	55.7143%		

Figure 4.1a Results from Logistic Regression and SVM on High Impact User Model: DJIA(t+1) ~ Happy_U(t) + Sad_U(t)

Model 3: DJIA (t+1) ~ Happy_U(t) + Sad_U(t) + Happy_U(t-1) + Sad_U(t-1) + label(t)				
Logistic Regression		55.3957%		
LIBSVM (LOOCV)	Linear Kernel	Radial Kernel	Sigmoid Kernel	
	55.3957%	56.1151%	55.3957%	
LIBLINEAR	10-fold CV	20-fold CV	LOOCV	
	57.5540%	59.7122%	58.9928%	

Figure 4.1c Results from Logistic Regression and SVM on Time-Series Sentiment and Outcome on High Impact User Model: DJIA(t+1) ~ Happy_U(t) + Sad_U(t) + Happy_U(t-1) + Sad_U(t-1) + DJIA(t)

4.2 High-Impact Tweet Content Results

With the same idea as in 4.1, we obtain results from filtering tweets by high-impact content keywords. We performed similar analyses using Logistic Regression and SVM with linear, radial and sigmoid kernels; along with 10-fold, 20-fold and Leave-one-out cross validation techniques. The best results came from Model 4:

DJIA(t+1) ~ Happy_W(t) + Sad_W(t) + Happy_W(t-1) + Sad_W(t-1) + Happy_W(t-2) + Sad_W(t-2),

using SVM with a linear kernel and cross validation. The results of models based on data filtered by high-impact tweet content are not as significant as those in 4.1. This achieved predictive power of is 55.7971%. The results of our best model is listed below:

Model 4: DJIA (t+1) ~ Happy_W(t) + Sad_W(t) + Happy_W(t-1) + Sad_W(t-1) + Happy_W(t-2) + Sad_W(t-2)			
Logistic Regression		56.5217%	
LIBSVM (LOOCV)	Linear Kernel	Radial Kernel	Sigmoid Kernel
	55.7971%	55.7971%	55.7971%
LIBLINEAR	10-fold CV	20-fold CV	LOOCV
	54.3478%	55.7971%	55.7971%

Figure 4.2d Results from Logistic Regression and SVM on Extended Time-Series Sentiment on High-Impact Content Model: DJIA(t+1) ~ Happy_W(t) + Sad_W(t) + Happy_W(t-1) + Sad_W(t-1) + Happy_W(t-2) + Sad_W(t-2)

4.3 Combined User/Content Model Results

Finally, we combined the high-impact user and high-impact keyword content models from 4.1 and 4.2, using the same modeling techniques. The best results again came from time series Model 4:

DJIA(t+1) ~ Happy_U(t) + Sad_U(t) + Happy_W(t) + Sad_W(t) + Happy_U(t-1) + Sad_U(t-1) + Happy_W(t-1) + Sad_W(t-1) + Happy_U(t-2) + Sad_U(t-2) + Happy_W(t-2) + Sad_W(t-2).

This model achieved an accuracy of 62.3188%. The best result is listed below:

Model 4.3d: DJIA(t+1) ~ Happy_U(t) + Sad_U(t) + Happy_W(t) + Sad_W(t) + Happy_U(t-1) + Sad_U(t-1) + Happy_W(t-1) + Sad_W(t-1) + Happy_U(t-2) + Sad_U(t-2) + Happy_W(t-2) + Sad_W(t-2)			
Logistic Regression		57.2464%	
LIBSVM (LOOCV)	Linear Kernel	Radial Kernel	Sigmoid Kernel
	61.5942%	62.3188%	56.5217%
LIBLINEAR	10-fold CV	20-fold CV	LOOCV
	60.8696%	61.5942%	60.8696%

Figure 4.3d Results from Logistic Regression and SVM on Extended Time-Series Sentiment on High-Impact Content Model: DJIA(t+1) ~ Happy_U(t) + Sad_U(t) + Happy_W(t) + Sad_W(t) + Happy_U(t-1) + Sad_U(t-1) + Happy_W(t-1) + Sad_W(t-1) + Happy_U(t-2) + Sad_U(t-2) + Happy_W(t-2) + Sad_W(t-2)

5. Conclusion

We achieved modest results in our models, with models based on high-impact users achieving slightly greater predictive power relative to models based on high-impact finance-related keywords, and combined modeling achieving greater predictive power than either in isolation. Both individual models achieved between 55-60% predictive power, which combined modeling peaking at 62.32%. While, we were unable to match Bollen's results of 78.6%, given the lack of public availability of the sentiment algorithm used by Bollen, our results were significant in showing that combining both high-impact user modeling and high-impact keyword content modeling and noteworthy techniques in processing Twitter sentiment. We also conclude that including time-series instruments for previous changes in the market improved predictive power. If we can have more data, a multinomial classification will give more useful results.

However, by Jack Sparrow [6], "My best trader makes money only 63 percent of the time. Most traders make money only in the 50 to 55 percent range. That means you're going to be wrong a lot." Our project focuses on the market performance influenced by market sentiment, and gives us a good prediction (62.32% accuracy) of the market in the short-term, compared to the above statistic.

In addition, our project is not trying to build a trading strategy, but to establish a fundamental market prediction model, which can be widely applied. After the project, we may add other financial indices to our model and make it a trading strategy. In addition, our result can be applied in different researches and strategies related to financial market.

Further analysis would involve applying more advanced linguistic analysis techniques to better measure sentiment for Twitter content and apply those preprocessing techniques to our model for sentiment-based stock market prediction.

References

[1] Bollen, J., Mao, H. and Zeng, X.-J. 2010. Twitter mood predicts the stock market. Journal of Computational Science 2(1):1–8

[2] Yang J., Leskovec J. Temporal Variation in Online Media. ACM International Conference on Web Search and Data Mining (WSDM '11), 2011.

[3] Yahoo Finance. DJI Historical Prices. http://finance.yahoo.com/q/hp?s=%5Edji+historical+prices

[4] Business Insider. 101 Finance People You Have to Follow On Twitter. April 12, 2012.

[5] Wall Street Journal. Best Tweets for Your Money. March 8, 2013.

[6] Mercenary Trader. The Siren Song of High Winning Percentage. March 30, 2011. http://www.mercenarytrader.com/2011/03/the-siren-song-of-high-winning-percentage/

[7] Davies, Alex. Twitter Sentiment Analysis Word List. http://alexdavies.net/twitter-sentiment-analysis/