Predicting the Stock Market with 250,000,000 Tweets

Mike del Balso, Alex Litoiu

University of Toronto
Rogers Department of Electrical & Computer Engineering, University of Toronto

Introduction

In the year 2000, the value of all trades in **US financial markets exceeded $500 Trillion** (U.S. Census Bureau, 2001), over fifty times the GDP of the United States for that year. **A third of all stock trades were driven by algorithms .**

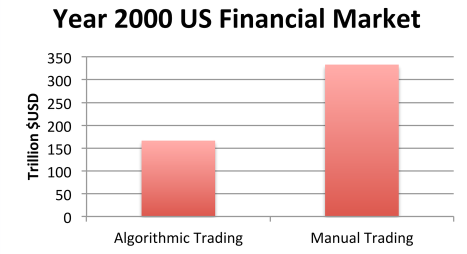

Figure 1: Year 2000 US Financial Markets

The Stock Market is greatly influenced by investor confidence and human emotion. This project analyzed 1% of the 250,000,000 million daily tweets for the past 100 days for human emotion in order to predict the stock market performance of various financial symbols.

Novelty: Our novel contributions are to experiment with identifying common pockets of sentiment using K-means clustering machine learning algorithms and drawing correlations to varied financial stocks and futures, as opposed to just a single index. There are few publications that accurately predict the stock market based on twitter. Those that claim to do so use questionable methods.

Functional Requirement: Predict whether a stock, index, futures contract, exchange traded fund (ETF), fixed-income security, indicator, or mutual fund will go up or down in a future time-frame (provide an uncertainty along with this prediction)

Objective: The lower the uncertainty, the better

Design

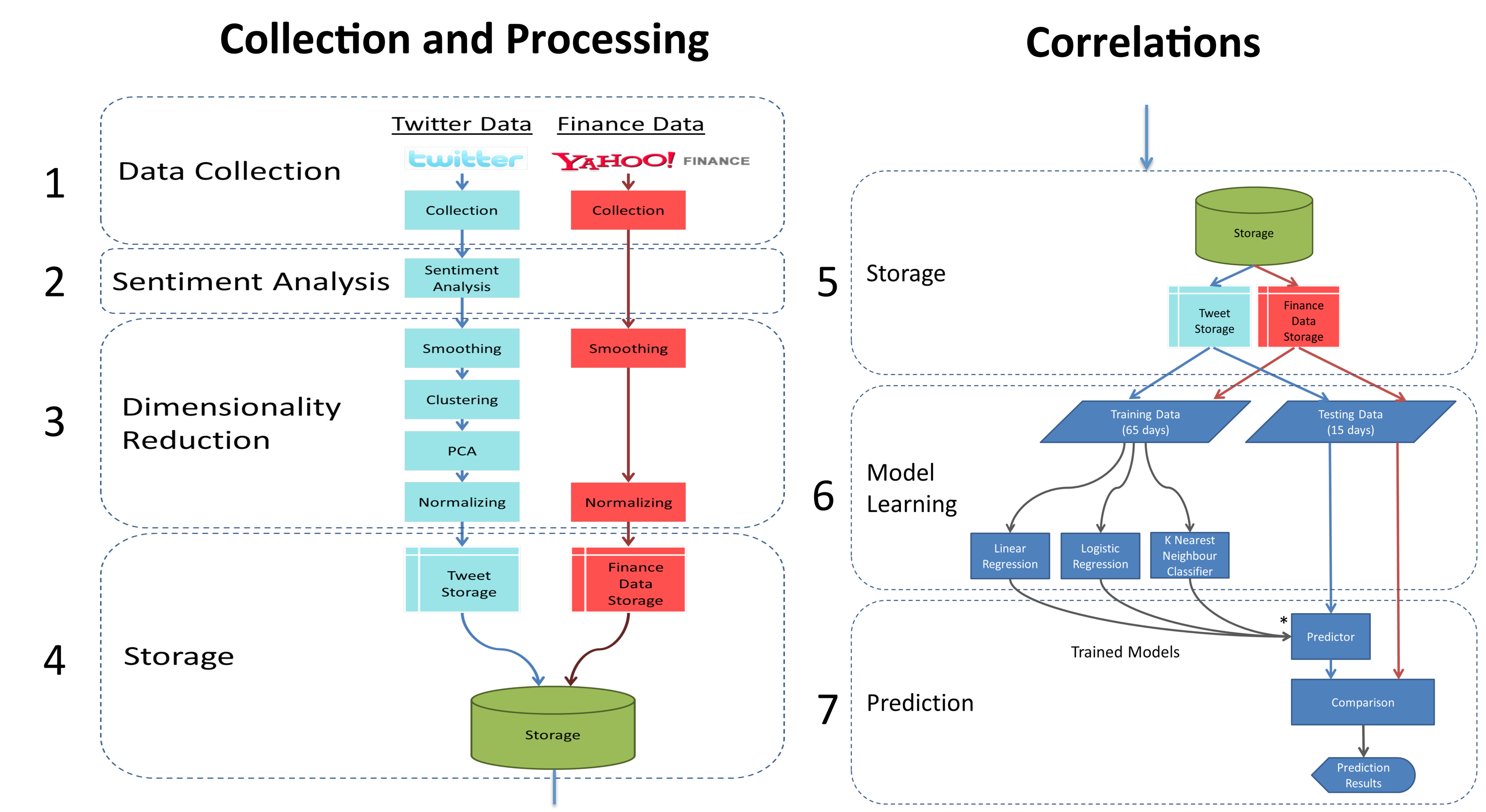

Figure 2: Project Design Diagram

1. **Data Collection –** Download 2.5 million tweets per day
2. **Sentiment Analysis –** Analyzes each tweet for 13 dimensions of sentiment: Tension, Depression, Anger, Fatigue, Vigor, and Confusion, etc. Assigns each dimension a numeric score
3. **Dimensionality Reduction -** extracts important and sometimes hidden signals of sentiment from the sentiment data
4. **Storage –** MongoDB database stores 450GB of Tweets. Holds the raw tweets, sentiment scores, clustering results, financial data, and all other intermediate results.
6. **Model Learning –** The predictive models are learned that are used by the predictor to create the stock price predictions. Models are learned based on training data. Testing data is then used by the prediction module.
7. **Prediction Module -** uses the derived Twitter data to make predictions about the stock price on the test data, the second partition of data mentioned in model learning module section. Uses linear and logistic regression and K-nearest neighbors prediction algorithm

CONCLUSIONS

- Collected 250 million tweets and followed 2747 stock symbols for 6 months
- Analyzed 3 months of data for correlations.
- Using K-Nearest Neighbor, achieved average 56% prediction rate across all 2747 stocks

Next Steps

- Confirm results on a larger data set
- Training data has different characteristics from testing data. We can likely achieve better results by extending the length of the experiment

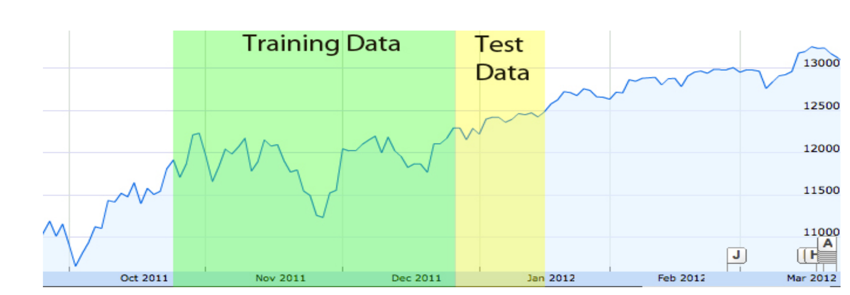

Figure 12: Figure of index fund performance over the date ranges tested

- All twitter users are not created equal. Filtering tweets by influence of user is likely to improve results
- Filtering tweets by subject matter is likely to improve prediction rate

TESTING AND VERIFICATION

All Modules were tested for correctness:

1. **Data Collection**
   a. **Twitter:** Tweets manually verified for correctness; graph of tweets collected per day
   b. **Finance:** Manually checked for correctness
2. **Sentiment Analysis**: Tweets analyzed in all dimensions by humans, and compared to machine values
   a. **Sentiment v1:** Tested manually, and determined to be noisy. Process repeated as Sentiment v2.
   b. **Sentiment v2:** Compared human ratings to predicted ratings.
3. **Dimensionality Reduction**: Checked for convergence, and visualized results
   a. **Clustering**
   b. **Grouping**

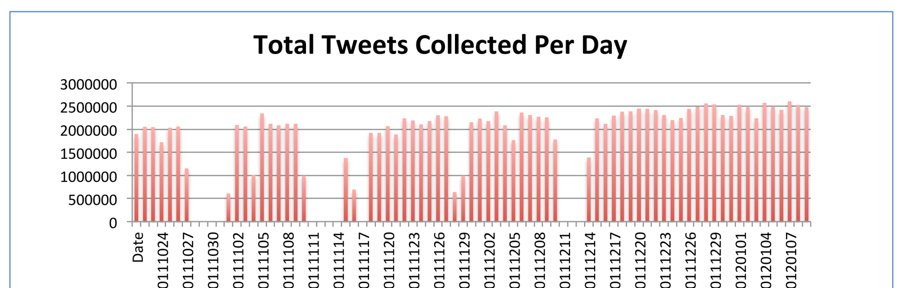

Figure 3: Module 1a – Total Tweets Per Day

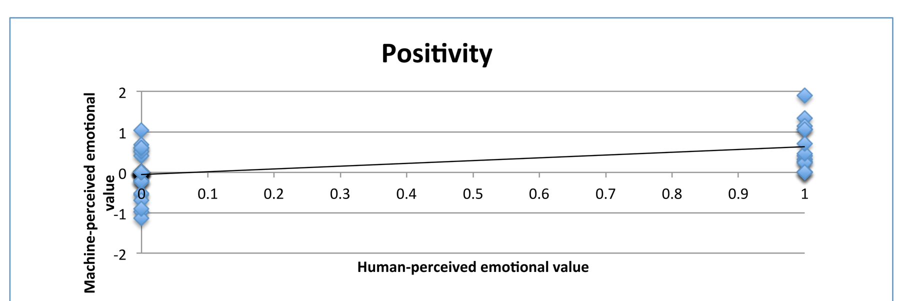

Figure 4: Module 2b – Positivity dimension in sentiment analysis

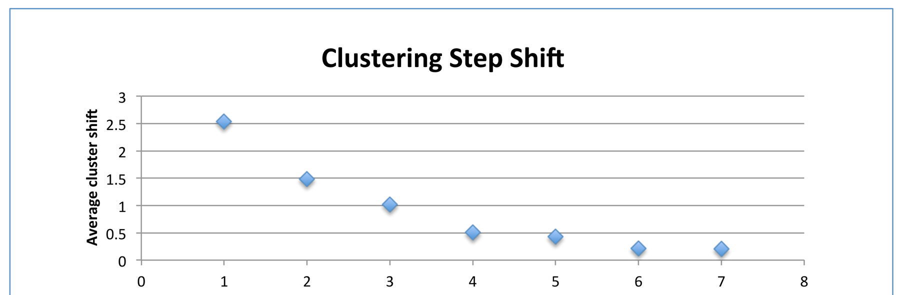

Figure 5: Module 3a – Clustering average shift distance

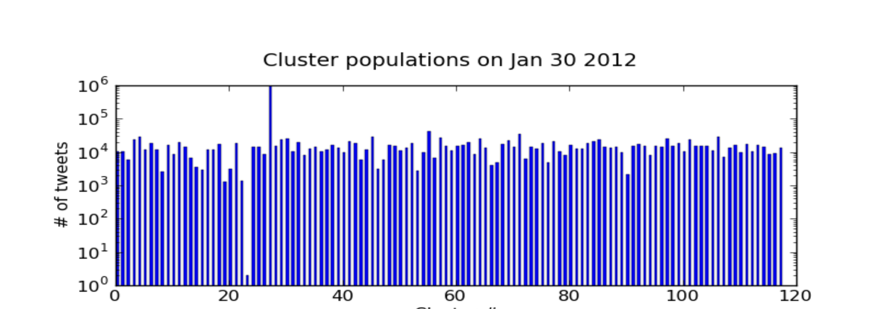

Figure 6: Module 3b – Number of tweets in each group on Jan 30, 2012

6 Model Learning and 7 Prediction Module results in the Results Section

RESULTS

Sentiment Version 1

- Tested 11 financial symbols
- Training Data: October 22nd 2011 to December 26th 2011 (65 days).
- Testing data: December 27th to January 10th 2012 (15 days).
- Twitter Sentiment v1 created noisy sentiment dimensions.
- When trying to predict a stock's performance based on the number of tweets in each group, the average predictive rates were **41%**

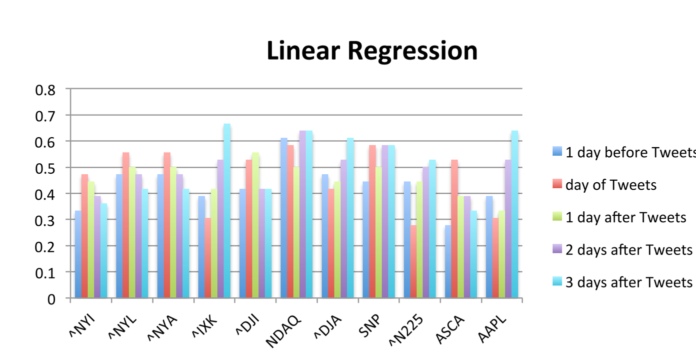

Figure 7: Linear Regression Predictive Rates

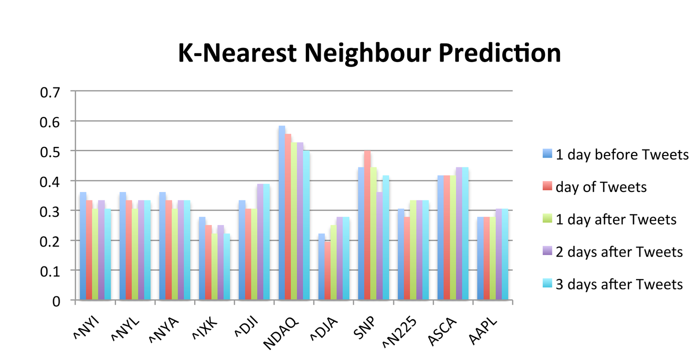

Figure 8: K-Nearest Neighbor Predictive Rates

Figure 9: Logistic Regression Prediction Rates

Sentiment Version 2

- Tested 2747 financial symbols
- Training Data: October 22nd 2011 to December 26th 2011 (65 days).
- Testing data: December 27th to January 10th 2012 (15 days).
- When trying to predict a stock's performance based on the number of tweets in each group, the average predictive rates were **54%.**
- Across all stocks, K-nearest Neighbor achieved **56%** prediction rate across all offsets, and **57.2%** prediction 1 day in advance

	Logistic	K-Nearest	Linear
0 days	0.556	0.576	0.505
1 day	0.547	0.572	0.567
2 days	0.544	0.562	0.492
3 days	0.532	0.537	0.492
Average	**0.544**	**0.561**	**0.502**

Figure 10: Predictive rates of all methods 0,1,2 or 3 days in advance

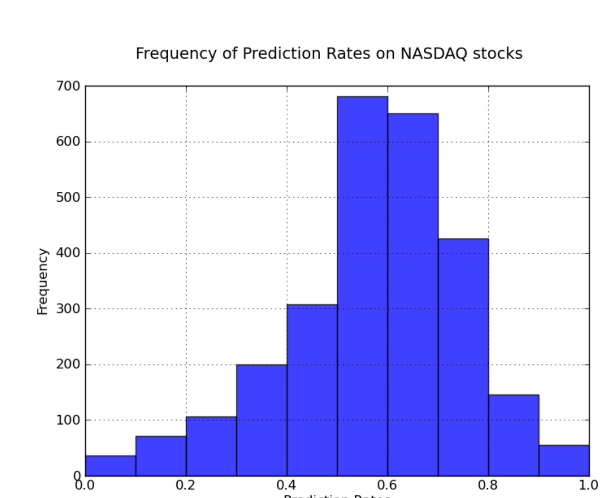

Figure 11: Frequency of predictive rates for K-Nearest Neighbor, 1 day in advance

REFERENCES

1. U.S. Census Bureau. (2001). *Statistical abstract of the United States: 2001.* U.S. Department of Commerce. Washington D.C.: United States Government.
2. 10gen, Inc. (n.d.). *mongoDB*. Retrieved 02 14, 2012, from http://www.mongodb.org/
3. Aite Group. (2011). *Algorithmic Trading in FX: Ready for Takeoff?* New York: Aite Group.
4. Facebook. (n.d.). *Statistics of Facebook.* Retrieved Octovwe 30, 2011, from Facebook: https://www.facebook.com/press/info.php?statistics
5. Gimpert, B. (2011, May 13). *Sour Grapes: Seven Reasons Why "That" Twitter Prediction Model is Cooked.* Retrieved February 14, 2012, from Some Ben?: http://blog.someben.com/2011/05/sour-grapes-seven-reasons-why-that-twitter-prediction-model-is-cooked/
6. Johan Bollen, H. M.-J. (2010). Twitter mood predicts the stock market. *Computing Research Repository , abs/1010.3003*.
7. Joliffe, I. (1986). *Principle Component Analysis* (2 ed.). New York: Springer.
8. Nielsen, F. Å. (2011, March). A new ANEW: Evaluation of a word list for sentiment analysis in microblogs. *Computing Research Repository* .
9. Maurice Lorr, P. D. (n.d.). *Profile of Mood States.* Retrieved February 14, 2012, from MULTI-HEALTH SYSTEMS INC: http://www.mhs.com/product.aspx?gr=cli&id=overview&prod=poms#description
10. Petzoldt, D. (2011, February 11). *Statistical flaws in "Twitter mood predicts the stock market" research paper.* Retrieved February 14, 2012, from http://petzoldt.tumblr.com/post/3236488086/statistical-flaws-in-twitter-mood-predicts-the-stock
11. Rosenberg, D. (2010, April 20). *What's (technically) in your tweets?* Retrieved February 14, 2012, from CNet News: http://news.cnet.com/8301-13846_3-20002924-62.html
12. Twitter. (n.d.). *Streaming API Methods*. Retrieved 02 14, 2012, from https://dev.twitter.com/docs/streaming-api/concepts
13. Twitter: @twittereng. (2011, June 30). *200 million Tweets per day.* Retrieved February 14, 2012, from Twitter Blog: http://blog.twitter.com/2011/06/200-million-tweets-per-day.html

ACKNOWLEDGEMENTS

- **Professor Brendan Frey** for his guidance with machine learning techniques and allowing us access to his lab computer throughout the duration of the term.
- **Professor Hans Kunov** for giving us great guidance with respect to project management and always pushing us in the right direction when we needed it.
- **Dan Astoorian** for helping us with special permissions on the ECF network and always making time to accommodate us and to troubleshoot.