Using Twitter as a source of information for stock market prediction

Ramon Xuriguera
(rxuriguera@lsi.upc.edu)

Joint work with Marta Arias and Argimiro Arratia

ERCIM 2011, 17-19 Dec. 2011, University of London, UK

UPC BarcelonaTech

Motivation and Goals

Vast amounts of new information **every ~~day~~ second** available in social networking platforms.

Can some of this information help improve time series' predictions for certain stocks?

Hypotheses

- Volume
  More messages $\rightarrow$ More variance (Volatility)
- Sentiment
  More positive/negative $\rightarrow$ Increase/Decrease benefits (Returns)

Motivation and Goals
Related Work

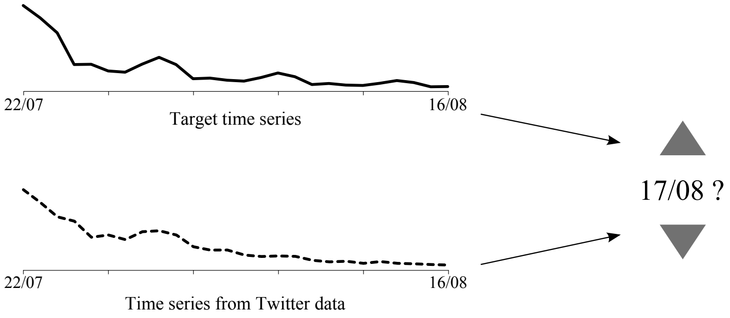
J. Bollen, et al. *Twitter mood predicts the stock market* [5]

- Assesses **general mood** by checking whether messages contain certain words (OpinionFinder and Google Profile of Mood States)
- Predict time series direction with a self-organizing Fuzzy Neural Network
- Twitter dataset from Feb. to Dec. 2008.

Our approach

- Different sentiment classifier
- Target different companies and indexes by only looking at tweets related to them.
- Test with other models

Motivation and Goals
The big picture



My right sock has THor. My left has Iron-Man. I'm missing **Captain America** pants however.

@Stormylyric I wnt to see tht Super 8 and I ain't sure bout Green Lantern. F*** **Captain America**

I need a Marvel obsessed friend to see Thor and **Captain America** with.

Captain America movie looks way better than the Thor movie.

Twitter

@TEDNews
TED News

RT @TEDchris: Mind-shifting #TED talk on the evolution of language from Mark Pagel http://on.ted.com/Pagel

3 Aug via TweetDeck

- Social networking and microblogging service
- Messages or *tweets* up to 140 characters
- Hashtags, retweets, mentions, URLs

Twitter

Data Retrieval

Lack of standard public datasets.
It is not allowed for third parties to redistribute Twitter Content.
We were forced to create our own dataset.

Twitter Streaming API

- An HTTP connection is kept alive to retrieve tweets as they are posted. Tweets are filtered by keyword or author.
- Limited amount of data. Can't go back in time.
- Began listening to Twitter's stream in March 2011.

Buying data from **official tweet resellers**

- **Extremely expensive**

Twitter

Relevance Filter

Latent Dirichlet Allocation

- Text as mixtures of (two) topics (*relevant / not relevant*)
- 300K tweet collection. Tested with 20K tweets
- High precision (83.3%), low recall (65.4%)

Sentiment Analysis

Determine whether a message contains **positive or negative opinions** on a given subject

A sentiment classifier is trained with an automatically labelled dataset where **smileys are used as labels**: [1, 2, 3]

`:-)` `:-D` $\rightarrow$ positive `[:=8][ -]?[)D]`
`:-(` $\rightarrow$ negative `[:=8][ -]?(`

How do we go about providing text to a probabilistic classifier?

- Bag of Words
  Vector of occurrences/frequencies of the words in a text.

Sentiment Analysis

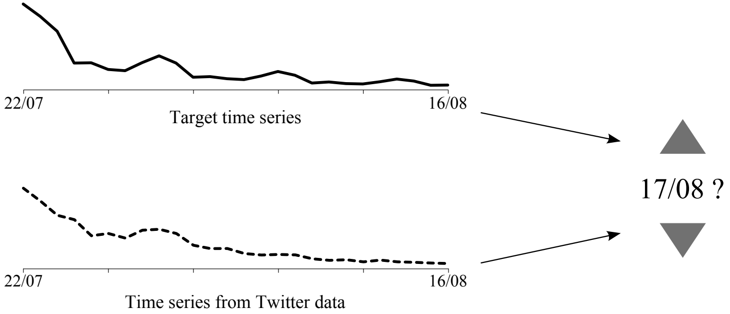
Classifier

- Multinomial Naïve Bayes
  Will assign a tweet, the sentiment with the highest conditional probability given that tweet
- Binary Classification (positive/negative)
- 82.5% for twittersentiment set

Sentiment Index:

- Time series of the daily percentage of positive tweets concerning a company

Sentiment Analysis
The big picture



My right sock has THor. My left has Iron-Man. I'm missing **Captain America** pants however.

@Stormylyric I wnt to see tht Super 8 and I ain't sure bout Green Lantern. F*** **Captain America**

I need a Marvel obsessed friend to see Thor and **Captain America** with.

Captain America movie looks way better than the Thor movie.

Forecasting Financial Time Series
Goal

Focus on companies: AAPL, GOOG, MSFT, YHOO
Two indices: OEX (S&P100) and GSPC (S&P500)
And their implicit volatility: VXO and VIX

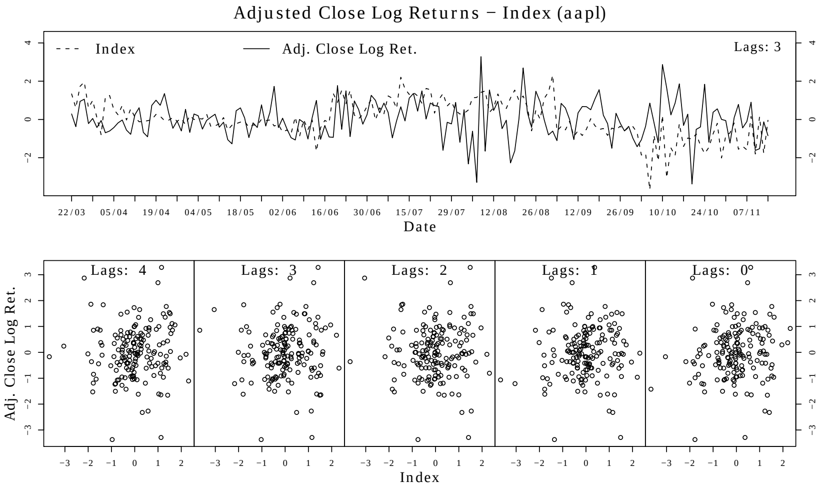
Price Returns:

- Adjusted Close
- Log-normally distributed
- Logarithmic Returns

Volatility:

- Computed from price returns
- Exponetially Weighted Moving Average

Forecasting Financial Time Series



Forecasting Financial Time Series

Model Adequacy

- Nonlinearity
  Test for nonlinear relationship between two series. [6]
- Causality tests
  Granger test of causality (parametric and non-parametric [7]) for the two time series and different lags
- Prequential evaluation
  Too few data to split into sets. Use all past data to re-train, then predict the direction of the series for the following day.

Models for prediction

- Linear Regression
- Feed-forward Neural Networks
- Support Vector Machines

Results

Compared **Accuracy** and **Directional Measure** for predictions **using** and **not using** Twitter data.

Large DM $\rightarrow$ Model outperforms chance of random choice

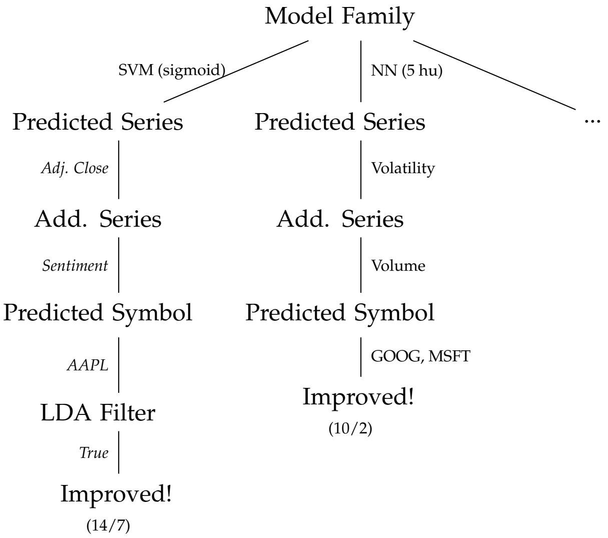
Predicted	**R.Filter**	**Model**	**Lags**	...	**Acc. w/o**	**Acc. w/**	**DM w/o**	**DM w/**
AAPL	TRUE	SVM(sigmoid)	2	...	0.560	0.628	2.398	10.739
AAPL	TRUE	SVM(sigmoid)	3	...	0.521	0.601	0.296	6.629
AAPL	TRUE	SVM(sigmoid)	4	...	0.537	0.635	0.835	11.932
...	...	...	...	...	...	...	...	...

Tested many parameter configurations and we end up with **thousands** of rows of results:

Need to summarise

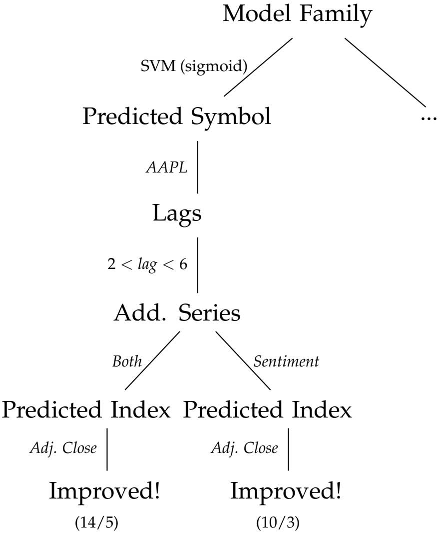
Results

Decision tree. Improved Accuracy (Weka's REPTree)



Results

Decision tree. Improved DM (Weka's REPTree)



Results

Some of the machine learning models we have evaluated, especially Support Vector Machines, present improvements in their predictive power when reinforced with our Twitter Sentiment Index

References I

[1] A. Go, R. Bhayani, L. Huang
Twitter Sentiment Classification using Distant Supervision,
2009.

[2] A. Bifet, E. Frank
Sentiment Knowledge Discovery in Twitter Streaming Data,
Discovery Science, 2010.

[3] J.K. Ahkter, S. Soria
Sentiment Analysis: Facebook Statuses Messages,

[4] R.S. Tsay
Analysis of Financial Time Series (pp. 216)
2002

References II

[5] J. Bollen, H. Mao and Xiao-Jun Zeng
Twitter mood predicts the stock market,
Journal of Computational Science, Vol. 2, Iss. 1, 2011.

[6] T. H. Lee, H. White, C. W. J. Granger
Testing for neglected nonlinearity in time series models
Journal of Econometrics, Vol. 56, Iss. 3, 1993.

[7] C. Diks, V. Panchenko
A new statistic and practical guidelines for nonparametric Granger causality testing
Journal of Economic Dynamics and Control, Vol. 30, Iss. 9-10, 2006.

Directional Measure

Given a contingency table with predicted results,

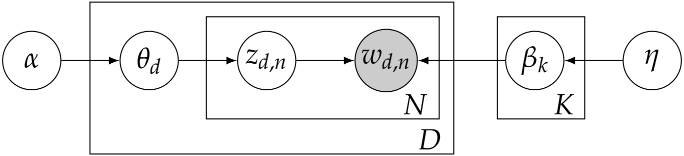
Actual	Predicted		
	Up	Down	
Up	m_{11}	m_{12}	m_{10}
Down	m_{21}	m_{22}	m_{20}
	m_{01}	m_{02}	m

The directional measure can be computed as:

$$\chi^2 = \sum_{i=1}^{2} \sum_{j=1}^{2} \frac{(m_{ij} - m_{i0}m_{0j}/m)^2}{m_{i0}m_{0j}/m}$$

Under some circumstances, it behaves like a χ^2 distribution with 1 degree of freedom [4]. Assuming a 5% error, we need to have $\chi^2 > 3.84$.

A bit more about LDA



- Generative model
- Words are only observed variables
- Topics are distribution over words
- Dirichlet: distribution over distributions