CS229 Project Report

Using Newspaper Sentiments to Predict Stock Movements

Hao Yee Chan *Anthony Chow*
haoyeec@stanford.edu ac1408@stanford.edu

Problem Statement

It is often said that stock prices are determined by market sentiments. Also, these stock prices are an instant reflection of the current market sentiments. Despite this, investors often use current news to inform their next investment decision. The problem is then that it is almost impossible to read through all the news available online. Even with a wealth of readings, market sentiments are difficult to be quantified and understood.

This project looks at news from Reuters Technology to be used as sources of data to generate a model to capture market sentiments. This model will be used to try to predict the movements of the NASDAQ Composite in the immediate future.

Dataset

We scrapped data Reuters to create a model for market sentiments. We will be using yahoo share prices to construct our classifiers, which will be the NASDAQ Composite, which is highly followed in the US as an indicator of the performance of stocks of technology companies and growth companies. (*see appendix for screenshot of Reuters Technology*) We created 2 separate datasets, one with 1 year of data (260 days of trading) and the second one with 3 years of data (690 days of trading). Note that there are less days of trading than there are in a year due to market closing during weekends as well as during public holidays.

Problem Formulation

We split the news data from Reuters Technology into headline and body feature sets. We aim to predict, given today's set of headline and body features, if the closing price of tomorrow's stock will be higher or lower than today's, using data from yahoo finance.

$$prediction = \begin{cases} 1 & if\ stock\ moves\ upwards \\ -1 & otherwise \end{cases}$$

We did some preprocessing to build our dataset. From the news, we first split them into 2 sets: Headline and body. For each set, we had stop words removed, the words lemmatized and selected 500 words using the following heuristics:

Bag-of-Words	Chi-squared	Mutual Information
Most frequent words	$X^2 = \frac{(N_{11}+N_{10}+N_{01}+N_{00})\,(N_{11}N_{00}-N_{10}N_{01})^2}{(N_{11}+N_{01})(N_{11}+N_{10})(N_{10}+N_{00})(N_{01}+N_{00})}$ where N_{tc} =Number of news articles with word t and class c	$MI = \sum_{y\in Y}\sum_{x\in X} p(x,y) log \frac{p(x,y)}{p(x)p(y)}$

Table 1: Feature selection techniques used

Our Methodology

We used a mixture of supervised and unsupervised machine learning techniques to figure out our data. Under the unsupervised techniques, we used factor analysis with EM to look at some of the key dimensions that described the data. We also compared the performance of the various supervised learning algorithms on the dataset. A summary of the supervised learning algorithms implemented in this paper is summarized in the table below.

Multinomial Naïve Bayes	Gaussian Discriminant Analysis	Support Vector Machines
Headline features (1 year)	-	Headline features (1 year)
Body features (1 year)	Body features (1 year)	Body features (1 year)
Headline features (3 years)	-	Headline features (3 years)
Body features (3 years)	Body features (3 years)	Body features (3 years)

Table 2: Table of summary of supervised learning algorithms implemented

Our Results

Factor Analysis of Data

We implemented factor analysis on the body feature sets. We present the results obtained from the body feature sets (1 year and 3 years) generated from frequent words. We find that there seem to exist 3 main dimensions in the data – Finance, Facebook and Apple that characterized the news from last year. For the 3 years data, it seemed to be – Finance, Apple and everyone else. Perhaps the Facebook IPO in this past year generated enough coverage to create this unique dimension in the data. Armed with the idea that there were special dimensions in the data, that it was not as random as we thought, we went to perform supervised learning with more confidence.

Factor Analysis using EM (Data Visualization)

Frequent Words, 1 year, body feature set

Dim 1 (apple and 2011)	Dim 2 (facebook and 2012)	Dim 3 (finance)
'apple'	'said'	'percent'
'said'	'company'	'company'
'job'	'billion'	'said'
'new'	'year'	'share'
'company'	'would'	'billion'
'iphone'	'percent'	'million'
'would'	'olympus'	'quarter'
'market'	'million'	'year'
'product'	'share'	'revenue'
'year'	'u'	'analyst'
'analyst'	'reuters'	'market'
'one'	'business'	'business'
'people'	'last'	'reuters'
'rim'	'also'	'profit'
'phone'	'investor'	'earnings'
'technology'	'facebook'	'cent'
'percent'	'new'	'investor'
'u'	'2012'	'sale'
'reuters'	'firm'	'forecast'
'also'	'could'	'loss'
'could'	'online'	'expected'
'sale'	'est'	'new'
'google'	'market'	'per'
'world'	'group'	'would'
'device'	'deal'	'maker'
'4'	'corp'	'wednesday'
'steve'	'executive'	'growth'
'many'	'internet'	'firm'
'mobile'	'stock'	'inc'
'billion'	'japanese'	'yen'
'samsung'	'board'	'street'
'time'	'one'	'also'
'2011'	'network'	'corp'
'service'	'editing'	'oct'
'editing'	'revenue'	'stock'

Factor Analysis using EM (Data Visualization)

Frequent Words, 3 years days, body feature set

Dim 1 (finance)	Dim 2 (apple)	Dim 3 (everyone else)
'said'	'apple'	'said'
'company'	'said'	'company'
'apple'	'company'	'year'
'share'	'job'	'reuters'
'year'	'year'	'new'
'quarter'	'one'	'service'
'million'	'new'	'mobile'
'sale'	'reuters'	'also'
'revenue'	'share'	'phone'
'new'	'investor'	'network'
'2011'	'also'	'internet'
'reuters'	'could'	'could'
'profit'	'service'	'million'
'tablet'	'two'	'one'
'expected'	'computer'	'government'
'last'	'steve'	'technology'
'first'	'chief'	'world'
'forecast'	'say'	'sale'
'investor'	'executive'	'say'
'margin'	'million'	'last'
'earnings'	'iphone'	'told'
'cent'	'2011'	'system'
'stock'	'take'	'nokia'
'job'	'think'	'consumer'
'also'	'last'	'firm'
'product'	'ipad'	'group'
'ipad'	'ceo'	'2010'
'month'	'internet'	'attack'
'per'	'state'	'rim'
'operating'	'stock'	'state'
'computer'	'product'	'data'
'iphone'	'user'	'maker'
'street'	'technology'	'device'
'price'	'going'	'user'
'according'	'board'	'make'

Table 3: Dimensions obtained using factor analysis on body feature set of 1-year (left) and 3-years (right) of data respectively

Supervised Learning 1 – Multinomial Naïve Bayes

We started the supervised learning with Multinomial Naïve Bayes. The datasets were split into 2/3 training and 1/3 testing sets. Observing that there exist imperfections in the market, we created a cumulative feature set that takes into account information from past news in the following fashion:

$$featureSet(t) = featureSet(t) + \frac{\delta_1}{2} featureSet(t-1) + \frac{\delta_2}{3} featureSet(t-2) + \frac{\delta_3}{4} featureSet(t-3)$$
$$\delta_1 = 1 \; if \; t = 1,2,3 \text{ , } and \; 0 \; otherwise$$
$$\delta_2 = 1 \; if \; t = 2,3 \text{ , } and \; 0 \; otherwise$$
$$\delta_3 = 1 \; if \; t = 3, and \; 0 \; otherwise$$
$$t = 0,1,2,3 \; days$$

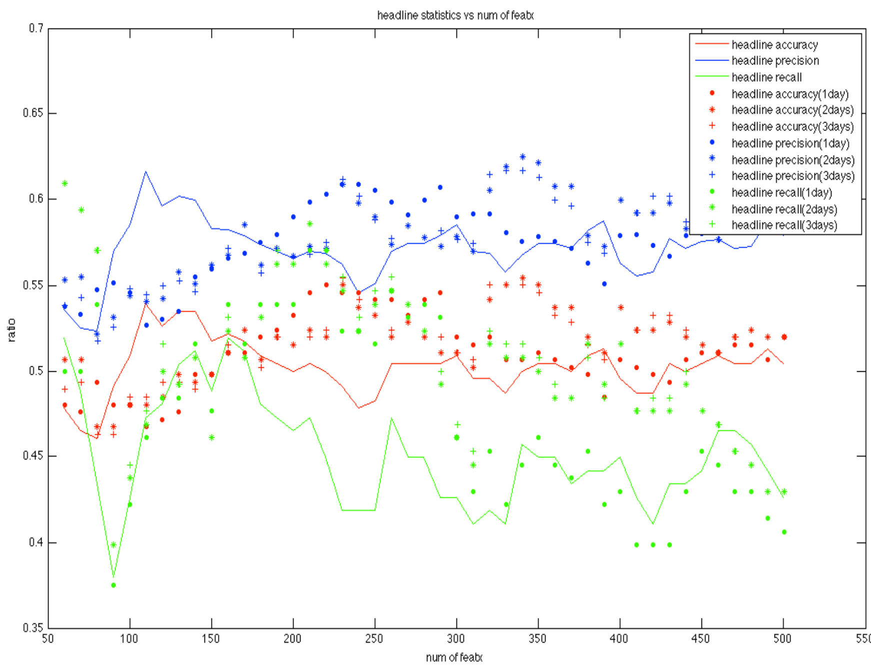


Figure 1: An example of test statistics trained using chi-squared feature selection.

Remarks: 1) Best performance comes from feature sets that contain information only from current day or with one extra day behind. 2) Higher dimension feature set needed for better performance for more number of days incorporated. (See figure, 120 features for 0 days, 240 features for 1 day, 350 features for 2 days)

It is interested that the chi-squared and mutual information feature selection did not perform as well as feature selection using frequent words. From our experiments, we find that the best performance came from feature set generated from 1 year of data with frequent word feature selection. The test results are summarized in the two tables below.

Multinomial Naïve Bayes - 1 year of data						
	Frequent Words		Chi Squared		Mutual Information	
	Headline	Body	Headline	Body	Headline	Body
Best Predictor	160 featx, 1 day	360 featx, 1 day	150 featx, 1 day	170 featx, 1 day	110 featx, 0 days	280 featx, 0 days
Accuracy (Test)	0.66	0.593	0.59	0.61	0.652	0.62
Precision (Test)	0.63	0.59	0.587	0.65	0.8	0.685
Recall (Test)	0.85	0.771	0.75	0.85	0.49	0.88

Table 4: Summary of statistics of MNB using 1 year of data

Multinomial Naïve Bayes - 3 year of data						
	Frequent Words		Chi Squared		Mutual Information	
	Headline	Body	Headline	Body	Headline	Body
Best Predictor	110 featx, 0 day	270 featx, 0 day	220 featx, 1 day	150 featx, 1 day	190 featx, 1 day	500 featx, 1 day
Accuracy (Test)	0.583	0.571	0.55	0.572	0.6047	0.57
Precision (Test)	0.6154	0.578	0.6	0.51	0.6667	0.51
Recall (Test)	0.6822	0.853	0.56	0.63	0.5275	0.63

Table 5: Summary of statistics of MNB using 3 years of data

<u>Supervised Learning 2 – Gaussian Discriminant Analysis</u>

We then tried using GDA to compare performance.

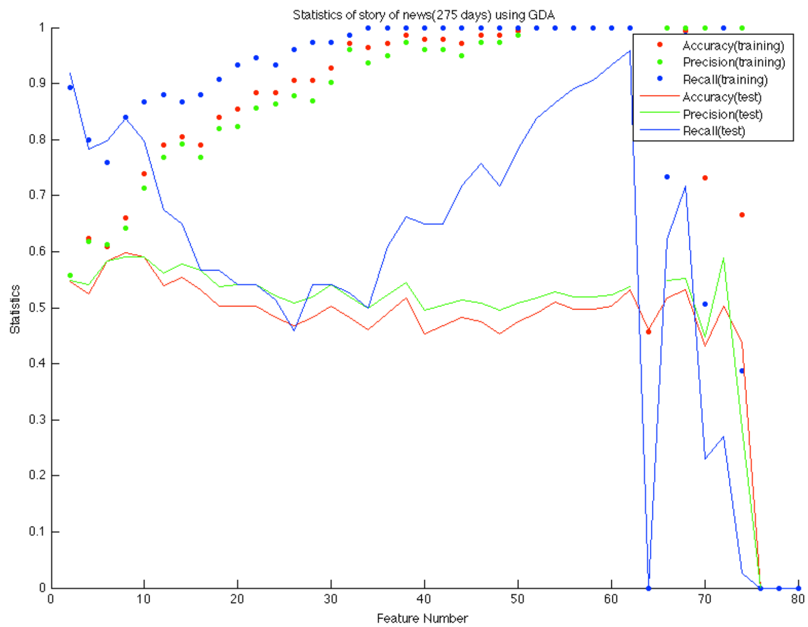


Figure 2: An example of test statistics trained using mutual information feature selection.

Remarks: 1) The covariance matrix rapidly becomes singular at higher feature spaces due to insufficient training data. This occurs when the number of features is approximately equal to the number of training data. 2) Also, we observe that the training data gets fitted very well with increasing number of features, perhaps leading to over-fitting. 3) We are able to obtain good accuracy (60%) on the test data set with low number of features. Thus when it is expensive to collect features, the GDA presents itself as a good alternative to MNB.

The tables below summarize our findings with GDA. We see that in general the best performance comes from number of features that are significantly lower than that required by the MNB.

Gaussian Discriminant Analysis - 1 year of data			
	Frequent Words	Chi Squared	Mutual Information
Best Predictor	24 features (Body)	4 features (Body)	8 features (Body)
Accuracy (Test)	0.5474	0.572	0.6
Precision (Test)	0.625	0.56	0.59
Recall (Test)	0.4	0.85	0.79

Table 6: Summary of statistics of GDA using 1 year of data

Gaussian Discriminant Analysis - 3 years of data			
	Frequent Words	Chi Squared	Mutual Information
Best Predictor	76 features (Body)	6 features (Body)	4 features (Body)
Accuracy (Test)	0.533	0.57	0.52
Precision (Test)	0.571	0.58	0.56
Recall (Test)	0.64	0.79	0.61

Table 7: Summary of statistics of GDA using 3 years of data

Supervised Learning 3 – Support Vector Machines

Unlike the MNB, there does not seem to be a clear trend in the results of the SVM. From the tables below, we observe that we get the best results from the SVM using the mutual information feature selection. All the results below were calculated using linear kernel.

SVM - 1 year dataset						
	Frequent Words		Chi Squared		Mutual Information	
	Headline	Body	Headline	Body	Headline	Body
Best Predictor	140 featx, 0 days	290 featx, 0 days	460 featx, 1 day	90 featx, 0 day	110 featx, 3 days	140 featx, 0 days
Accuracy (Test)	0.554	0.596	0.554	0.602	0.627	0.615
Precision (Test)	0.58	0.598	0.577	0.614	0.714	0.627
Recall (Test)	0.644	0.755	0.667	0.778	0.555	0.7111

Table 8: Summary of statistics of SVM using 1 year of data

SVM - 3 years dataset						
	Frequent Words		Chi Squared		Mutual Information	
	Headline	Body	Headline	Body	Headline	Body
Best Predictor	200 featx, 1 days	150 featx, 0 days	140 featx, 0 days	190 featx, 3 days	80 featx, 0 day	310 featx, 0 days
Accuracy (Test)	0.556	0.558	0.523	0.57	0.665	0.609
Precision (Test)	0.578	0.612	0.573	0.615	0.7097	0.69
Recall (Test)	0.605	0.5659	0.572	0.643	0.682	0.543

Table 9: Summary of statistics of SVM using 3 years of data

Summary and Future Work

We compared the performance of 3 supervised classifiers on the Reuters Technology news section's ability to predict the stock movements of the NASDAQ composite. All 3 were able to perform better than random, with SVM and NMB being able to perform better than 65% accuracy under certain conditions of feature selections, number of features and number of days of information included. Also, with the dimensions learnt from the factor analysis, we can show convincingly that our learning algorithms did pick up hidden trends in the data to aid in prediction. This showed that there is ability of news sentiments to predict stock market movements in the imperfect market conditions we live in today.

The next step would be to build a stronger classifier based on the 3 weak classifiers we presented in this paper. Also, other classification techniques like random forests could be attempted as well. More interestingly, we could incorporate news from other sections, to see which sections provide best prediction capabilities for tomorrow's stock price movements.

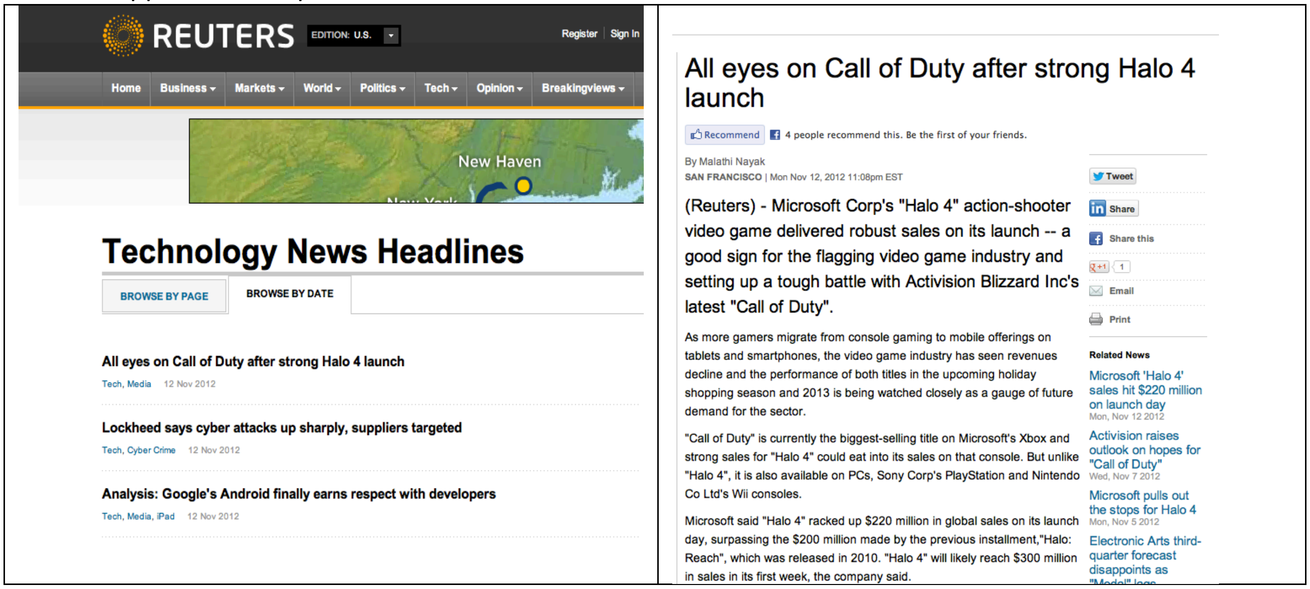
Bibliography

1. Andrew Ng, CS 229 Machine Learning, Stanford University 2012

Appendix

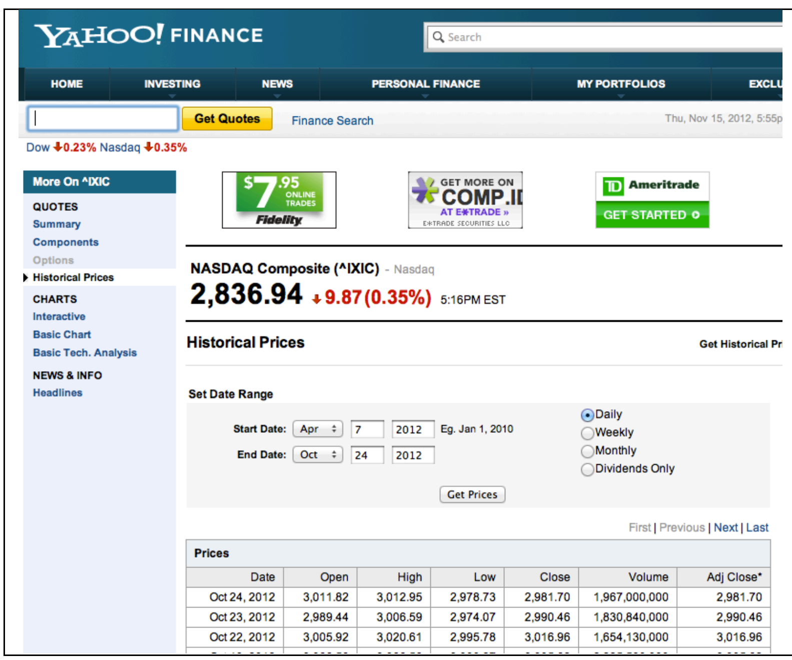
Screenshots of Reuters Technology

Data scrapped from http://www.reuters.com/



Example of Headline and story from Reuters Technology News

Screenshots of Yahoo Finance



Example of screen shot from yahoo finance. This shows the ticker for the NASDAQ Composite. We also experimented with others such as the Dow Jones Industrial Average and the Nikkei Index. However, it appears that the news we were scraping (ie, technology news) were better predictors of the NASDAQ Composite due to the large number of technology companies in this stock market index.