CS224N Final Project: Sentiment analysis of news articles for financial signal prediction

Jinjian (James) Zhai (jameszjj@stanford.edu)
Nicholas (Nick) Cohen (nick.cohen@gmail.com)
Anand Atreya (aatreya@stanford.edu)

***Abstract*—Due to the volatility of the stock market, price fluctuations based on sentiment and news reports are common. Traders draw upon a wide variety of publicly-available information to inform their market decisions. For this project, we focused on the analysis of publicly-available news reports with the use of computers to provide advice to traders for stock trading. We developed Java data processing code and used the Stanford Classifier to quickly analyze financial news articles from The New York Times and predict sentiment in the articles. Two approaches were taken to produce sentiments for training and testing. A manual approach was tried using a human to read the articles and classifying the sentiment, and the automatic approach using market movements was used. These sentiments could be used to predict the daily market trend of the Standard & Poor's 500 index or as inputs to a larger trading system.**

I. Introduction

Traders are using powerful computers to speed-read news reports, editorials, company Web sites, blog posts and even Twitter messages and then letting the machines decide what it all means for the markets. [1] The ultimate goal is to build an automatic trading system which looks for market opportunity and trades on it. Markets often move based on human emotion, and so sentiment may provide a useful signal for trading. We seek to test whether sentiment may be gleaned from news articles describing the market.

The work described in this project can be a module in a larger trading system. This module would extract sentiment from natural language and provide it as an input to that larger system. This project describes two approaches to predicting sentiment from news articles: a manually trained method, and an automatically trained method using market movements. The results of both of these approaches are reported, and the efficacy of using sentiment to predict profitable trading opportunities is discussed.

Due to the time and resources limitation of this project, the problem is open-ended. It gives the foundation of the basic ideas of how to solve the problem using natural language processing (NLP) techniques, but does not describe a complete trading system. This work gives an example of how to parse information and link that information with the market, which in this case is defined as the S&P 500 index. A collection of New York Times articles in 2006 relating to business and finance was used as content for training and testing. These articles were used with the Stanford Classifier utilizing maximum entropy to train and predict sentiment.

A. Previous Work

The initial idea of this project was to exploit online information using net crawler. There have been several papers written on sentiment analysis for social network media. The paper of Alec Go, Lei Huang and Richa Bhayani (2009) [5] discusses issues relevant to the algorithm, classification and query of Twitter messages as positive or negative term. They assumed sentiment classification in Twitter messages can be obtained using machine learning techniques. The authors mentioned the general goal of sentiment analysis for consumer research, product / service or market opinion collection. It was concluded that machine learning techniques perform reasonably well for classifying sentiment in tweets.

The paper of Ravi Parikh and Matin Movassate (2009) [6] uses Unigram and Multinomial Bigram Naive Bayes to detect sentiment. They concluded that sentiment classification on tweets is a significantly different problem than performing similar analysis on movie or product reviews. The former is challenged by the nonstandard and diverse language employed by Twitter users. And the latter

is more concerned with problems of long-format review in determining sentiment.

Because the workload was exceeding our work force, we focused on existing databases and try to use them in financial prediction. Researchers have also worked on detecting stock movement from sentiment in text. The paper of Qicheng Ma (2008) [7] presents a scheme in which 29 stock names are picked from Dow Jones Indistrial Average and related articles in *Wall Street Journal* and *Reuter* Corpus are classified as positive or negative using Naive Bayes and Maximum Entropy classifier.

The paper of Naveed Ahmad and Aram Zinzalian (2010) [8] explores stock volatility forecasting from quarterly earnings call transcripts of the 30 DOW component stocks. They concluded that with large enough training and test sets, historical volatility together with n-gram features, when used together with stop word filtering, improves over the historical baseline. Though summarization was not found to improve model performance and the incorporation of POS adjective tag, handpicked word features do slightly improve over the historical baseline.

The paper of Ryan Timmons and Kari Lee (2007) [9] assumes financial news articles to play an important role in influencing the movement of a stock and there is some lag between when the news article is released and when the market has moved to reflect this information. They built a text classification system to categorize news articles containing references to a list of companies and predict the impact of the news on the stock price.

The paper of Gyozo Gidofalvi [10], [11] predicts the short-term stock price movements using financial news articles. The articles were given one of three labels based on the stocks movement compared to its expected movement. A naive Bayesian text classifier is trained to predict which movement class an article belongs to. In experiments they find predictive power for the stock price movement in the interval starting 20 minutes before and ending 20 minutes after news articles become publicly available.

The paper of Gabriel Fung, Jeffrey Yu and Hongjun Lu (2005) [12] predicts the movements of stock prices based on the contents of the news stories. They investigate the immediate impacts of news stories on the stocks based on the Efficient Markets Hypothesis. Features were the words in a document weighted using term frequency. And the optimization problem was solved using a Support Vector Machine. The resulting system was appropriate for predictions within 3-5 days.

The paper of Robert Schumaker and Hsinchun Chen (2009) [13] proposed a framework called Arizona Financial Text System (AZFinText) to examine discrete stock price prediction using a linguistic and statistical techniques to partition articles by similar industry and sector groupingsand compared the result against quantitative funds and human stock pricing experts. A predicted Directional Accuracy of 71.18% the system gave a trading return of 8.50%, placing the fifth of the top 10 quantitative mutual funds of 2005.

Although we learned a lot from the aforementioned papers, we built a model with our own speciality and differed with them in either dataset size, time frame, market marker, stock index or general methods.

II. Dataset

Our articles are taken from The New York Times Annotated Corpus [4]. This corpus contains every article published in The New York Times from Jan 1987 to Jun 2007. Each article is annotated with date, category, and set of tags describing the content of the article. For the purposes of this project, we have chosen to focus exclusively on the articles that are directly related to the stock market, although future work could examine the value of sentiment analysis on other types of content.

III. Manual Classification

In order to classify natural language sentiment of news articles, two methods were tested for determining sentiment: manual and automatic ones using stock market results. Manual classification involved reading each article and assigning it a sentiment tag: positive, neutral, or negative. A class, NYTManualClassifier, was built to aid in this process. When pointed at a folder containing New York Times articles from the LDC corpus, it presents the headline and first thousand characters of each article for the reader to classify. It also filters articles on the fly so that only articles in the Business section are shown to the user. The XML file for each article is put into a destination folder with subfolders for each sentiment class.

Manual classification is obviously time consuming and we were only able to classify two months

worth of articles. January and June 2006 were chosen. Some temporal diversity was desired because news situations affecting the market can change, causing different overall sentiments, and because we noticed journalists focus on different types of stories at different times of the year. For example, January contains many articles summarizing the results of the previous year and speculating on the upcoming year. These articles may contain sentiment relating the previous year's market return, or they may contain very speculative sentiment about the upcoming year. Sentiment may be very different in June, when half the year has passed and journalists may be more focused on day-to-day movements of the stock market. January and June were manually classified by different team members, so the standards for classification may have been different. This is discussed in the results.

Some general guidelines were used when manually classifying articles. Mergers were generally considered positive because they indicate companies have cash on hand. Technology and general interest articles were considered neutral as they are not directly related to stocks. Lawsuits were generally considered negative, as was corruption. Rising interest rates were considered negative and declining interest rates were considered good because they indicate more cash in the general economy.

The Stanford Classifier v. 2.0, utilizing Maximum Entropy and Quasi-Newton optimization, was used as a tool to classify. First a simple approach was taken with unigrams and bigram features. Words in article headlines were used as one set of features and the words in the body were used as another set. As an additional feature we tried using lists of words considered to have positive and negative sentiment that we found on the Internet [2], [3]. These sentiment words were counted in the articles and four bins were set up for different word counts. These bins were used as additional features.

A. Results

The results from manual classification were promising, as shown in Table I. When January and June 2006 were combined, the classifier achieved F1 scores for positive, neutral and negative classes of 0.581, 0.614, and 0.568, respectively, using unigram and bigram features. 10% of the dataset was reserved for testing, and the results were compared with ground truth to compute these scores.

Adding counts of positive and negative sentiment words from the Internet had inconsistent results: F1 scores changed to 0.556, 0.701, and 0.545. The classifier improved in the neutral category but actually worsened in the positive and negative categories, which is where we had hoped it would improve since it was using positive and negative words. Part of the reason for this could be that these lists were devised for psychological purposes and relate to feelings and relationships, which are not necessarily the same words that might be used to describe business issues. Devising lists of positive and negative sentiment words that have a closer relation to business articles might be productive.

January and June 2006 were classified by two different team members and so it is reasonable to assume classification standards were slightly different. F1 scores for January were 0.588, 0.744, and 0.588 for positive, neutral, and negative sentiment articles using unigram and bigram features. F1 scores for June were 0.567, 0.000, and 0.577. Interestingly, the classifier did not classify any neutral articles correctly in June. This may be a result of the small data set size: 68 articles were used to test, so it is not entirely unreasonable that the classifier will guess them all wrong.

The individual months can be compared to the combined results. The combined results are slightly worse which would indicate that slightly different classification standards were used. The classifier is not as good at classifying a mixture as when it is able to specialize on a single team member's classifications. The differences in classification among the different months can be seen in Table II.

TABLE I
F1-SCORES FOR MANUAL CLASSIFICATION RESULTS

Dataset	*Test ex.*	*Positive*	*Neutral*	*Negative*
Combined	141	0.581	0.614	0.568
Combined incl. +/- words	141	0.556	0.701	0.545
January only	73	0.588	0.744	0.588
June only	68	0.567	0	0.577

TABLE II
NUMBER OF CLASSIFIED NYT ARTICLES IN JAN 2006 AND JUN 2006

Month	Negative	Neutral	Positive
Jan 2006	197	364	171
Jun 2006	263	153	260

IV. Automatic Classification

Movements of the stock market were also used to generate classifications. We realized that our data set would not be large enough to study the impact of article sentiment on individual stocks or even industries. For that reason we decided to use S&P 500 index data to capture movements of the market as a whole. We used the log return: the log of today's close divided by yesterday's close. We designated log returns of greater than 0.0031 as positive, less than -0.0045 as negative, and anything in between as neutral. The value was set to make 36% of the dates positive, 28% neutral and 36% negative as we would like to test with higher volatility. We realized that there is a lag in news publishing and news articles will typically be discussing the previous day's market movements. With this in mind, for each day we had S&P 500 data, we used the market log return to classify the following day's news articles. We also filtered news articles for anything containing *stock*. 200 days worth of data from 2006 were used, starting January 3.

A. Results

The results are considerably worse for articles classified based on stock market movements. F1 scores for positive, negative, and neutral articles using unigrams and bigrams were 0.269, 0.386, and 0.368, respectively. This could be an indication that the stock market itself does not correlate well with what journalists are writing about. It is possible journalists try to find a good mix of positive and negative articles to create interest. For example, on the same day the following pairs of article headlines appeared (shown with the classifications our model chose):

> "47 years after father son wins a nobel too" (positive)
> "profit falls but intel looks at bright side" (negative)
>
> "advanced micro earnings beat estimates" (positive)
> "intel fails to meet its targets" (negative)

The second example is particulary interesting as it involves two direct competitors. On the same day they reported opposite results: this is of interest to readers of the newspaper but obviously very confusing when trying to make predictions.

In addition, it appears the market can be fickle and the market as a whole may not respond the same as individual stocks. Consider the following two headlines on the same day:

> "daimler earnings rise despite setbacks"
> "greenhill plans a secondary share offering"

Both are positive but the market the previous day returned neutral results. The classifier labeled the first article as positive. This is a correct labeling, but does not help our system as the market was neutral. The classifier incorrectly labeled the second article as negative. It is difficult to tell why the classifier labeled this as negative. The article contains language about selling stock as part of the offering, which may have thrown off the classifier. Stock offerings would likely come up frequently and it might improve results to add features delineating stock offerings from articles about stock selloffs.

In addition, feature writing does not have much to do with day to day stock movements. Consider the following headlines:

> "two feuding gallery owners admit the art of tax evasion"
> "hidden cost of shark fin soup its source may vanish"

These are both negative sentiment but occurred after a positive market day. There is not necessarily anything to be gleaned from these articles though: they likely took weeks to prepare, are specialized in topic, and do not appear to discuss or have much bearing on the market as a whole. It might be interesting to study sentiment on larger timescales: we were looking at results of only one day at a time, but it might be possible that newspapers better reflect long term trends. One approach to this might be using moving averages.

V. Article Selection

One of the explanations for the poor performance in the automatic classification task may be due to inappropriate selection of articles from the newspaper. Since this work makes the assumption that there is a correlation between news articles and stock market movements, it is important to select articles that can reasonably be expected to exhibit this correlation. As implied in the previous section, selecting articles simply containing the word "stock" may be insufficient. For example, one of the articles returned by this filter contained:

> Sherlock Holmes is 100 years old and still gets many requests from real people to solve crimes. The **stock** reply is: "Mr. Holmes thanks you for your letter. At the moment he is in retirement in Sussex, keeping bees."

This article likely has no correlation with the stock market's performance on that particular day. Thus, we would like to examine the effect on our results of a better article selection process.

A. Filtering by metadata

To do this, we carefully examined the breadth of articles in the NYT corpus, looking for title, body text, and metadata features that would narrow down the articles to those likely to contain information of direct import to the stock market. The first feature that was identified was the "desk" metadata tag. Analysis showed that articles from The New York Times' "Financial Desk" were highly relevant to the stock market, while those from other departments at the newspaper were of significantly lower importance.

This new filter improved our automatic classification results significantly, as shown in Table refarticleresults. The F1 scores for two classes, positive and neutral, improved by 10-20%. However, the negative score decreased, indicating that the new articles have lower correlation with negative stock market movements.

TABLE III
F1-SCORES FOR AUTOMATIC CLASSIFICATION RESULTS FOR THREE DIFFERENT ARTICLE SELECTION METHODS

Article words	*Test ex.*	*Positive*	*Neutral*	*Negative*
Stock	282	0.269	0.368	0.386
Financial Desk	223	0.454	0.462	0.206
Financial Desk + Finance	78	0.407	0.507	0.258

B. Filtering by content

Given the improvement that the metadata filtering made on our classification results, we wanted to examine further ways to improve article selection and thus potentially further improve our results. Having carefully examined the metadata of the articles, we turned our attention back to the content of the article. While the word "stock" did not effectively filter articles, we found that the word "finance" (as distinct from "financial") did a very effective job of filtering. Empirically, the articles that contain "finance" tend to be focused more on discussions of news that is directly applicable to the stock market, while "financial" articles often mention stock market information in passing.

Combining this with the metadata filtering, we again improved our performance. In this case, while the performance on the positive class went down slightly (though it remained far above the original performance), the performance on neutral and negative increased appreciably. These improved filtering methods allowed us to get much better performance from relatively simple changes.

VI. PREDICTIVE CAPABILITY OF THE MODEL

We now examine the model's ability to predict stock market movements based upon sentiment from a given day. Above, we analyzed the relationship between today's stock return and positive or negative classifications of tomorrow's news articles, under the assumption that news articles posted tomorrow morning are typically about events that occurred today.

Now, we look at whether the model's classifications of articles from the morning of a given day correlate with the stock market return for that day. This type of correlation is what provides the potential value in this model in making predictions about stock market movements during a given day based upon the news reports that have been received so far.

To perform this analysis, we assigned a $+1$ score to each positive article, a -1 score to each negative article, and a 0 score to each neutral article. We then summed up the scores for all articles in a given day to compute an overall sentiment value. This value was then compared against the stock market return for that day to compute a quality score, where positive values indicate a positive correlation between the stock market movement and sentiment value, and negative values indicate a negative correlation. For a model that is effective at predicting the market based on news sentiment, we would expect to see consistently positive scores.

The results of this analysis are shown in Figure 1. As we can see, the results are distributed rather evenly around the origin, thus indicating very poor correlation between the two data series. This result holds true for both of the improved article selection methods.

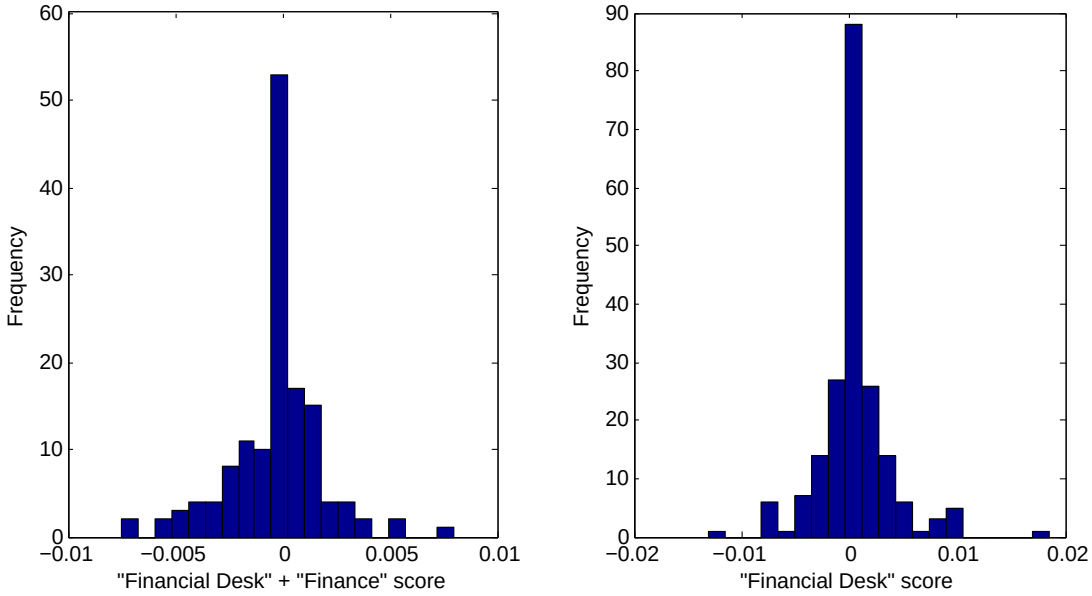

Fig. 1. Histogram of stock prediction quality scores for two different article selection methods. Positive values indicate a positive correlation with stock market movement and article sentiment.

VII. Discussion of Results

A. From Movement to Trading

Although the prediction accuracy is only in 70s percent, we can not say that the result is bad or the idea of extracting information from newspaper is useless. Because we have to keep in mind if we don't have any information from the newspaper, we will at most bet 50% correctly under the assumption of martingale property if considering the stock index as Brownian motion. For example, the paper *A Quantitative Stock Prediction System based on Financial News* [13] got an average Directional Accuracy of 71.18% and still got an average return of 8.5% (the 5th of top 10 hedge fund of year 2005.) in simulated trading. Although we don't have a trading system (because it is beyond the scope of CS224N class) and the method to pick stock and do classification is different from theirs, we can see that our accuracy is still promising compared to previous research. Shumaker's result is obtained using the optimal classification method - "Sector" of the five methods. The other classification methods reported accuracies in lower 50% and simulated trading return at about 2%.

B. Larger Data Set

Intuitively, we know that *New York Times* can not cover every single news in stock market. It can only improve but not determine the prediction. Instead it is an integrated news of social life, technology and sports which are further away from stock prediction. The reason why we chose this corpus is long period(20 consecutive years). So it is good to annotated with lots of historical news. But this data set is used as the purpose of a demonstrative prototype. In the real life, other information such as blogs, tweets, company reports and TV/audio news are all related to stock market trend. A larger data set will definitely promote the prediction accuracy based on the performance of this corpus.

C. Industry Oriented and Time Frame

Another reason to keep the accuracy result from perfect is the nature that one industry is very different and some times negatively related from another. The breaking news of Middle East chaos might be good for oil industry and oil companies, but will be negative to automobile industries. And the lower oil price might be positive for power generation industries, but not necessary for high tech and clean industries. This part of work was called statistical arbitrage based on industries and is also an interesting research topic to apply artificial intelligence in. The time frame of a news can also be very different. The breaking news of Korean War change the economics for years, while the release of dividends typically influences the market for several days. Some news of the same type might also be different in lasting time to influence the market. For example, the merger of Delta and Northwest is more influential than merger in a remote country

such as the merger of Air France and KLM Royal Dutch Airlines.

The tool developed in this project will definitely help the traders to prediction market trend, but in a different level than we expected. The perfect performance of hedge fund is a complexity of hardware, software, algorithm and people.

VIII. Future Work

There are a number of ways this work could be improved. One idea tries to address the fact that most articles are much more nuanced than simply expressing positive or negative sentiment. As a result, it is worth analyzing the potential performance improvements if we perform classification at a more nuanced level - either by separately classifying different portions of an article or by allowing for classifications such as "somewhat positive".

Another promising idea concerns article selection. Since we were focusing on the S&P 500 index, it is likely that we could improve our accuracy by further selecting articles exclusively about companies that are actually in that index. Furthermore, we could weight the importance of such articles based on their relative proportions in the index. This would allow us to better model the relationship between sentiment and market performance.

More broadly, there are many other potential improvements. For example, we could pay attention to the particular time of day that an article was released and correlate that with market movements throughout the day. For an actual trading system, this would be important since decisions have to be made at a fine timescale. In addition, we could examine Twitter posts both to gain an idea of overall investor sentiment as well as learn relevant information about specific companies in a timely manner.

IX. Conclusions

This work has demonstrated the difficulty of extracting financially-relevant sentiment information from news sources and using it as a market predictor. While news articles remain a useful sort of information for determining overall market sentiment, they are often difficult to analyze and, since they are often focused on conveying nuanced information, may contain mixed messages. Furthermore, the success of this model relies largely upon the exploitation of market inefficiencies, which often take a great deal of work to identify if they are to be reliable.

Thus, while our system provides interesting analysis of market sentiment in hindsight, it is less effective when used for predictive purposes. Nonetheless, given the coarse signals produced by our model, it is important to note that it is not necessary to trade directly using the values produced from our model. The sentiment results we produce could instead be an input to another trading system or simply be given to human traders to aid their judgments.

X. Collaboration Statement

This work was done by Nick Cohen, Jinjian James Zhai and Anand Atreya.

Nick performed manual classification for January 2006, wrote the classes NYTManualClassifier (the manual parsing assist program), PreprocessNYT (cleans and formats New York Times article data), SP500Data (stores and retrieves S&P 500 data), and AutoClassifyNYT (integrates New York Times data with S&P 500 results for automatic classification), collected and analyzed results, and wrote sections III and IV of this report.

James proposed idea, reviewed peer work, collected/processed S&P 500 data (SPMovement.log, run0 and ReadWriteTextFileWithEncoding.java), manually classified 600+ articles (run1, run3, Jun2006.tar.gz, manual_parse.log), wrote initial article selection script (run4, FastParsingAllstock.zip) and wrote sections I, II and VII in report.

Anand optimized article selection to improve classification performance, integrated classification results with S&P 500 data to examine predictive capability (AnalyzeClassifications class), collected and analyzed results, and wrote sections V, VI, VIII, and IX of the report.

All team members contributed to the structure of the report.

References

[1] Graham Bowley, "Wall Street Computers Read the News, and Trade on It", *New York Times*, Dec 21, 2010.

[2] An Energizing List of Positive Words, http://www.the-benefits-of-positive-thinking.com/list-of-positive-words.html

[3] Negative Feeling words, http://eqi.org/fw_neg.htm

[4] http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2008T19

[5] Alec Go, Lei Huang and Richa Bhayani, "Twitter Sentiment Analysis", *CS224N Final Report*, 2009.

[6] Ravi Parikh and Matin Movassate, "Sentiment Analysis of User-Generated Twitter Updates using Various Classification Techniques", *CS224N Final Report*, 2009.
[7] Qicheng Ma, "Stock Price Prediction Using News Articles", *CS224N Final Report*, 2008.
[8] Naveed Ahmad and Aram Zinzalian, "Predicting Stock Volatility from Quarterly Earnings Calls and Transcript Summaries using Text Regression", *CS224N Final Report*, 2010.
[9] Ryan Timmons and Kari Lee, "Predicting the stock market with news articles", *CS224n Final Report*, 2007.
[10] Gidofalvi, Gyozo. "Using News Articles to Predict Stock Price Movements". Department of Computer Science and Engineering, University of California, San Diego, 2001.
[11] Gyozo Gidofalvi and Charles Elkan, "Using News Articles to Predict Stock Price Movements", *Technical Report*, 2003
[12] Gabriel Fung, Jeffrey Yu and Hongjun Lu, "The Predicting Power of Textual Information on Financial Markets", *IEEE Intelligent Informatics Bulletin*, Vol. 5. No. 1, June 2005.
[13] Robert Schumaker and Hsinchun Chen, "A Quantitative Stock Prediction System based on Financial News", *Information Processing and Management*, Vol. 45. Iss. 5, pp. 571-583, Sept 2009.