Learning Business Article Sentiment Based on Stock Market Performance

Jeffrey Schlosser and David Li

Introduction

The authors' ultimate goal is to predict stock market trends based on a corpus of business news articles published in real time. It is believed that stock brokers, as human agents, will affect market prices by emotionally responding to financial news. As a result, articles that are published on a certain day will have a strong correlation with financial events in the near future. However, given that the scope of this problem is too large to face in a single quarter, the focus of this project is on analyzing the sentiment of business articles published in the past.

Identifying article sentiment can generally be a difficult and time consuming procedure. Fortunately in finance, the sentiment of an article can be reliably linked to the trend in relevant stock market prices in the time period around publication. In the proposed algorithm, an article is classified as positive when it occurs during a time period associated with a favorable market response. Since market response is quantitative and can be easily determined, our procedure allows for labels to be applied automatically to a large corpus of training data. Machine learning is applied to learn a model of the training data, and to make predictions about the sentiment of previously unread business articles.

Methodology

Our proposed business article classification system has three major aspects: data collection, model training, and testing and evaluation.

<u>Data collection</u>: Several sources of training data were considered as the project matured, and their progression is shown in the figure below.

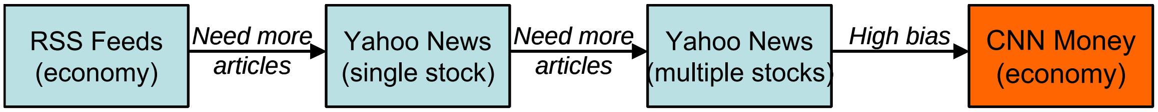


Figure 1: Progression of data sources.

The data sources are shown in blocks, and the needs addressed by moving from one source to the next are shown beside the arrows. Shown in parenthesis below the sources are the search criteria used to gather articles. RSS feeds and CNN money were mined for general articles about the economy, while Yahoo news was searched for specific company(s). Overall, CNN Money was found to have a large volume of relevant financial articles concerning the economy, so it was chosen as the preferred and only data source.

Data harvesting is accomplished by searching cnnmoney.com for business articles, then filtering by date. A web interface is used to extract the title, summary, and/or full text of each article, and the Porter stemming algorithm is used along with a specific vocabulary list (manually or automatically specified) to generate a histogram of tokens. Manual selection was performed by picking the tokens which subjectively seemed most relevant to the authors. Automatic vocabulary selection was performed by eliminating all tokens with frequency less than a certain threshold, then training on this data set. The most relevant positive words were those that had the largest ratio between $\square_{j|y=1}$ and $\square_{j|y=0}$ after Naïve Bayes training, and the reverse was true for the most relevant negative words.

Figure 2 summarizes the implemented data collection process, and its relationship to the learning algorithm. .

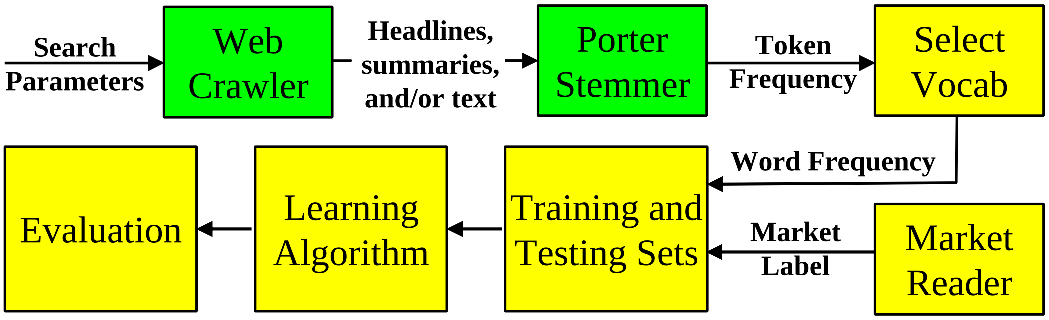


Figure 2: Software Flow Chart.

Training: Both Naïve Bayes and Support Vector Machine (SVM) algorithms were used to learn a predictive model from the collected data. Each article was labeled according to its date of publication using the following metric based on the DOW Jones Industrial Average (DJI) stock quote: *Opening value the following day minus closing value the previous day*. To ensure that a close number of positive and negative training examples were used, thresholds for the DJI metric increase and decrease were implemented. Articles that did not show an increase or decrease greater than the thresholds were discarded, and the thresholds were adjusted until an even number of positive and negative training examples were acquired. Note that no causal relationship was assumed between the article publication and the stock prices; some articles could have *caused* the stock to rise or fall, while other articles could have *described* a stock rise or fall in a given day. Since the project goal was to classify the sentiment of financial articles and not to predict future stock performance, this ambiguity in the causal relationship is acceptable.

Testing and Evaluation: Naïve Bayes testing yields a measure between zero and one that a single article is classified as positive. One way to act upon these results would be to place a threshold at 0.5, classifying the article as positive if the output is greater than 0.5 and negative otherwise. Another choice would be to label an article as positive, negative, or uncertain—the case where the measure is close to 0.5. Using the latter metric, articles on which the algorithm is uncertain are removed from the testing dataset. The remaining articles are labeled as confident positive or confident negative. Keeping in mind that an investor is interested in making decisions using all articles published in a given time period, the arithmetic mean of all confident articles in a given day yields information about a single day. A similar thresholding and prediction routine can then be applied to data for each day, resulting in confident positive and confident negative days. These predictions can then be validated from stock market data.

Summary of Terminology:

•*Fraction of Correct Articles* **:** The number of testing articles classified correctly divided by the total number of testing articles

•*Normalized Confident Days (Confidence Measure)***:** Number of days which are given a high confidence value when testing with Naïve Bayes divided by the total number of days considered

•*Normalized Correct Confident Days (Accuracy Measure)*: Fraction of the days deemed confident that have correct classifications

The fraction of correctly classified training examples (fraction of correct articles) is a standard evaluation metric in machine learning. The two other metrics relate to the ultimate goal of predicting future stock trends using business article sentiment. A human agent will only invest in a stock if he or she is confident that a prediction will be correct, and will only profit on a correct prediction. Therefore a good investor is one that makes correct predictions on a high percentage of days.

Results

Naïve Bayes: The classification system proposed above was trained using Naïve Bayes on 17848 articles, from 1/1/2003-1/1/2006, and tested on 9845 articles from 1/1/2006-1/1/2007. Figure 3 represents the classification accuracy and confidence of a model trained on this data.

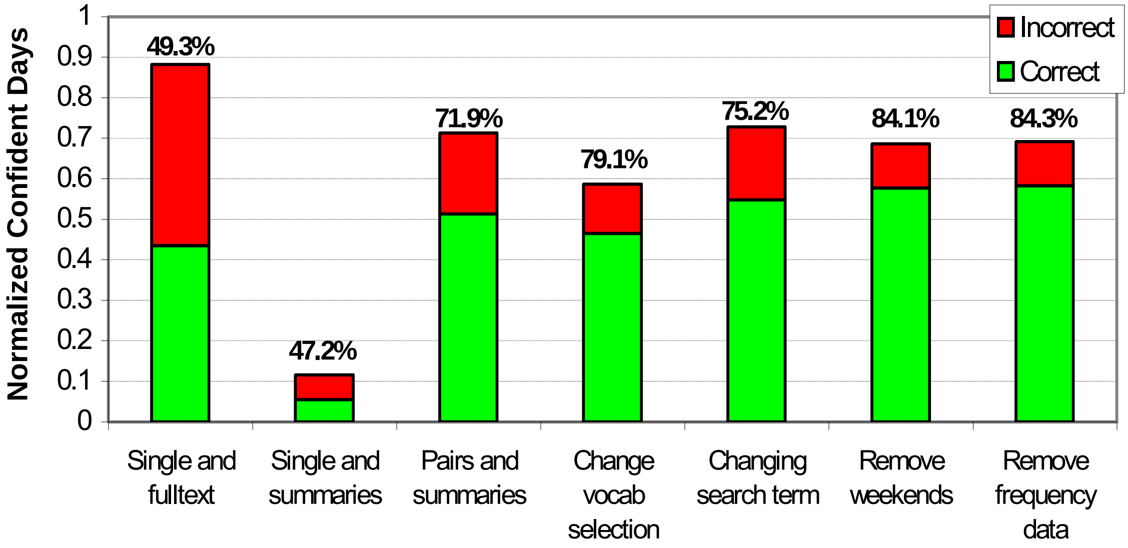
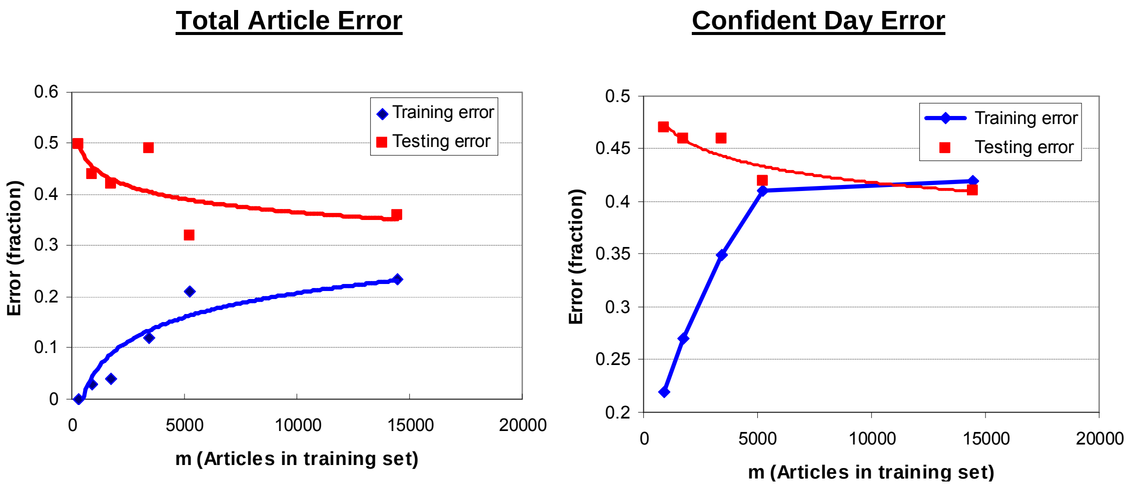


Figure 3: Algorithm Performance.

The vertical axis of Figure 3 denotes the normalized confident day metric, which measures the confidence of the model. The accuracy of the model—the normalized correct confident day metric—is presented here as the ratio between the colors of the graphs. The percentage of correct predictions made on confident days is presented above each bar. The horizontal axis represents a progression in features and test parameters. The test parameters are detailed in Table 1 below. While exact weights on confidence and accuracy are decided by the user, the data from these experiments indicates that certain test parameters yield clear performance increases.

Table 1. Test parameters from Figure 3.

Label	Part of Article	Token Type	Vocab Selection	Search Terms	Removed Days	Frequency Data Used?
Single and fulltext	Full text	Single	Frequency based	Dow, Economy, Stock	None	No
Single and summaries	Summaries	Single	Frequency based	Dow Economy Stock	None	No
Pairs and summaries	Summaries	Pairs	Frequency based	Dow Economy Stock	None	No
Change vocab selection	Summaries	Pairs	Human based	Dow Economy Stock	None	No
Changing search term	Summaries	Pairs	Human based	Dow Stock	None	No
Remove weekends	Summaries	Pairs	Human based	Dow Stock	SMFS	Yes
Remove frequency data	Summaries	Pairs	Human based	Dow Stock	SMFS	No


Total Article Error

Confident Day Error

Figure 4: Learning curves.

The learning curves presented in Figure 4 on the predictions for articles (left) and confident days (right) show that sufficient training data was collected. While the error may seem high in the absolute sense, using thresholds yields good classification performance with the evaluation metrics presented in the last section. The rate of decrease of the testing error has slowed for both articles (left) and days (right) to the point where a larger training set would not yield substantial returns. In addition, the shape of the curves suggests that bias could be further decreased. Future test would benefit from a larger set of features, such as those resulting from groups of three or four words.

Support Vector Machine: A series of tests employing the Support Vector Machine method were applied to the same data used in Naïve Bayes in order to compare results from discriminative and generative learning algorithms. Different kernels, including radial basis, sigmoid, and linear, yielded results that were within one percentage point of the Naïve Bayes results.

Future Prediction: The previously described training metric, involving stock prices before and after article publication, cannot be used to predict the future because it would require buying stock before the article is published. A more practical technique would only consider price trends on the day after the article was released. Training on this modified metric yielded a model that tended to make a high percentage of confident labels with low accuracy. For example, a typical result using word pairs on all days of the week yielded a confidence of 71.3% in the past and 66.5% in the future, with an accuracy of 71.9% in the past and 52.9% in the future.

Discussion

Test Parameter Selection: Examination of the problem parameters can provide insight into this machine learning classification problem (Figure 3). Testing shows that single token analysis on articles with full text yields a high confidence, but low accuracy. A move to article summaries drastically reduces the confidence. One explanation is that full text articles provide many words that are not relevant to sentiment analysis and so accuracy drops. Moving to examining token pairs (instead of single tokens) and article summaries leads to a large performance gain. A word pair offers inherently more information because it captures both words, as well as their relationship. Table 2 shows the 5 most relevant positive and negative word pairs and single words, as determined by Naïve Bayes training. Clearly, the word pairs give more relevant information than single words about economic article sentiment.

Table 2. Most relevant words and word pairs

Negative Pairs	Positive Pairs	Negative Single	Positive Single
stock slump	stock surge	even	tough
take profit	stock gain	future	or
stock fell	stock rally	today	world
stock slip	edge higher	deal	how
weak in	investor cheer	into	higher

Analysis of the vocabulary selection shows that the single tokens are ambiguous because they are taken out of context, while word pairs retain more context and more meaning. Further reducing the vocabulary with human input decreases confidence, but increases accuracy. The intuition is that there is a smaller set of words that yield meaning, but these words are more relevant. Taking a subset of search terms (from “Economy Stock Dow” to “Stock Dow”) yields gains in accuracy because “Economy” has weaker ties to the Dow Index. Restricting consideration to articles published on Tuesday, Wednesday, and Thursday has a modest increase in accuracy. Since the DJI only contains information on weekdays (trading is stopped on weekends), the training metric for Friday, Saturday, Sunday, and Monday contains information about days that are not adjacent. For example, news published on Friday is labeled according to the opening stock price *Monday* minus the closing price Thursday. News articles lose relevancy after a certain time period, so accounting for only Tuesday, Wednesday, and Thursday increases accuracy. Finally, analysis based on word presence in the article showed virtually no change from analysis with word frequency. Using article summaries, enough relevant information is encoded in a short amount of text that repeated words are less common, and therefore less significant to consider.

Future Training Metric: The future training metric mentioned in the results section was not expected to perform well in testing because of the ambiguity in the causal relationship between articles and stock prices in the training set. If a training set was obtained that contained precise times of article publication, the training articles could be labeled using stock market prices a short time after article publication. This would ensure that stock prices were a *reaction* to article publication, and testing results would likely improve.

Acknowledgement

We would like to thank Andrew Ng, Paul Baumstarck, and Ian Goodfellow for their help and insights in our project.