Chapter 22

Good News or Bad News? Let the Market Decide

Moshe Koppel and Itai Shtrimberg
Dept. of Computer Science
Bar-Ilan University
Ramat-Gan, Israel
koppel@netvision.net.il, ishtrimberg@iai.co.il

Abstract

A simple and novel method for generating labeled examples for sentiment analysis is introduced: news stories about publicly traded companies are labeled positive or negative according to price changes of the company stock. It is shown that there are many lexical markers for bad news but none for good news. Overall, learned models based on lexical features can distinguish good news from bad news with accuracy of about 70%. Unfortunately, this result does not yield profits since it works only when stories are labeled according to cotemporaneous price changes but does not work when they are labeled according to subsequent price changes.

Keywords: sentiment analysis, financial analysis, automated labelling.

1. Introduction

The assessment of sentiment in written text is inevitably subjective and subject to considerable disagreement (Wiebe et al. 2001a). In some instances, such as starred movie (Turney 2002, Pang et al. 2002), restaurant (Finn and Kushmerick 2003) or product (Kushal et al. 2003) reviews, the author provides self-assessment. But in most cases, we require human judges to provide an assessment of a document's sentiment. As a result, one of the main research bottlenecks in sentiment analysis has been the procurement of large reliably labeled corpora.

The case of business news concerning a publicly-traded company is a special one, though, because price movements of the company's stock can serve as an objective measure of the valence of a news item (although, not every movement in price is a direct consequence of a given news story). The market effectively serves as judge. Thus the use of price movements correlated with the appearance of news items is a promising method for automatically generating a labeled corpus without directly invoking individual human judgments (though, of course, stock movements themselves are a product of collective human judgment).

In this paper, we classify news stories about a company according to its apparent impact on the performance of the company's stock. We will check the extent to which a learned model can be used to classify out-of-sample texts in accordance with market reaction. That is, we will determine how much of the information in a news story that drives market reaction can be gleaned from textual features alone. We will also test the degree to which such learned models can be used to turn a profit. (To prevent letdown, let us note already that our conclusions in this regard will be somewhat pessimistic.)

It is important to distinguish our approach from others (Das and Chen 2001, Seo et al. 2002) which directly judge whether a story is good news or bad news in the hopes of turning a profit based on the assumption that what human judges deem to be good (bad) news leads to exploitable price increases (decreases). In this work, we avoid such assumptions by making no judgment regarding a story itself. We are interested only in the market's reaction to the story. A similar approach was previously considered by Lavrenko et al. (2000).

2. Experiments

As our initial data set, we gathered news stories concerning each of the stocks in the Standard & Poor index of 500 leading stocks (S&P500) for the years 2000-2002. The stories were taken from the Multex Significant Developments corpus (which had been found at http://news.moneycentral.msn.com/ticker/sigdev.asp but has since been removed). The advantages of this particular corpus include that it covers only significant stories and eliminates redundancy. The total number of stories in our database is just over 12,000 – an average of 24 stories per stock. The average length of a story is just over 100 words – short enough to be focused but long enough to permit harvesting of statistics.

We used two approaches to labeling a story as having positive/negative impact on stock price. In the first approach, we matched each story with the change in price of the relevant stock from the market close the day preceding the publication of the story to the market open the day following the story. Thus for example if the news appeared on January 15, we compared the price of the stock at the close of January 14 and the open of January 16. This period is long enough to reflect market reaction to the news regardless of the particular hour of January 15 when the news became public but short enough to minimize the chances that other significant market or company news might mask the impact of this story.

A different approach, which we did not try, would be to assume that the time-stamp on the story accurately reflects the precise hour when the news became public and to check price changes during a narrower time band around that hour. Unfortunately, such an assumption would be unduly optimistic.

In the second approach, we matched each story with the change in price of the relevant stock from the market open the day *following* the publication of the story and the market open the day following that one. Thus for example if the news appeared on January 15, we compared the price of the stock at the open of January 16 and the open of January 17. Although, the first approach provides a more reliable assessment of the story's impact, the second offers a more exploitable one since, if the story is published after market close, the next day's open is the first price at which an investor might be able to purchase the stock.

In both cases, we defined a story as positive if the stock in question rose 10% or more and as negative if the stock declined 7.8% or more. We used these rather high thresholds because such dramatic price moves can be safely assumed to be reactions to news stories and not mere reflections of general market moves or random fluctuation. The lower threshold for downward moves was chosen so as to provide an equal number of negative examples as positive examples. Using our first approach, these thresholds resulted in 425 positive examples and 426 negative examples.

We also considered a subset of these stories that satisfied two additional conditions:

1. The base price of the stock was at least $10.
2. The percentage change in the stock price is in excess to the percentage change in the S&P.

Such stories can be more reliably linked to the price change of the stock than those that do not satisfy these conditions. Approximately half the stories satisfied both conditions.

We used as our feature set all words that appeared at least 60 times in the corpus, eliminating function words (with the exception of some obviously relevant words such as *above*, *below*, *up* and *down*). Since the texts are quite short, we represented each text as a binary vector reflecting whether or not a feature is present in the story, but ignoring frequency. Previous work (Pang et al. 2002) has indicated that presence information is superior to frequency information for sentiment categorization.

Our categorization methodology consisted of selecting the 100 features with highest information gain in the training corpus and then using a linear SVM (and other learners) to learn a model.

3. Results

Using our first labeling approach, 10-fold cross-validation experiments yielded accuracy of 70.3% using a linear SVM. Training on the entire 2000-2002 corpus, while testing on the 2003 corpus, yielded accuracy of 65.9%. Other learners, including Naïve Bayes and decision trees, yielded essentially the same results. Boosting and selection of other kernels for the SVM also had little effect on results. In addition, use of the narrower set of more reliable examples yielded essentially the same results despite the fact that the number of training examples was considerably smaller.

Closer analysis of the results offers some interesting insights. There are a number of features that are clear markers of negative documents. These include words such as *shortfall*, *negative* and *investigation*. Documents in which any of these words appear are almost always negative. However, unlike Tolstoy’s happy families, every happy stock story is happy in its own way. There are no markers of positive stories; positive stories are characterized only by the absence of negative markers. In fact, of the twenty words in the corpus with highest information gain, all are negative markers. As a result, recall for positive stories is high (83.3%) but precision is much lower (66.0%); the misclassified documents are mostly those negative stories which fail to have any of the standard markers. This trend is evident regardless of which learner is used.

Since 77.5% of stories classified as negative really are negative, one might hope to develop a strategy which could leverage this information to make short investments based on such stories.

But recall that the labeling approach we have been using measures price moves from *before* the appearance of the story. To invest at that base price on the basis of a not-yet-published story would involve look-ahead unavailable to an investor not in possession of inside information. The honest investor interested in exploiting published news to select an investment vehicle, must use our second labeling approach which reflects price moves *subsequent* to the publication of the story. Unfortunately, 10-fold cross-validation experiments using this labeling approach yield much more modest results of just above 52% – probably too small a margin to overcome the cost of trading. This bears out the so-called Efficient Markets Hypothesis (Fama 1970): "prices fully reflect all available information".

4. Conclusions

The main contribution of this short paper is to suggest a new method for automatically collecting labeled data for sentiment analysis. The use of stock price movements offers several large advantages over hand-labeled corpora. First, the entire procedure is automatic and thus a large corpus can easily be generated. Second, the collective judgment of the market is a more reliable determiner of sentiment than that of a small number of judges. Finally, if the objective of the analysis is to maximize profit, the method of labeling directly matches the objective.

Having assembled a modest corpus in this way, w have found that we can learn to automatically characterize news stories about stocks according to their market impact with moderate success. At the very least, we can reliably identify certain stories as negative. More sophisticated features, such as word collocation (Wiebe et al. 2001b), might improve matters, but other learning methods are unlikely to improve results significantly.

However, our experiments leave little room for optimism that analysis of news stories might be successfully parlayed into an investment strategy. It is very likely, though, that labeling news stories according to their price impact in a period of several seconds or minutes following first publication of the story would yield more promising results. This should be the main direction of future research in this area.

5. References

Das, S. and Chen, M. (2001) Yahoo for Amazon: Extracting Market Sentiment from Stock Message Boards. In *Proceedings of the 8th Asia Pacific Finance Association Annual Conference (APFA 2001)*, Bangkok, Thailand.

Fama, E. (1970) Efficient Capital Markets: A Review of Theory and Empirical Work. *Journal of Finance 25*, 383-417.

Finn, A. and Kushmerick, N. (2003) Learning to classify documents according to genre. *In IJCAI-03 Workshop on Computational Approaches to Style Analysis and Synthesis*, Acapulco, Mexico.

Kushal D., Lawrence, S., and Pennock, D. M. (2003) Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In *Proceedings of the Twelfth International World Wide Web Conference (WWW-2003)*, 519-528, Budapest, Hungary.

Lavrenko, V., Schmill, M., Lawrie, D., Ogilvie, P., Jensen, D., and Allan, J. (2000) Mining of Concurrent Text and Time Series. In *Proceedings of Text Mining Workshop of the Sixth ACM*

SIGKDD International Conference on Knowledge Discovery and Data Mining, 37-44, Boston, MA.

Pang, B., Lee, L. and Vaithyanathan, S. (2002) Thumbs up? Sentiment Classification using Machine Learning Techniques. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 79-86, Philadelphia, PA.

Seo, Y., Giampapa, J.A., and Sycara, K. (2002) *Text Classification for Intelligent Portfolio Management*. Technical report CMU-RI-TR-02-14, Robotics Institute, Carnegie Mellon University.

Turney, P. D. (2002) Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews. In *Proceedings of ACL 2002*, 417-424, Philadelphia, PA.

Wiebe, J., Bruce, R., Bell, M., Martin, M., and Wilson, T. (2001) A Corpus Study of Evaluative and Speculative Language. In *Proceedings of 2nd ACL SIGdial Workshop on Discourse and Dialogue*. Aalborg, Denmark.

Wiebe, J., Wilson, T., and Bell, M. (2001) Identifying Collocations for Recognizing Opinions. In *Proceedings of ACL 01 Workshop on Collocation*. Toulouse, France.