Stock Price Prediction from Natural Language Understanding of News Headlines

Vasilios Daskalopoulos   VASIL@PAUL.RUTGERS.EDU
Rutgers University, Department of Computer Science

Abstract

This paper describes the design of a natural language understanding agent able to predict the movement of stock prices due to the occurrence of news headlines pertaining to a stock symbol. Using prospective training data, a Naïve Bayes model is built, that is used to score the impact that new instances of headlines will have on the value of a stock price. Linguistic features of the headlines are used for model learning and can provide human understandable interpretations of learned rules. The agent is designed to be domain independent so that it may be included as a sub-module to any machine-learning system that needs a reinforcement signal from natural language data.

1. Introduction

This paper proposes a design for a machine-learning agent that is trained to predict short term fluctuations in the value of a stock price that are reactions to recent news headlines having to do with a particular company. The agent will utilize language models to determine if an utterance in the form of a news headline about a company will affect its stock in a positive or negative way.

A Naïve Bayes classifier built from the training data is used to determine if News Headlines should be classified as those indicative of a forthcoming rise in a stock price or a fall.

1.1 Motivation

This agent could potentially be used as a sub-component of a stock-trading agent that could use human language data such as analyst's reviews and news headlines as an input feature to a more complex learned asset-trading policy. More generally, an agent successfully designed to learn whether natural language utterances provides a positive or negative indication of reinforcement can be used as an off-the-shelf component to a general natural language dialogue agent capable of learning new things about the world through a reinforcement channel provided entirely in natural language. Consider a chat-bot capable of modifying the way it communicates with another interlocutor through reinforcement learning, and that is able to ascertain whether the other party can understand it. The natural language understanding provides the most natural and direct feedback channel to the agent. Further, crossing this bridge enables reinforcement learning agents to learn from human-language instructions.

1.2 Source of Data

The finance section of Yahoo.com (http://finance.yahoo.com) provides an interface to a large quantity of historical data about thousands of stock symbols with both dated price information and dated news headlines having to do with the ticker symbol. This provides a potentially rich corpus of data with which to train an agent to correlate its interpretation of a headline with the actual change in a stock price. The website's historical data can be mined and used to train the agent. It may also then be checked daily to make predictions once it is trained.

In this experiment, 8885 headlines between November 15th, 2002 and November 15th, 2003 across 100 different stock symbols were collected along with 25201 instances of daily price quotes through the same time frame. The raw HTML data was parsed and filtered to obtain raw tuples of <date, headline> and <date, quote> instances. These were then aligned by date and combined to provide the instances the agent would learn on.

1.3 Training

The agent is autonomously trained on 90% of the financial data corpus, reserving 10% for testing. The data chosen for training and testing is chosen randomly at run time. This assumes that any temporal information that may exist in the data that may bias the classification is abstracted away. That is, the training instances are taken to be independent of each other and are only dated so that quotes may be matched with headlines.

A particular quote instance is classified as either a RISE or a FALL if it has risen or fallen by more than 10% of the price of the previous day. Otherwise the instance was regarded at a NOTHING. A similar trigger was used by Macskassy (2003) to signal a prospective data instance deemed to be interesting. The prior probabilities of instances that indicated a RISE, FALL or NOTHING

across all the data collected were 0.013055, 0.007976 and 0.978969 respectively.

2. Results

The agent was tested on an increasing amount of training data (10% of which was reserved for testing and NOT used for training) in the increments shown in Table 1. These results are plotted in Fig 1. For each round, 10 sub rounds of random 90/10 train/test data was selected and trained on. The average prediction accuracy across the 10 sub rounds is shown in Table 1 for each round.

Table 1. Classification accuracies for each 10 rounds of increasing training data.

ROUND	# OF HEADLINES	AVERAGE ACCURACY (10 SUB-TRIALS)
1	878	**0.988183**
2	2049	0.959354
3	2309	0.945009
4	3711	0.952243
5	4329	0.952420
6	5222	0.941922
7	5992	0.950521
8	7032	0.948095
9	7795	0.950015
10	8885	0.954537

Note that the scarcity of the instances that were considered RISE or FALL relative to the NOTHING instances had a significant effect on the observed results. While, at first glance, the range of the accuracy may seem to be impressive (between 0.988 and 0.945), it actually reflects poorer performance than that of a trivial agent design which always chooses to classify examples as NOTHING since this would yield 97.8% accuracy. One way improve this situation would be to reduce the scarcity of the RISE and FALL instances by loosening the definition of those classifications. (For example, a RISE could be considered any quote that is greater than the previous day's quote by more than 5% rather than 10%) This, however, must be done in such as way as to not introduce too much noise in the meaning of a "RISE" or "FALL". Further experimentation might illustrate how much of this tradeoff to make.

Furthermore, the results unfortunately do not show a significant increase in classification accuracy as the amount of training increases. In particular, the first few increases in training examples yield a relatively strong reduction in the agent's accuracy. One way to explain these results is to note that the way the data was increased was by adding more companies' data to the training. This

had the unintended effect of diluting the strength of the learned language model since the dictionary collected was used across all companies. To illustrate this further, consider the first, most successful, point on the graph. Its 878 training instances all come from one company. This language model is especially strong since it is specific to the language used for that particular company. Further training examples come from other companies, and so learned words may have different interpretations for different contexts. A way to test this theory and improve the performance of the agent would be to choose incremental amounts of training data randomly from different companies.
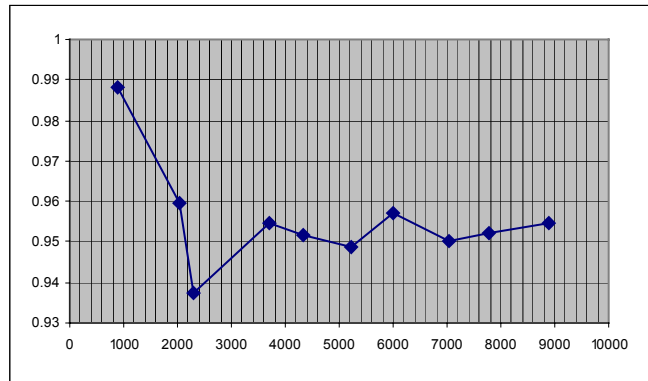

Fig 1. Classification accuracy vs. training data.

2.1 Dictionary

A word dictionary was generated consisting of each word from every headline read in the experiment. There were 65168 occurrences of 8377 distinct words that appeared. These were scored according to how often they appeared in a headline aligned with a quote that was a RISE, FALL or a NOTHING. Some interesting words that were highly indicative of either positive or negative trends are shown in Table. 2. Notice neutral words such as "despite" often wound up having an influence in an either positive or negative direction even though they do not conventionally carry the learned connotation in everyday English language use. This might be indicative of the fact that not enough data was used to smooth out noise in the language model built.

There were also very high counts for common English prepositions (in, from, to, of) that intuitively should not be considered in the language model unless higher-level syntactical parsing is done on the headlines. However, for the purpose of this experiment and simplicity of implementation, no filtering was done. It also remains to be seen whether this type of filtering would bias the data to the point where is it detrimental to the accuracy of the classification. A control for this experiment to determine the benefit of preposition filtering should be conducted

and might perhaps improve the language model used across all company headlines.

Table 2. Example words highly indicative RISEs or FALLs.

TOTAL OCCURANCES	DURING RISE	DURING FALL	WORD
98	11	0	upgraded
59	0	5	downgraded
97	21	1	bank
58	6	1	big
15	3	0	boosts
133	11	3	deal
9	0	2	disappoint
38	2	8	drop
31	3	1	despite
189	8	4	disclosure
96	3	1	Growth
53	7	1	gains

Note that most words in the dictionary appeared only once and that most had a 0 for both the RISE and FALL column (P(w|h)= 0). This data scarcity can be problematic for Bayesian models but a technique similar to that used by Lavrenko et al, (2000) for approximating P(D|h) was employed.

3. Conclusion

While the agent performed fairly well with a very basic Naïve Bayes design and a trivial language model (word occurrences), it seemed to not improve significantly with increased training data. This may reflect un upper limit to how robust a classifier can be yielded with a simple "Bag-of-words" approach. Further work should attempt to utilize more grammatical structure in the headlines to indicate a positive or negative statement. Also, simple word-based decision-stubs as inputs to the Boosting algorithm might provide a more robust classifier with the sparse data that words in headlines provide

Acknowledgements

References

Fawcett, Thomas & Provost, Foster (1999). *Activity monitoring: Noticing interesting changes in behavior.* In Proceedings of the Fifth International Conference on Knowledge Discovery and Data Mining, 1999.

Lavrenko, Victor et al, (2000) *Language models for Financial news recommendation.* In Proceedings of the Ninth International Conference on Information and Knowledge Management, (pp 389-396), 2000.

Macskassy S. (2003) *New Techniques In Intelligent Information Filtering*, PhD Thesis, Rutgers, The State University of New Jersey

Yu, H. Hatzivassiloglou, V. (2003) *Towards Answering Opinion Questions: Separating Facts from Opinions and Identifying the Polarity of Opinion Sentences*, Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing, (pp. 129-136)