Stock Market Polarity Extraction

Justin Nafe
School of Engineering and Computer Science
University of North Texas
Denton TX, 76203-5017.
jjn0032@unt.edu

Abstract

This paper explores the detection of polarity of news articles pertaining to the stock market, or the individual stocks, and how that polarity correlates with the actual stocks in the stock market. The program that this paper discusses is not a stock market forecaster. Stock market forecasters typically analyze numerical data over history, but this prototype explores the textual information from news articles discussing various public stocks. The assumption is that if the consensus expresses positive or negative polarity toward a stock, then the value will go up, or down, respectively. This program looks for subjective terms within news articles and determines the polarity of the document. This unigram method and the use of scales closely relates to work done by Pang et al 2002 and 2005.

1 Introduction

The problem is determining the trend of a stock based on the polarity of the textual data such as news articles.

It is within an individuals interest to figure out what makes a stock tick, but the answer to this problem, being able to automatically classify a document to a topic (a particular stock) and determine the polarity or sentiment of news articles, is more valuable to Natural Language Processing (NLP) researchers and marketing professionals than to the focused domain of the Stock Market.

Marketers or business professionals could use the polarity determining algorithm to gather a consensus on the opinion of a product. For example, if Google decides to go into business with either AT&T or T-Mobile with a new phone, then Google may gather documents that express opinions about the products that each company produces, then rate these products on a scale based on the polarity of the opinions. Google may factor this into their decision.

Small business may want to run the algorithm over the web about a product before they deploy a similar product, in hopes that a common flaw that consumers despise will be exposed, and then the small business can fix the problem before it goes public.

Polarity detection involves the determination of the context's sentimentality, either by looking for key words that hold polarity by themselves (unigram), or by a more contextual approach of considering phrase, sentence, or document polarity (Wilson et al. 2005). For the purposes of this program, unless explicitly stated, a document is an online news article, forum, or blog. The product summation of the polarities for each company using each method will determine the flow of the stock of choice. Then an overall assessment of the stock market will be calculated.

2 Related Works

Many studies in the past retrieve a binary (*Positive – Negative)* sentiment of reviews, classifying documents based on sentiment instead of topic (Pang et al., 2002). This is relevant to the project in that the stock can either go up or down or stay the same.

Pang and Lee latter use a scaling system for the sentiment classification (Pang et al., 2005). My degree of confidence will be used as my scale of polarity.

Another approach to determine polarity is to determine contextual polarity by analyzing polar phrases (Wilson, 2005), (Kwon, 18th Annual International Digital Government Research Conference).

Different types of writing call for different methods of sentiment retrieval. An empirical study on the variations of different types of writing suggests that there is a difference (Biber, 1988).

3 Method

The algorithm is a simple calculation of the number of subjective words and the intensity of the polarity of the word when a stock is discussed in the news.

The program reads in subjective terms provided by the OpinionFinder (Weibe et al. 2006) and the annotated polarity for the term. This information is stored in a vector for later retrieval.

The program will have to crawl the web to retrieve information pertaining to stocks. I will restrict my initial domain to yahoo domains such as http://biz.yahoo.com/topic/us-markets/ and http://finance.yahoo.com. The crawler grabs the links from the visited pages and stores the links onto a stack for later retrieval.

The current program does not contain a separate algorithm to help determine the topic of the web page, so the program simply looks for the stock symbol or the company name pertaining to the symbol. If the contents of the web page consist of the symbol or the company name, then the program proceeds with the following algorithm.

Given the problem of not having an annotated corpus specific for the domain of business or of the stock market, I decided to use a unigram method of utilizing the polarity of subjective terms provided by the OpinionFinder lexicon from the University of Pittsburg. This lexicon contains a list of subjective terms annotated with the degree of polarity.

After the crawler finds a stock symbol or company name of a stock, the program looks for subjective terms. If a subjective term is found, then the corresponding numerical weight is added to the symbol's polarity scale. Adjacent words are considered, and if an adjacent word is negative, then the polarity, not the strength of the polarity is inverted. The numerical weight associated with the subjective term, gathered from the lexicon, is as follows:

"strongpos" = 5
"weakpos" = 4
"weakneg" = 2
"strongneg" = 1

The number 3 is reserved for latter development to better match the common terms of the stock market ("Strong Buy, Buy, Hold, Underperform, and Sell) (http://www.nasdaq.com/reference/glossary.stm). The sum of these weights is associated with the stock symbol in a vector form. The number of subjective terms for the symbol is also recorded in the vector. If the crawler comes across another page about the particular stock, the weights and counts are summed with the weights and counts already calculated for the stock. For the output, the summation for each stock is divided by the count of subjective terms associated with the stock. This results in a scale of 1 to 5, where 1 and 2 mean that the stock is decreasing and 4 and 5 mean that the stock is increasing.

4 Evaluation

An evaluation of your algorithm on a gold-standard data set. Describe the data set, how it was created (or include a reference to previous work, if the data set was created by other people), provide some statistics relevant to the data set.

The gold standard used to measure the precision and recall of this algorithm was the closing prices of each stock in the Nasdaq Index for November 18, 2008 and December 9, 2008. The data is downloaded from http://www.eoddata.com/ and consists of the values of the stocks. I chose to use the aforementioned dates (the span of two weeks from the current date) to give the stock time to adjust to the news. To determine whether the stock is increasing or decreasing, the closing value of November 18th is subtracted by the 9th of December's closing value. For example, if I use the same date (the difference between the opening value and the closing value for December 9th), then the precision drops to the range of 55% and 48%.

For the two week span, the precision ranges from 66% to 51%, and this value changes throughout the day as the news articles change.

5 Discussion of Results

Nearly every time that the program runs during normal business hours of weekdays, the precision and recall vary, likely due to the change in the news articles visited by the crawler.

From Rada Mihalcea's suggestion of just considering the strongly positive and strongly negative, after commenting out the stocks with the scale ranging from 2 to 4, the precision of the limited crawl jumped to 100%. So I conducted a complete crawl, resulting in 84%!

Developing the crawler was the most difficult component of this program to develop. As explained before, I used the HTML library to grab textual data and browse to new links gathered by LinkExtor (developed by D.H PodMaster 2004), and to keep from getting stuck on a page I set the parser to timeout. Revisiting sites, entering loops, and leaving the desired domain were other problems that were dealt with by better code design.

6 Conclusions and Future Work

and symbol, and the industry that the company is in will be used to detect the relevant topic of the document.

The extraction of the polarity is only one factor in determining whether or not to purchase stocks or sell stocks.

Another feature that might help with the polarity extraction would be a Bayesian classifier or some other method to classify or index the document about a company that's been analyzed. If I were to accurately determine the topic of the document, the precision would increase. If I can accurately tag documents for the topic and the polarity, then dump the positive tagged documents into the positive bucket and the negative type documents and the negative bucket. Then maybe I can build my own corpus for polarity detection. There's some research being done about this.

Once the natural language processing computation is perfected, I can integrate this program into a renowned stock market forecaster, preferably open source, and build a simulator. The final product of the natural language processing aspect of forecasting in the stock market will only be a factor in determining investment in the stock market.

7…References

Douglas Biber,. Variation across speech and writing. Imprint Cambridge [Cambridgeshire]; New York, NY, USA : Cambridge University Press, 1988.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. Thumbs up? Sentiment Classification Using Machine Learning Techniques. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP-02),* pages 79–86, Philadelphia, Pennsylvania, July 2002. ACL

Bo Pang and Lillian Lee. Seeing Stars: Exploiting Class Relationships for Sentiment Categorization with Respect to Rating Scales. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*, 2005.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis. In *Proceedings of Human Language Technologies Conference/Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP* 2005), pages 347–354, Vancouver, BC Canada, October 2005. ACL Press.

Soo-Min Kim and Eduard Hovy, Determining the Sentiment of Opinions. In *Proceedings of the Coling conference, Geneva, 2004*

Peter D. Turney. Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), Philadelphia, July 2002, pp. 417-424.*

Namhee Kwon, Liang Zhou, Eduard Hovy, Stuart W. Shulman, Identifying and Classifying Subjective Claims. In *Proceedings of the 8th Annual International Digital Government Research Conference.*

Janyce Wiebe, Theresa Wilson, and Claire Cardie (2005). Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation* (formerly *Computers and the Humanities*), volume 39, issue 2-3, pp. 165-210.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann (2005). Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis. *Proceedings of HLT/EMNLP 2005*, Vancouver, Canada.

Jan Wiebe's Natural Language Processing (NLP) Group at the University of Pittsburgh 2006. (Provider of the OpionionFinder Lexicon.

Theresa Wilson, Janyce Wiebe and Paul Hoffmann (2005). Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis. Proceedings of Human Language Technologies Conference/Conference on Empirical Methodsin Natural Language Processing (HLT/EMNLP 2005), Vancouver, Canada.

Janyce Wiebe and Ellen Riloff (2005). Creating subjective and objective sentence classifiers from unannotated texts. Sixth International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2005).