A Comprehensive Review of Sentiment Analysis of Stocks

Pranav Bapat
Department of Information Technology
Pune Vidyarthi Griha's College of Engineering and Technology

ABSTRACT

This paper comprehensively studies the sentiment analysis of stock market news and explains the maturity of sentiment analysis in the stock market scenario. I explain in its entirety the procedure, difficulties, and limitations concerning sentiment analysis of financial news to predict stocks. The stock market movements are regarded as highly unpredictable and a large number of factors contribute to that unpredictability. Factors such as market sentiment, government policies and company announcements are some of the major contributing factors however the list is not exhaustive. Technological advancements over the past two decades has enabled researchers and market professionals to develop mathematical models to optimize their returns and keep the risk in check. These advancements have given way to social media platforms especially Twitter to more conveniently express opinions and reviews. Narrowing it only to the stock market and financial scenario, Twitter is an attractive platform for the user community to discuss company health, company announcements, major news, and government policies etc to name a few. Companies or organizations in turn also boast their success on Twitter. This entire process of sharing news and opinions yields a large amount of financial data to be searched for an overall sentiment or a possible prediction of the stocks. This paper attempts to dive deep into the specifics of procedure, the current stage of its maturity, and the importance of Machine Learning in finance and the stock market.

General Terms:

stock market, sentiment analysis

Keywords:

stock market analysis, sentiment analysis

1. INTRODUCTION

An active area of research is the prediction of the stock market. Efficient Market Hypothesis (EMH)[1] dictates that the stock market is largely driven by new information and it is a central paradigm governing stock markets around the world. In addition to the stocks, the same treatment can be applied to the currency market. Studies have shown the correlation between market sentiment and weekly or monthly returns. [4] The major challenge when dealing with this information is the extraction of patterns from a large data set and attempting to predict the value of a security. This information is available on the Internet on a plethora of platforms however Twitter is the most popular of those platforms given the amount of users and comments it attracts. A procedure called sentiment analysis [2] is carried out on a set of Tweets about an organization and further it is used for price prediction of its security. It is therefore sufficient to say that a correlation between investor sentiment and market sentiment exists and can be exploited to optimize investment strategy. Sentiment analysis aims to identify and extract user generated opinions given a certain topic. The sentiment analyzer outputs the polarity of users about a certain topic as positive, or negative. This polarity lends perspective on the overall sentiment about an organization and can be a catalyst in determining the stock price of that company. [3][5] The rest of the paper is organized as follows. The second section briefly describes the algorithms and procedures in place to handle user generated content. The section that follows concentrates on the shortcomings on the current analysis model. Final section attempts to draw conclusions and make way for future work and explain its scope.

2. SENTIMENT ANALYSIS MODEL

2.1 Basic Sentiment Analysis

This section briefly describes the sentiment analysis procedure. The process begins with maintaining a corpus of financial words of about 5000 words. The words are organized according to their polarity i.e. positive or negative. Then given a set of Tweets, each Tweet given the pre-processing and cleaning treatment wherein the message is separated from stop-words, white-spaces, punctuation and emoticons. The clean message is then forwarded to a machine learning algorithm to predict the sentiment of that message. The algorithm then checks the occurrence of positive and negative words in a message and outputs the sentiment accordingly. The specifics of the algorithms will be explained in later sections. The algorithm should be selected on the basis of the following criteria

(1) Fast: The algorithm selected must be fast and efficient for it has to handle massive amounts of real time data

(2) Accurate: The algorithm must also be accurate in predicting the sentiment of a message. Incorrect sentiment will lead to a deviation in results.

(3) Real-valued: The algorithm should produce a real time quantity and therefore must be able to quantify stock movements. Generally regression algorithms are implemented for such requirements.

2.2 Algorithms

Sentiment analysis makes use of machine learning algorithms. Machine learning deals with a training data set against which new incoming data is compared to predict the result. This process is known as supervised learning where the system is first supplied with correct answers and further expected to classify based on past experience or knowledge. Another category of machine learning algorithms is regression. Unlike classification algorithms which produce discrete outputs, regression algorithms produce real values. Stock market analysis makes use of both classification and regression algorithms as follows. Classification algorithms produce discrete values such as positive or negative given the content of the message. The algorithm makes use of the corpus of words against the words used in a message to output whether the message is positive or negative. A tweet that contains more positive words than negative words from the corpus is said to be positive. There are mainly two classification algorithms used in sentiment analysis of the stock market namely Naive Bayes, and Support Vector Machines (SVM).

(1) Naive Bayes: Bayesian classifiers are statistical classifiers which can predict class membership probabilities such that a given message belongs to one of the classes. Naive Bayes algorithm can be used to predict the index of the stock market using previous index values. The algorithm maintains volume, value, and index among numerous other factors. Using this as training data, the algorithm can predict the value of the index to a certain extent. The same procedure can be applied to a single stock based on the content of the message. If a sequence of messages about an organization indicate positive news or negative news, the algorithm can accurately predict the overall sentiment about that organization. Generally speaking, Naive Bayes assigns a document dj (represented by a vector dj) to the class ci that maximizes by applying Bayes rule as follows, Thus Naive Bayes classifier classifies each message into the most relevant class by performing probabilistic analysis.

(2) Support Vector Machines (SVM): Another popular machine learning algorithm is SVM mainly used for classification however can be altered to produce a continuous value. SVM plots all features in a message on an axis and attempts to classify the set of messages by drawing a hyper-plane. Messages on one side of the hyper-plane indicate positive messages or sentiment and messages on the other side indicate otherwise. Referring to the stock market scene, given a certain topic, the algorithm plots each message on the basis of occurrence of positive and negative words in each message. After the plotting, SVM will output a hyper-plane that splits evenly the messages and thus classify messages. In SVM, the further distance between the point and hyper-plane is, more confident we are for the prediction we made, whereas, our prediction cannot be very accurate when the point is close to hyper-plane.

3. LIMITATIONS

Sentiment analysis of stocks makes management of portfolios simple however the process is far from being rock solid. The limitations that exist are categorized as follows:

(1) Algorithmic limitations:

(a) The algorithms that play an important role in the actual analysis are far from perfect. Naive Bayes and SVM are sufficient for sentiment analysis however do not perform on all the conditions. It is observed that depending on the sentence structure of the message i.e. usage of punctuation and emoticons both algorithms perform analysis with varying accuracy. Naive Bayes performs probabilistic analysis but does not factor in the phrasing of the sentence and the aggregated meaning of that sentence. Words when used in conjunction with other words might have varied meanings and Naive Bayes algorithm fails to take account of it. Words in a sentence are an important indicator of the sentiment but the collaborative meaning and significance in the sentence should also be considered equally. SVM also has its own different shortcomings. SVM is very sensitive to the size of the training data set. When the training data set is inadequate, the hyper-plane found by SVM might be a deviation from what is expected. Both algorithms are highly dependent on feature selection which leads to different results on changing the feature set. The stock market is increasingly volatile and what might seem an irrelevant piece of news might turn out to have the closest relation. Thus the algorithms have to deal with a varying number of features and finding the optimum set of features is a major challenge. The machine learning algorithms also fail to output specific polarities of the message. A piece of news can be extremely positive or extremely negative but the algorithms give a vague sentiment.

(2) Market limitations

(a) It is generally said that whatever is already in the news is already factored into the market. So if company ABC has decided to buy an oil block which will make them millions, the market has already filtered the results. This provides a very small leeway for investors to capitalize on that information. This is where Efficient Market Hypothesis comes into the picture. EMH states that market prices immediately reflect all available information. Three popular versions of EMH exist namely weak, semi-strong and strong. The weak form of the hypothesis is rather pessimistic. It states that no profit can be made by reviewing at publicly available information thus rendering sentiment analysis moot. The semi-strong hypothesis argues that profit can be made by analyzing data that is not publicly available thus not reflected in the market. Thus EMH presents a big problem and working around it to optimize returns is a major challenge. Another possible limitation to sentiment analysis is failure of a message to actually be factored into the market. The possibility of a message being positive or negative not affecting the market exists and sentiment analysis currently cannot mitigate this problem.

(3) Data limitations

(a) Data limitations refer to the inadequacy of the data to be analyzed. Hundreds of thousands of users express opinions about companies, policies and announcements. However not all messages are relevant to the happenings in the market. Maintaining relevancy of the data to the current market scenario is a difficult task. Noise is another negative when it comes to analyzing messages online. The signal to noise ratio is very low and current sentiment analyzers are not prepared for it. Also current sentiment analysis tools base their findings on specific sources such as news, tweets, and blogs etc. However only one platform as a data source provides incomplete information about the market. Thus a collaboration of sources must be utilized for efficient and unbiased market analysis. The format of the data is biggest of the problems. News agencies usually employ info-graphics for their news, organizations circulate newsletters and investors use online platforms. Each user uses a variety of formats ranging from articles, and tweets to videos and info-graphics. The failure to incorporate the different formats might deviate the result and does not do justice to the analysis.

4. CONCLUSION AND FUTURE WORK

Sentiment analysis of stock market for predicting the indexes and values is very young however improving at light speed. Sentiment analysis is the procedure of analyzing text based information from online messages or news articles and predicting the overall sentiment. In the stock market situation text messages are analyzed in order to gain insight into the current happenings in the market and predicting returns. New platforms and tools utilizing machine learning algorithms among other techniques are emerging and bound to change the stock market scenario. The algorithms, Naive Bayes and Support Vector Machine (SVM) are basic machine learning algorithms currently used however hybrid versions are upcoming. It is correct to state that sentiment analysis techniques are keeping up with the advancements in technology but the market is ruthless. The stock market is volatile and unpredictable to say the least. With a large number of fundamentals such as P/E Ratio, volume, EPS, and Dividends and yield the market presents many challenges along the way and analysis is extremely tough. The limitations mentioned above are evidence of the multitude of problems that exist when dealing with the market. Striking a perfect balance between the fundamentals, real-time continuous news feed and investor sentiment is key for sentiment analysis to make its mark and become an important tool in analysis. Given the maturity of the current analysis tools available, there is room for improvement on many fronts. Improvements to mitigate the limitations and bettering performance is key to optimum sentiment analysis. Researchers are looking into hybrid algorithms, hybrid pre-processing techniques, and understanding more about the stock market in order to optimize the returns. These approaches attempt to deal with a variety of data sources and its collaboration. Overcoming EMH is also a huge challenge and market analysis tools cannot ignore its significance. Thus the problem of finding an optimal formula for handling all the market variants is tough and only after a thorough analysis will sentiment analysis thrive. In conclusion, the sentiment analysis of the stock market has potential to change the investing world but requires a lot of research to mitigate the limitations. Current tools are adequate for basic analysis in order to assess the health of the market but for investors to make profit certain adjustments have to be incorporated into the analysis.[6]

5. REFERENCES

[1] Azar, Pablo D. Sentiment Analysis in Financial News. Thesis. Harvard College, 2009. N.p.: n.p., n.d. Print.

[2] Bo Pang, Lillian Lee, Shivakumar Vaithyanathan, 2002. Thumbs up?: sentiment classification using machine learning techniques, Proceedings of the ACL-02 conference on Empirical methods in natural language processing, p.79-86

[3] More than words: Quantifying Language to Measure Firms' Fundamentals". Paul Tetlock, Maytal Saar-Tsechansky, and Sofus Macskassy Journal of Finance. Forthcoming.

[4] K. Rao and A. Ramachandran, "Exchange Rate Market Sentiment Analysis of Major Global Currencies," Open Journal of Statistics, Vol. 4 No. 1, 2014, pp. 49-69. doi: 10.4236/ojs.2014.41006.

[5] Xiang, Bing, and Liang Zhou. "Improving Twitter Sentiment Analysis with Topic-Based Mixture Modeling and Semi-Supervised Training." Association for Computational Linguistics (2014): 434-39. Print.

[6] Zheludev, I., Smith, R. and Aste, T. When Can Social Media Lead Financial Markets? Sci. Rep. 4, 4213; DOI:10.1038/srep04213 (2014).