STM JOURNALS

Journal of Artificial Intelligence Research & Advances
ISSN: 2395-6720 (Online)
Volume 6, Issue 2
www.stmjournals.com

Sentiment Analysis and Stock Market Prediction: Using News to Predict Stock Markets

Shah Abrar Amin[1,]*, Dhajvir Singh Rai[2]
[1]*Student, Department of Computer Science and Engineering, Dev Bhoomi Institute of Technology, DBIT, Uttarakhand Technical University, Naugaon, India*
[2]*Professor, Department of Computer Science and Engineering, Dev Bhoomi Institute of Technology, DBIT, Uttarakhand Technical University, Naugaon, India*

Abstract
The aim of this paper is to propose a relatively new approach regarding the influence of financial news on stock market prices. Previously, a lot of effort and research has been done to study the effect of historical stock market prices. This approach mainly involved time series numerical data to predict the future trend of stock markets. But, with the advent of superior computing power and computing techniques such as Natural LANGUAGE Processing, SVM etc., it becomes imperative to look beyond just numerical data. Thus, paper is one such effort to use textual data in the form of financial news to predict the market trends. The idea behind using textual data is that text inherently contains words that convey the sentiment of that text. The sentiment may either be positive or negative. Therefore, the idea is to analyse the sentiment of the text to classify the text as being positive or negative with the assumption that positive news will affect markets positively and negative news will affect markets negatively. However, textual data alone cannot be a good predictor of stock markets as there a number of other factors that influence stock markets. Hence, we do not focus on just textual data in this paper. We also take into account numerical aspects such as opening and closing prices of markets in confluence with the textual data to predict stock markets. This paper also presents a review of some of the most recent and most efficient research that has been done to find out the effect of public mood or sentiment on the stock markets. This paper also presents various findings of such researches.

Keywords: *Stock markets, social media, sentiment analysis, machine learning*

***Author for Correspondence** E-mail: shahabraramin@gmail.com

INTRODUCTION

Sentiment analysis or opinion mining is a name given to a host of techniques whose end goal is basically efficient classification. The object of such a classification is either text or speech. The aim of such a classification is to determine the opinions expressed in the given documented text. In other words, sentiment analysis involves the classification of a piece of text into three broad categories-positive, negative or neutral. Sometimes, these three dimensional classifications may not suffice, in which case, classification takes place along several dimensions of mood such as sad, angry, calm, happy etc. [6-8].

Sentiment analysis entails the following:

1. **Identification of subjective/objective content:-** Here, the goal is to classify the given text into two broad categories-subjective and objective. Subjective text is that text, which is subject to multiple interpretations, whereas, objective text involves factual information which has singular meaning. An important task here is the removal of objective sentences from the text which helps improve performance of the classifier [9-15].
2. **Aspect and features:-** An aspect or feature is an attribute. Every entity has different components or attributes which constitute its aspects or features. Here, we look at various aspects/features of an entity in isolation. This helps in getting a better insight into the entity as a whole and also helps in capturing its various nuances e.g., a mobile phone may have very good build quality but poor camera

performance. So, camera and build quality form two features of the entity namely mobile phone.

STOCK MARKET PREDICTION

Stock market prediction approaches can be categorised into the following broad categories:-

1. **Rudimentary/minimalistic approach:-** The basic premise here is that if a company performs well, more capital should be pumped into that company which in turn would cause a rise in the company's stock values/prices. This is a reasonable and plausible argument and is widely used .
2. **Time series analysis:-** Here, the future value of a company's stocks is approximated merely on the basis of the company's historical stock prices i.e., historical values of stocks are used to predict/calculate future projections of stocks.
3. **Machine learning:-** Machine learning is a name given to a variety of algorithms/techniques that are pivotal in any classification and stock market prediction is no exception. Machine learning techniques such as neural networks and genetic algorithms are used for stock market prediction. Neural networks deserve a special mention because of their ability to produce highly accurate prediction models. Commonly, feed-forward networks are used that make use of back-propagation algorithms to reduce error by re-assigning network weights , thereby increasing accuracy.

Applications of Sentiment Analysis

At the outset, there may not seem to be a direct relationship between sentiment analysis and stock market prediction. But, in light of this paper, these two go hand-in-hand. Sentiment analysis refers to the process of very closely and very minutely analysing any text in order to get a sense of the writer's/author's mood, so that a decent conclusion could be drawn as to whether the author/writer is speaking for or against the topic at hand. Sentiment analysis has far reaching applications, a few of which find mention below:-

Health

Emotion detection/sentiment analysis finds use in detecting certain psychological conditions such as anxiety.

Elections and Politics

This is one of the most critical applications of sentiment analysis/opinion mining where the mood of the general public towards politicians, their affiliations, like or dissent towards a certain govt. policy or the general mood of the electorates can be gauged.

Businesses

Large businesses like Amazon are spending billions of dollars to mine people's opinion in a useful manner which in turn would help them to better serve their customers. Social media is the *de-facto* place to get this data which could then be used for sentiment analysis/opinion mining [16-18].

Stock Market Prediction

Social media feed, especially Twitter feed, has been used for stock market prediction by subjecting the Twitter feed to opinion mining/sentiment analysis tools. The motivation for such an exercise is to investigate the role of public mood in predicting stock market prices (behavioural economics).

STOCK MARKET PREDICTION

Stock market prediction has mainly and traditionally been based upon the historical statistical/numerical data, such as stock market indices. Using the same type of data, i.e., stock market indices, different machines/algorithms have been devised with the common goal of achieving higher accuracy in predicting stock markets.

The efficient market hypothesis asserts that the stock market prices follow a random walk and thus cannot be predicted using any historical data. Another obstacle that comes in the way of predicting stock markets is the day-to-day financial news. The day-to-day news greatly impacts stock markets. Since this news is dynamic and therefore, cannot be predicted, there seems to be no way to find a correlation between the news and stock market prices. In simpler terms, news cannot be used to predict stock markets [19, 20].

Using News to Predict Stock Markets
This study undertakes this Herculean task of using news to predict stock markets. However, news cannot be directly used for prediction. Hence, the following two modules are used:
Module-I
The foremost step is the classification of news. This means developing a classifier which would classify news. News can be classified as market friendly or market unfriendly along different dimensions of friendliness or unfriendliness. This is where sentiment analysis plays a big part. So, module-I mainly deals with developing a classifier that would classify news according to its favourability to the market.

Module-II
This module deals with the attainment of a dataset which would consist of historical news over a certain period of time and then using the results of the classifier from module-I to find the closest match from this dataset of historical news e.g., If a price hike of one rupees in petrol is today's news. After proper sentiment analysis of this news, the closest match to this news would be found from the already classified historical news dataset in terms of its resemblance to the historical news with respect to the effect on the stock market prices, with the assumption being that similar news has almost similar effect on the stock markets prices.

Levels of Sentiment Analysis
Sentiment analysis can be carried out at various levels differing in their degrees of minuteness. The three main classification levels are:

Document Level
Here, the denomination of classification is the entire document which usually contains information about only one topic.

Sentence-Level
Documents contain sentences which are either subjective or objective. At this level, objective sentences are gotten rid of and classification is carried out on subjective sentences.
Aspect-Level
Here, the unit of classification is an aspect/attribute/feature of an entity, e.g. a restaurant may have great interiors but distasteful food.

CLASSIFICATION OF SENTIMENT ANALYSIS APPROACHES
Machine Learning
Machine learning based approaches are preferred because of their adaptability and accuracy. The machine learning technique comprises of the following steps:
1. Data Acquisition: The data that is to be analysed is collected from mostly social media sources such as blogs, twitter, Facebook, etc.
2. Data Pre-processing: Having collected the data, this data needs to be refined and gotten rid of unwanted elements such as URLs, usernames, grammatical mistakes, stop words, punctuations, etc.
3. Training data: Data is collected using crowd-sourcing methods and is then hand-labelled. This data is then fed to the classification algorithm for learning purpose.
4. Classification: After training a suitable technique such as SVM or Naïve Bayes is employed for the analysis. Having completed the training, the classifier is ready to be deployed for sentiment extraction of real time tweets/posts.
5. Results: Results are displayed using charts, bars, graphs, etc. and performance tuning is done accordingly.

Machine learning makes use of two sets of documents/datasets: training set and a test set. The training set is used by the machine to learn those features/characteristics of a document which stand out and can be pivotal for differentiating purposes. The test then is used to check how well the learning was done and is also used to check the performance of the classifier.

Features of Machine Learning Technique for Sentiment Classification
The features of machine learning technique for sentiment classification are:
i. Presence of various n-gram terms such as uni-grams, bi-grams, tri-grams and their frequency.

ii. Parts of speech information is used to remove any ambiguity which acts as a guide in feature selection, e.g., adjectives and adverbs are observed to play a great role in expressing sentiments/opinions.
iii. Negations: They tend to overturn the sentiments expressed by positive or negative sentiment words.

Lexicon Based Approach

In this approach, dictionaries containing opinion words are used. These dictionaries are called sentiment dictionaries. The sentiment words in these dictionaries are matched with the data for classification purposes so that the data can be classified into positive or negative categories. A tokenizer is used to convert the text (used as input) into tokens. Different techniques/approaches used to build a sentiment lexicon are:

i. Manual construction,
ii. Corpus-based methods, and
iii. Dictionary-based methods.

Manual construction is labour intensive, arduous and takes time. Highly accurate opinion words are produced due to corpus-based methods. Finally, in dictionary based approach, seeding is used. A small group of words is collected to form a set. The polarity/inclination/orientation of the collected words is already known. This set of words is then allowed to grow and mature and this is done by searching for synonyms and antonyms of the seed words. WordNet is consulted for finding synonyms/antonyms.

Hybrid Approach

Finally, the hybrid approach uses an amalgamation of machine learning and lexicon-based approaches. The idea is to improve performance by taking advantage of the accuracy of machine learning techniques and the speed of lexicon-based techniques, resulting in an efficient hybrid technique. We create imitative/pseudo documents which consist of all the words from the chosen lexicon. We calculate cosine similarity measure between the imitative documents and the unlabelled documents. Depending upon the degree of similarity, the documents are given either a positive or a negative label. This training dataset is then trained by feeding it to a Naïve-Bayes classifier.

CHALLENGES IN SENTIMENT ANALYSIS

Sentiment analysis has been on the radar of researchers for the past many years now. Although significant advancements have been made in sentiment analysis, there are a few issues that remain to be dealt with and they constitute the scope for future research. With the use of social media for sentiment analysis, these challenges have been growing. Some of the challenges faced by sentiment analysis are:

1. Named Entity Recognition: NER comes under the purview of information extraction and its goal is to identify/recognise named entities in a text and then classify them into appropriate and pre-defined categories such as persons or organisations, e.g., identifying whether apple refers to the tech company or the fruit.
2. Anaphora resolution: It involves resolution of what a pronoun or noun refers to, e.g., we went to a show and then had some dinner. It was distasteful. The answer to the question 'What does "it" refer to here?' is called anaphora resolution.
3. Parsing: Parsing involves resolving a sentence into its component parts and describing their syntactic roles. In the context of sentiment analysis, parsing means identifying the subject and object of a sentence and resolving issues like which one does the verb or adjective actually refer to.
4. Sarcasm: Sarcasm is one of the hardest challenges for sentiment analysis. Very often in social media, people say things they do not mean with the aim of mocking or reiterating their point of view. Identifying sarcasm in a text is a daunting task that still remains a challenge. Also, if we do not know the context of the text, we will have no way of knowing whether a particular means good or bad, e.g., unpredictable when used in a movie review is a positive thing. But, if the review of a vehicle contains the word 'unpredictable', it is a negative thing.

5. Finally, Social media slangs and abbreviations and grammatical and spelling mistakes pose a very big challenge for sentiment analysis.
6. Utterances may convey emotions in a variety of ways and to varying degrees of intensity.
7. Utterances may represent emotional events without making it clear how the speaker feels about the event. This problem is more significant when dealing with texts as there is no way of inferring sentiments unlike in speech where the tone of the voice could provide a clue about the emotional state of the speaker.
8. Lack of para-linguistic information: Written text does not contain annotations of stress and emotions which are conveyed through tones of voice or facial expressions.
9. Lack of large amounts of labelled data: Machine learning algorithms require large amounts of training data (labelled data). However, the sheer range of human emotions is so great and much of the work on sentiment analysis has been carried out using only a handful of emotions and valence categories.

RELATED WORK

Johan *et al.* investigated the role of behavioural economics i.e., whether public mood affects economies at large [1]. Two mood assessment tools were used, namely ‘opinion Finder’ and ‘Google-Profile of mood states (GPOMS)’ and public mood was categorised along seven mood dimensions and the effect of each mood dimension was studied by finding correlation between twitter feeds and the Dow Jones Industrial average (DJIA). Certain mood dimensions had a greater impact as compared to others. A SOFNN model provided better results using a combination of calm and happy, thereby proving the non-linear relation among different mood dimensions (Table 1).

Khan *et al.* proposed a hybrid approach for twitter feed analysis and classification [2]. Twitter feeds were collected and pre-processed. After pre-processing, the classification of these tweets was carried out using the following three classifiers:

1. EEC: Tweets were classified on the basis of emoticons. EEC achieved 70% accuracy.
2. IPC: It used a bag of words approach. 2006 positive and 4784 negative words were taken from Bing Liu list. 354 positive words and 2349 negative words were taken from Bill Mc Donald list. Each word in the list has been classified as positive or negative. Words found in positive are labelled positive. Similarly, negative words were labelled. Words found in both lists were labelled as neutral.
3. SWNC: It used senti-word and assigned weights to different words. The sentiment value of each word is calculated by calling the senti word net.

Each tweet was then subjected to EEC, IPC and SWNC for classification purposes (Table 2).

Table 1: *Statistical Significance (P-Values) of Bivariate Granger-Causality Correlation Between Moods and DJIA in Period February 28,2008 to November 3, 2008.*

Lag	OF	Calm	Alert	Sure	Vital	Kind	Happy
1 day	0.085	0.272	0.952	0.648	0.120	0.848	0.388
2 days	0.268	0.013	0.973	0.811	0.369	0.991	0.7061
3 days	0.436	0.022	0.981	0.349	0.418	0.991	0.723
4 days	0.218	0.030	0.998	0.415	0.475	0.989	0.750
5 days	0.300	0.036	0.989	0.544	0.553	0.996	0.173
6 days	0.446	0.065	0.996	0.691	0.682	0.994	0.011
7 days	0.620	0.157	0.999	0.381	0.713	0.999	0.150

Table 2: *Accuracy of Classifier on Dataset.*

	EEC	IPC	SWNC	Proposed
Dataset1	20%	68%	62%	90%
Dataset2	23%	55%	70%	84%
Dataset3	22%	49%	65%	87%
Dataset4	9%	45%	75%	87%
Dataset5	18%	60%	71%	86%
Dataset6	15%	52%	52%	88%

**each cell entry denotes accuracy of a particular classifier on that dataset.*

White, at University of California, investigated the role of neural network modelling techniques to find any deterministic chaos in economic time series [3]. The focus was on gathering enough evidence against efficient market hypothesis. The IBM daily stock returns were used for this purpose. To begin with, an autoregressive linear model was used, but the EMH could not be refuted. So, a three-layered feedforward network with five inputs and five hidden units was used. Although, better results were achieved, but further evidence was needed to refute EMH.

Alena *et al.* described methods to automatically generate a sentiment lexicon called SentiFul. They proposed the following methods for lexicon expansion [4]:

1. Synonymy relation: Weights of the known lexemes are propagated to the first/direct synonym of already known lexemes.
2. Antonym relation: The first/direct antonym of an already known lexeme is found and the positive and negative scores of known lexemes are assigned to the newly found lexeme but the two scores interchange places.
3. Hyponymy relation: A word is taken is taken from SentiFul and its corresponding hyponyms are retrieved. Sentiment features of the SentiFul are then transmitted to its hyponyms.
4. Morphological modification: This involves suffixing and prefixing the original word and if the newly formed word is already in SentiFul, it is dropped, else it is included.
5. Compounding: Bases existing independently are combined to give rise to new lexemes, e.g. good-for-nothing.
6. Modifiers: Modifiers impact the sentiment features of neighbouring words and modifiers were categorised as intensifying type or reversing type.
7. Modal operators: Modal operators reflect confidence, possibility, permission and modal operators were categorised into modal verbs and modal adverbs.

Compounding achieved highest performance of 99%, 99.5% in assigning dominant positive or negative labels followed by hyponymy relations (97.9%, 98.9%) followed by synonymy (93.3%, 95.4%) and the least performance was yielded by antonymy relation (66.7%, 94.5%) (Ttable 3).

Zhang *et al.* proposed a method for sentiment analysis that used sentiment dictionaries and the calculation of sentiment scores and weights, thereafter [5]. The proposed method has two stages:

1. Construction of Sentiment Dictionaries: It involves pre-processing of text and the subsequent construction of sentiment dictionaries. Pre-processing means refining the text and getting rid of unwanted elements such as http links, usernames, sop words, etc. The following dictionaries were developed after pre-processing was done:

Table 3: *Highest performance Accuracy in Percent.*

Method	**Results on GS-1 Accuracy**	**Results on GS-2 Accuracy**	**Precision Pos Neg**	**Recall Pos Neg**	**F-score Pos Neg**
Synonymy	93.3	95.4	86.2 100	100 93.6	92.6 96.7
Antonymy	66.7	94.5	97.0 90.7	94.2 95.1	95.6 92.9
Hyponymy	97.9	98.9	96.7 100	100 98.4	98.3 99.2
Compounding	99.0	99.5	98.9 100	100 99.0	99.4 99.5
Derivation	94.2	97.8	95.7 99.1	98.5 97.4	97.1 98.3

**each cell represents accuracy in %.*

i. Basic sentiment dictionary:
It is constructed as:

word	part of speech	weight	number
reprove	verb	0	0
honest	adjective	+-3	3

ii. Degree adverb dictionary:-It is constructed as:

word	grade	weight	number
a little over, unduly	very	3	30
a little less, relative	insufficiently	0.5	12

iii. Network word dictionary:-It is constructed as:

word		weight	number
negative word	in vain, not	-1	31
rhetorical word	could it be said that	-2	10

iv. Emoticon dictionary: It is constructed as:

emoticon	weight	number
☺, : D	2	33

v. Relational conjunctional dictionary:-

word	property	weight	number
and, even	progressive	2	9
however, yet , but	turning	0.5	10

2. Sentiment Value Calculation and Classification:
Addition was used for the weigh calculation of emoticons; multiplication was used for the

weight calculation of emotion words with modifiers.

Experimental results showed an 11.86, 10.86 and 8.5% increase over methods that used only one basic sentiment dictionary in positive, negative and neutral classification respectively.

CONCLUSION

Sentiment analysis has been the topic of many researchers in the recent past and a variety of approaches have been proposed as a result of these researches. The fact that sentiment analysis is drawing so much attention could be attributed to its many widespread applications. From predicting certain psychological ailments to politics to large businesses, sentiment analysis seems to have the answer to it all. However, there is one particular field where sentiment analysis is yet to make any significant mark and that is the prediction of stock market prices. Although, much work has been done on stock market prediction but the techniques proposed have been limited to either statistical techniques or more recently, machine learning techniques. The literature about sentiment analysis contains only one paper that used sentiment analysis for stock market prediction, which further reiterates the need and the essence of this research namely, "Sentiment Analysis and Stock Market Prediction". The road to achieving fairly accurate stock market prediction using sentiment analysis is bumpy and one that has been far less travelled. The problem is two-fold:

1) It involves the resolution of the many challenges posed by sentiment analysis and obtaining a reasonably good sentiment analysis algorithm.
2) The other challenge entails dealing with the sheer dynamic and unpredictable nature of the stock markets.

This research undertakes this seemingly difficult task of stock market prediction using sentiment analysis, with the aim of achieving what might seem to be improbable but definitely not impossible.

REFERENCES

1. Bollen Johan, Mao Huina, Zeng Xiao Jun. Twitter Mood Predicts the Stock Market. *J Comput Sci*. 2010; 2(1).
2. Khan Farhan Hassan, Bashir Saba, Qamar Usman. TOM: Twitter Opinion Mining Framework Using Hybrid Classification Scheme. *Decis Support Syst.* Jan 2014; 57: 245-257p.
3. White Halbert. Economic Prediction Using Neural Networks: The Case of IBM Daily Stock Returns. *IEEE 1988 International Conference on Neural Networks.* 24-27 Jul 1988.
4. Neviarouskaya Alena, Prendinger Helmut, Ishizuka Mitsuru. SentiFul: A Lexicon for Sentiment Analysis. *IEEE Trans Affect Comput.* Jan-Jun 2011; 2(1): 22-36p.
5. Zhang S, Wei Z, Wang Y, *et al.* Sentiment Analysis of Chinese Micro-Blog Text Based on Extended Sentiment Dictionary. *Future Gener Comput Syst.* Apr 2018; 81: 395-403p.
6. Mizuno Hirotaka, Kosaka Michitaka, Yajima Hiroshi, *et al.* Application of Neural Network to Technical Analysis of Stock Market Prediction. *Studies in Informatic Control.* Jul 2001.
7. Mohammad Saif M. Sentiment Analysis: Detecting Valence, Emotions, and Other Affectual States from Text. *Emotion Measurement*. 2016; 201-237p.
8. Alessia D, Ferri Fernando, Grifoni Patrizia, *et al.* Approaches, Tools and Applications for Sentiment Analysis Implementation. *International Journal of Computer Applications (IJCA).* Sep 2015; 125(3): 26-33p.
9. Cootner, Paul H (1964) The random character of stock market prices.
10. Fama Eugene F. Efficient Capital Markets: II. *J Finance.* 1991; 46(5): 1575-1617p.
11. Miller George A. WordNet: A Lexical Database for English. *Commun ACM*. 1995; 39(1): 1-3p.
12. Moilanen Karo, Pulman Stephen. The Good, the Bad, and the Unknown: Morphosyllabic Sentiment Tagging of Unseen Words. *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers*. 2008.

13. Balahur Alexandra. Sentiment Analysis in Social Media Texts. *Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*. 2013.
14. Read Jonathon. Using Emoticons to Reduce Dependency in Machine Learning Techniques for Sentiment Classification. *Proceedings of the ACL Student Research Workshop*. 2005.
15. Wiebe Janyce, Mihalcea Rada. Word Sense and Subjectivity. *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*. 2006.
16. Baillie RT. Econometric Tests of Rationality and Market Efficiency.8(2) 1988: 151-186p.
17. Turney Peter D, Littman Michael L. Measuring Praise and Criticism: Inference of Semantic Orientation from Association. *ACM Trans Inform Syst.* Oct 2003; 21(4): 315-346p.
18. Feng Shi, Song Kaisong, Wang Daling, *et al*. A Word-Emoticon Mutual Reinforcement Ranking Model for Building Sentiment Lexicon from Massive Collection of Microblogs. *World Wide Web*. Jul 2015; 18(4): 949-967p.
19. Agarwal Deepak, Chen Bee-Chung, Pang Bo. Personalized Recommendation of User Comments via Factor Models. *Proceedings of the Conference on Empirical Methods in Natural Language Processing.* 2011.
20. Bermingham Adam, Smeaton Alan. On Using Twitter to Monitor Political Sentiment and Predict Election Results. *Proceedings of the Workshop on Sentiment Analysis where AI meets Psychology (SAAIP 2011)*. 2011.
21. Ghazi Diman, Inkpen Diana, Szpakowicz Stan. Detecting Emotion Stimuli in Emotion-Bearing Sentences.*International Conference on Intelligent Text Processing and Computational Linguistics*. 2015.

Cite this Article

Shah Abrar Amin, Dhajvir Singh Rai. Sentiment Analysis and Stock Market Prediction: Using News to Predict Stock Markets. *Journal of Artificial Intelligence Research & Advances.* 2019; 6(2): 17–24p.

