SentAMaL- A Sentiment Analysis Machine Learning Stock Predictive Model

SA Bogle[1], **and WD Potter**[2]
[1]Computer Science Department, University of Georgia, Athens, Georgia, USA
[2] Artificial Intelligence Institute, University of Georgia, Athens, Georgia, USA,

Abstract - *Social media comments have in the past had an instantaneous effect on stock markets. This paper investigates the sentiments expressed on the social media platform Twitter and their predictive impact on the Jamaica Stock Exchange. A hybrid predictive model of sentiment analysis and machine learning algorithms including decision trees, neural networks and support vector machines are used to predict the Jamaica Stock Exchange. The architecture created, SentAMaL, investigated the impact of sentiments on medical marijuana legalization on relevant stock indices. Due to the unstructured nature of tweets, a customized pre-processing routine was developed prior to determining sentiment and to perform the prediction. Experimental results show 87% accuracy in the movement prediction and 0.99 correlation coefficient for price prediction.*

Keywords: *sentiment analysis, stock prediction, pre-processing*

1 Introduction

Social media comments have in the past had a rapid effect on stock markets [1], [2]. This paper describes the use of sentiment analysis and machine learning (ML) techniques applied to social media comments coupled with historical stock data to predict the Jamaica Stock Exchange (JSE) *in the short term.* It builds on an earlier work [3] which gained 90% accuracy in the movement prediction and 0.95 correlation coefficient for price prediction on applying machine learning approaches to the JSE.

1.1 Background

The JSE was incorporated as a private limited company in August 1968 and the stock market began operations six months later in 1969. While the JSE is not considered a major market, it has been defined by Standard & Poor's as a frontier market. Supervised learning algorithms such as Support Vector Machine (SVM) and Artificial Neural Network(ANN) are more accurate than generic statistical models such as regression and presents a more accurate stock prediction model on the JSE dataset [3].

1.2 Contribution

Social media is inherently assistive in predicting stock trading volumes since it captures the views of many within the population and tweets and posts often go viral in very miniature increments of time. While studies such as [4], [5] and [6] show that volume shifts can be correlated with price movements, text mining prediction studies have not generally focused on the regression problem of predicting prices. Also, while social data have been used to predict economical outcomes in studies such as [7], these predictions are not in the context of financial markets.

2 Literature Review

This section outlines the research to date on sentiment analysis from sources such as news and social media data and use of semantic web architectures for stock prediction. It also compares these approaches with traditional non-sentiment based ML methods.

2.1 News Analysis for Stock Prediction

Wuthrich's group [8] analyzed news articles, collected from five popular financial websites, available before the opening of the Hong Kong stock market with text mining techniques including k-nearest-neighbor and a variety of neural networks. In predicting the trend of the Hong Kong market, they achieved an average accuracy of 46%, which proved better than the accuracy of a random predictor, which achieved a maximum 33% accuracy. News Categorization and Trading System (NewsCATS), was designed to predict stock price trends for the time immediately after the publication of press releases and achieved an average weighted recall of 54% [9]. These accuracy values are low in comparison to most ML-based studies which report between 80%-90% accuracy on binary predictions of market trends.

Wang et al [10] developed an ontology for knowledge about news in financial instrument markets and suggested the framework can be used as input to stock price prediction algorithms. In their later work, [11], which utilized the previous ontology, an expert reasoning system was designed to integrate the domain knowledge in the data mining process through building data mining models consisting of multi news variables with certain financial instrument trading activity and suggesting the potential polar ("positive", "neural", "negative") effect of each news variable, on trading activities.

However, no results were presented regarding accuracy of actual price predictions based on the model.

The AzFinText system [12] investigated whether subjectivity and objectivity impacted stock news article prediction. Based on sentiment analysis it was found that subjective news articles were easier to predict in price direction by 9%, while articles with a negative sentiment were easiest to predict by 1%.

Most web mining or ontology based approaches do not report accuracy of predictions, nor error rates, which is typical of statistical and machine learning papers. This makes the comparison between the models challenging as standard benchmarks are necessary for comparative analysis. However, a conceptual or qualitative analysis is given in the proceeding section since quantitative measures are not available from most semantic related financial forecasting studies.

2.2 The Case for Social Media Input

Markets will react quicker based on the efficient market hypothesis since the investor knowledge and intentions are publicized on the web. This is evident by a false tweet, when the Associated Press Twitter account was hacked on April 23rd, 2013, which alluded to two explosions at the White House and President Obama being hurt. This caused the market to plunge within minutes [1], [2]. Fortunately however, it reacted after the Associated Press denounced the false tweet. Although the USA market is strong form efficient based on the efficient market hypothesis, the market did not correct itself momentarily after the information went public. This is critical with electronic trading as it could have resulted in market crashes since trades are executed within nanoseconds and several minutes passed before it was corrected. CNN reported that built in circuits to facilitate 'trading halts,' failed to execute. According to Subrahmanyam [13]: "under reasonable conditions, discretionary [or] randomized, trading halts may be less susceptible than rule-based halts to 'gravitational' or 'magnet' effects which occur when traders concerned about an impending closure accelerate their orders (p.1)". The real time impact of the semantic web model can serve as an informant to markets so they can reflect factors with immediate impact.

One of the advantages of sentiment based models is that they consider qualitative factors which are missing from most ML models in the literature. However, the proposed approach in section 3 would consider qualitative factors similar to the prior knowledge artificial neural networks (ANN) used by [14] and the qualitative ANN approach used by [15]. In [15] a functional link ANN architecture was used for both long and short term stock forecasting, which utilized a standard least mean square algorithm with search-then-converge scheduling to compute a learning rate parameter that changes temporally and required less training experiments. Kohara et al [14] focused on the impact of qualitative factors including social and economic change and was less rigorous on the quantitative factors. The model proposed in section 3 includes a mix of qualitative and quantitative factors.

As evidenced by the instantaneous impact of the false tweet, [16] concurs that semantic stock prediction models provide identification of early warnings of financial systemic risk, based on the activity of users of the WWW and the query volume dynamics that emerges from the collective but seemingly uncoordinated activity of many users.

In [17], variants of dynamic social network analysis were used to predict movie stock values on the Hollywood Stock Exchange (HSX). They predicted the daily changes in prices and explored the effectiveness of sentiment analysis and web matrices in predicting trends. No explicit measure of accuracy or value was given for predicting stock price. After examining the nature of messages, [18] found that Web talk does not predict stock movements, although it is a good predictor of volatility.

2.3 Semantic Web Architectures

A semantic web approach to establish a correlation between daily trading volumes of stocks traded and volumes of queries related to the same stocks in the NASDAQ-100 was used in [16]. An OWL based application, Stockwatcher [19], tracked relevant news items on the NASDAQ-100 and predicted one of three possible effects it will have on the company: positive, negative or neutral. The news items were extracted from RSS feeds. The Stockwatcher architecture utilizes natural language processing (NLP) and text mining techniques such as tagging and morphologic analysis.

A generic news based stock prediction system using tagging and classification is outlined in [20]. Feature selection and features weighting are performed on the categorized news and the weighted vectors inputted to the classifier. A survey of eight developed classifier systems were compared. The classifiers used were SVM (or a SVM-variant) and decision trees. The datasets ranged from one month to eight years. The most accurate classifier among that survey was decision trees with 82% on a three month dataset in the system developed by [21]. SVM had a directional accuracy of 70% on a 15 month dataset used by [22].

2.4 Comparison of Machine Learning and Sentiment based Web Approaches

Sentiment and semantic web based prediction models may be more accurate in predicting trends than actual prices. Compared to non-sentiment based approaches, ML approaches have gained superior accuracy in predicting several stock attributes: trend, price and volumes. For example if a company is at the brink of crisis, such as: the loss of a law suit, explosion on the compounds or is approaching bankruptcy; this is likely to be spread rapidly by social media and traders using the semantic model would be able to react

more quickly by selling off the stock than if they relied on a traditional non-sentiment based ML based predictor.

In summary, sentiment and semantic knowledge based approaches may be able to offer faster predictions than traditional ML based approaches as well as give a good indication of volatility in the markets and predict potential trading volumes with reasonable accuracy. Although their learning or processing time is slower than semantic approaches, ML based approaches are better able to handle regression problems and give a good prediction not just on volume traded or market trends, but also on stock prices which is key to determining potential profits.

The semantic web knowledge based approach heavily incorporates human perceptions and inklings and is similar to fuzzy based reasoning on degrees of uncertainty. Unlike other ML approaches that take a non-human like approach to learning focusing mainly on the numbers, the semantic web KB approach favorably considers non-quantitative factors which often have a quick and direct impact on trading.

The traditional ML models are likely to outperform the semantic model if there are no major social or economic changes, since as lengthy supervised learning approaches they place greater emphasis on the relations between quantitative factors over a time series. The semantic web models are likely to be more accurate in short term forecasting whilst ML approaches, especially those that incorporate qualitative factors, may be more accurate on long term forecasts. If ML hybrid approaches were to include the real time inputs from the semantic approach, then the accuracy is expected to improve, as long as it is not given a false positive.

2.5 Summary

Sentiment based models seem better suited for short term forecasts while machine learning approaches will likely outperform them on long term forecasts. As discussed, a variety of architectures emerged within the last eight years on sentiment analysis and semantic web to predict stocks. However, except for the AzFinText system, most have not focused on price prediction. The next section will focus on the experimental and architectural design.

3 Experimental Design

This section details the research design used. A hybrid (qualitative and quantitative) research approach was employed in this study. The architectural design of SentAMaL is outlined in Figure 1. Machine learning algorithms are used both to classify sentiments from social media mining as well as to predict stocks based on qualitative and quantitative inputs, denoted by the shaded objects.

The qualitative data acquisition involves collecting data from the Twitter social network using a customized software developed in the open source R programming language and the Twitter application programming interface (API). Of note, the data on Twitter was fairly readily available through a connection to its API barring the restrictions on its API. A function was created and used to extract tweets from relevant Twitter timelines using hashtags, for example #marijuana, screen names (@tvj and keywords, for example, "weed" and "legalize".

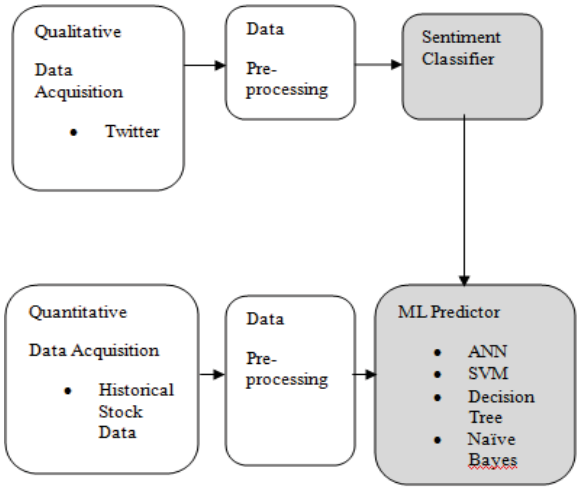

Figure 1: Architectural Diagram of SentAMaL

3.1 Cleaning of Tweets

Figure 2 shows snippets of code used to perform qualitative data pre-processing. The procedures used in this study compared three machine learning classifiers by utilizing a supervised classification approach. The classifier that was best suited for the problem was used to analyze the tweets obtained from Twitter between January and February 2015, in order to determine the sentiments being expressed about marijuana and its legalization.

The qualitative data pre-processing involves removal of duplicate tweets, numbers, punctuation and symbols. A pre-processing function developed in R, allowed for the initial filtering of unwanted or unnecessary verbiage from the tweets, while being extracted from Twitter. It also included replacing certain emoticons with words (e.g. :-) with "happy"). Figure 3 shows samples of tweets cleaned using the R programming application. The tweets downloaded were stored as comma separated values (.csv) files for further processing.

Normalizing Tweets

The normalizing tweets function used in a spreadsheet application, entailed the removal of duplicate tweets. This was done by comparing two rows of tweets at a time to determine whether they were similar. If so, then a user-defined identifier was used to mark one of them and this record was subsequently removed from the dataset.

The quantitative aspect of the analysis was realized using data classification tools that quantified and classified data instances based on the sentiments expressed in each tweet. The analysis was conducted predominantly based on the establishment of sentiment polarity (positive, negative or neutral) of the tweets.

The quantitative data acquisition involves historical data from S&P 500, NASDAQ and JSE over the trading period that the tweets were acquired.

3.2 Population and Sample

The population for the study was comprised of a corpus of approximately 1941 tweets that were extracted from Twitter pages between January and February 2015.

```
clean.text = function(x)
{
  # to lower
  x = tolower(x)
  # remove rt
  x = gsub("rt", "", x)
  # remove at
  x = gsub("@\\w+", "", x)
  # remove punctuation
  x = gsub("[[:punct:]]", "", x)
  #
  x = gsub('[[:cntrl:]]', '', x)
  #
  x = gsub('\\d+', '', x)
  # remove numbers
  x = gsub("[[:digit:]]", "", x)
  # remove links http
  x = gsub("http\\w+", "", x)
  # remove tabs
  x = gsub("[ |\t]{2,}", "", x)
  # remove blank spaces at the beginning
  x = gsub("^ ", "", x)
  # remove blank spaces at the end
  x = gsub(" $", "", x)
```

Figure 2: Tweet Cleaning Function Code Snippet

utech receives machine to advance medical marijuana research
jamaica legalizes medical marijuana amp decriminalizes all weed
legalize jamaica legalizes marijuana jamaica
rt weedfeed jamaica legalized medical marijuana on bob marleys birthday
jamaica legalized medical marijuana on bob marleys birthday
rt whaxyapp jamaica legalized medical marijuana amp decriminalized possession on bob marleys bdayâ€
jamaica legalized medical marijuana amp decriminalized possession on bob marleys bday€
jamaica legalizes medical marijuana amp decriminalizes all weed
medical marijuana could be jamaicaâ€™s economic legacy says businessman joe issa theradioshow itsyourlifestyle
florida and pennsylvania work on new medical marijuana bills and jamaica makes history on bob marleys birthday

Figure 3: Samples of Tweets Cleaned using R Programming Application

This population represented a sample of tweets expressing polarities of possible positive, negative or neutral sentiments about the legalization of medical marijuana. A stratified random sampling method was used to select approximately1000 tweets from the entire corpus for building the classification models. From this, ten fold cross validation was used to classify the dataset.

Of the training dataset, the positive tweets were approximately six times less than those deemed to be negative or neutral. In order to balance the polarity representation, a Synthetic Minority Oversampling Technique (SMOTE) filter was applied to stabilize the unbalanced data instances with synthetic data. Thus, a more equitable classification model was expected to be derived for use in determining the sentiment polarity of new data instances.

3.3 Data Pre-processing

Further pre-processing of the tweets was conducted to remove values not captured by the R tweet cleaner. This involved conversion of the data string into vectors of words by applying a *StringToWordVector* filter. This filter converts string attributes into a set of attributes representing word occurrence (depending on the tokenizer). It also sets parameters that were relevant to the information being retrieved. Such parameters include limits placed on the number of terms repeated (term frequency); the number of words to output (word count); the tokenizer which delimits words within the string; and stemmer that facilitates conversion of terms to their base forms, for example, the base term *love* for the words like lovely, lovable, loving.

3.3.1 Sentiment Classification

In order to determine sentiment of tweets, several machine learning classifiers were evaluated to identify the data mining classification model that is best suited for the problem. Three machine learning classifiers were explored : the Naive Bayes Multinomial Text Classifier, the Support Vector Machine (SVM), and the J48 Decision Tree, which are all said to work well on text categorization.

After training the three classification models to correctly categorize the tweets into positive, negative and neutral classes they were explored to validate their accuracy as well as their efficiency. Naive Bayes Multinomial Text emerged as the best performing model derived from these classifiers and was applied to unknown instances of tweets that were extracted from Twitter.

4 Results

Table I shows the non-sentiment based price prediction of all companies that traded on the JSE within the two month period while Table II shows the price prediction of only the five relevant companies that are perceived would be impacted by sentiments derived from the marijuana tweets. These companies are distributors of medical, pharmaceutical or tobacco products denoted on the JSE by the symbols: MDS, LASM, LASD, JP and CAR.

Table III shows movement prediction of the various indices of the various indices using SentAMaL. The number of instances for the seven indices combined was 2233, while each of the seven indices had 68 instances. While ANN had a slighter higher accuracy for than SVM for all the indices tested, the error was also marginally higher. This denotes the robustness of the SVM classification technique. Decision Trees also proved the superior binary classifier among the three.

Table I
Price Prediction of All Companies on the JSE

	Multilayer Perceptron	SVM
CC	0.9924	0.9954
MAE	3.0385	1.3479
RMSE	5.3307	4.1644
RAE	20.9405 %	9.2894 %
RRSE	12.3317 %	9.6337 %
Total Number of Instances		3547

KEY
CC-Correlation coefficient
MAE- Mean absolute error
RMSE-Root mean squared error
RAE-Relative absolute error
RRSE-Root relative squared error

Table II
Price Prediction of Drug Related Companies on the JSE

	SVM SentAMaL	SVM without SentAMaL
CC	0.9993	.9994
MAE	0.2577	0.2631
RMSE	0.6466	0.6004
RAE	1.7819 %	1.8214
RRSE	3.8413 %	3.563 %
Total Number of Instances		169

5 Conclusion

It is likely that there was not a great difference in values between SentAMaL and its counterpart because the contents of the tweet corpus did not contain shocking content or a newsflash as in the case of the false tweet which temporarily threatened the stability of the US market [2]. Although SentAMaL receives a similar correlation coefficient to its non-sentiment based counterpart, its error is significantly lower than the purely quantitative model. This indicates the accuracy of the SentAMaL model in using sentiment for its qualitative input to complement the traditional quantitative input of most ML stock forecasting models.

6 Acknowledgement

The authors acknowledge the use of the JSE dataset available on their website at http://www.jamstockex.com/ Investor centre => downloads.

Table III
SentAMaL Movement prediction of various JSE indices : Main JSE index(1), JSE Select (2), All Jamaican Composite (3), Cross Listed Index(4), Junior Market Index(5), Combined Index (6) and US Equities Index (7)

Index	**Scheme**	**M**	**Accuracy (%)**
All*	DT	0.09	99.7
All*	ANN	0.076	89
All*	SVM	0.251	87
1	DT	0.089	97
1	ANN	0.167	86.7
1	SVM	0.132	86.7
2	DT	0.014	98.5
2	ANN	0.15	92.6
2	SVM	0.088	91.1
3	DT	0.029	97
3	ANN	0.164	89.7
3	SVM	0.117	88.2
4	DT	0.022	95
4	ANN	0.117	85.2
4	SVM	0.025	84.1
5	DT	0.02	97
5	ANN	0.107	89.7
5	SVM	0.251	88.2
6	DT	0	100
6	ANN	0.192	88.2
6	SVM	0.102	89.7
7	DT	0.01	98.5
7	ANN	0.11	86.7
7	SVM	0.268	82.3

KEY
*All refers to all seven indices
M -Mean absolute error

7 References

[1] Blaine, C.(2013) *Stocks drop on false tweet of White House explosion* Retrieved 4/23/2013 from http://money.msn.com/top-stocks/post.aspx?post=a4b8243d-94ec-432b-ae75-750f850240f

[2] CNN, (2013) *Stocks plunge after twitter account hacked* Retrieved 4/23/2013 from http://www.cnn.com/video/#/video/us/2013/04/23/nr-stocks-plunge-after-ap-twitter-acct-hacked.cnn

[3] Bogle, S.A. and Potter, W.D. (2015) A Machine Learning Predictive Model for the Jamaica Frontier Market *Proceedings of the 2015 Int'l Conference of Data Mining and Knowledge Engineering* by IAENG(ISBN978-888-19253-4-3)

[4] Podobnik B, Horvatic D, Petersen A, Stanley H E (2009) Cross-correlations between volume change and price change. Proceedings National Academy Science USA 106: 22079–22084.

[5] Plerou V, Gopikrishnan P, Rosenow B, Amaral L, Stanley H E (2000) Econophysics: financial time series from a statistical physics point of view. Physica A 279: 443–456.

[6] Yamasaki K, Muchnik L, Havlin S, Bunde A, Stanley H E (2005) Scaling and memory in volatility return intervals in financial markets. Proceedings National Academy Science USA 102: 9424–9428.

[7] Asur, S., & Huberman, B. A. (2010). Predicting the future with social media. *Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference on 1* 492-499.

[8] Wuthrich, B., Cho, V., Leung, S., Permunetilleke, D., Sankaran, K., & Zhang, J. (1998). Daily stock market forecast from textual web data. *Systems, Man, and Cybernetics, 1998. 1998IEEE International Conference on, , 3* 2720-2725.

[9] Mittermayer, M. (2004). Forecasting intraday stock price trends with text mining techniques. *System Sciences, 2004. Proceedings of the 37th Annual Hawaii International Conference on,* 10 pp.

[10] Wang, S., Zhe, Z., Kang, Y., Wang, H., & Chen, X. (2008). An ontology for causal relationships between news and financial instruments. *Expert Systems with Applications, 35*(3), 569-580.

[11] Wang, S., Xu, K., Liu, L., Fang, B., Liao, S., & Wang, H. (2011). An ontology based framework for mining dependence relationships between news and financial instruments. *Expert Systems with Applications, 38*(10), 12044-12050

[12] Schumaker, R. P., Zhang, Y., Huang, C., & Chen, H. (2012). Evaluating sentiment in financial news articles. *Decision Support Systems, 53*(3), 458-464.

[13]Subrahmanyam, A. (1995). On rules versus discretion in procedures to halt trade. *Journal of Economics and Business, 47*(1), 1-16.

[14] Kohara, K., Ishikawa, T., Fukuhara, Y., & Nakamura, Y. (1997). Stock price prediction using prior knowledge and neural networks. *Intelligent Systems in Accounting, Finance and Management, 6*(1), 11-22.

[15] Padhiary, P. K., & Mishra, A. P. (2011). Development of improved artificial neural network model for stock market prediction. *International Journal of Engineering Science, 3*

[16] Bordino, I., Battiston, S., Caldarelli, G., Cristelli, M., Ukkonen, A., & Weber, I. (2012). Web search queries can predict stock market volumes. *PloS One, 7*(7), e40014.

[17] Doshi, L., Krauss, J., Nann, S., & Gloor, P. (2010). Predicting movie prices through dynamic social network analysis. *Procedia-Social and Behavioral Sciences, 2*(4), 6423-6433.

[18] Antweiler, W., M. Frank. 2004. Is all that talk just noise? The information content of Internet stock message boards. J. Finance 59(3) 1259–1295.

[19] Micu, A., Mast, L., Milea, V., Frasincar, F., & Kaymak, U. (2009). Financial news analysis using a semantic web approach. *Semantic Knowledge Management: An Ontology-Based Framework,* , 311-328.

[20] Nikfarjam, A., Emadzadeh, E., & Muthaiyah, S. (2010). Text mining approaches for stock market prediction. *Computer and Automation Engineering (ICCAE), 2010 the 2nd International Conference on, , 4* 256-260.

[21]Rachlin, Last, Alberg, and Kandel, 2007 "ADMIRAL: A Data Mining Based Financial Trading System," *2007 IEEE Symposium on Computational Intelligence and Data Mining*, pp. 720-725.

[22] Zhai, Hsu, and Halgamuge,2007 "Combining News and Technical Indicators in Daily Stock Price Trends Prediction," *Lecture Notes in Computer Science*, pp. 1087-1096.