Using Words from Daily News Headlines to Predict the Movement of Stock Market Indices

Branko Kavšek
University of Primorska, Slovenia
branko.kavsek@upr.si

Stock market analysis is one of the biggest areas of interest for text mining. Many researchers proposed different approaches that use text information for predicting the movement of stock market indices. Many of these approaches focus either on maximising the predictive accuracy of the model or on devising alternative methods for model evaluation. In this paper, we propose a more descriptive approach focusing on the models themselves, trying to identify the individual words in the text that most affect the movement of stock market indices. We use data from two sources (for the past eight years): the daily data for the Dow Jones Industrial Average index ('open' and 'close' values for each trading day) and the headlines of the most voted 25 news on the Reddit WorldNews Channel for the previous 'trading days.' By applying machine learning algorithms on these data and analysing individual words that appear in the final predictive models, we find that the words *gay, propaganda* and *massacre* are typically associated with a daily increase of the stock index, while the word *iran* mostly coincide with its decrease. While this work presents a first step towards qualitative analysis of stock market models, there is still plenty of room for improvements.

Key Words: stock markets, text mining, machine learning, predictive modelling, natural language processing
JEL Classification: C38, C52
https://doi.org/10.26493/1854-6935.15.109-121

Introduction

Predicting the movement of stock market indices is of great importance to entire industries. The investors determine stock prices by using publicly available information to predict how the stock market will react, where 'publicly available information' means mostly (financial) news. Nowadays, news come almost exclusively via web sources in the form of text. This is the reason why many researchers have proposed methods that use text information for analysing the stock market leading to the establishment of an entirely new sub-field of data mining called text mining

(Fawcett and Provost 1999; Permunetilleke and Wong 2002; Thomas and Sycara 2000; Wuthrich et al. 1998).

Recent research in predicting stock market from textual information incorporates knowledge from the fields of economy, statistics, data mining and natural language processing. There are a few main directions that the researchers tend to follow. Shynkevich et al. (2015) focus on improving the predictive power of generated models by carefully choosing the modelling algorithm while simultaneously increasing and diversifying the news sources. Gidófalvi (2001) concentrates on the time series nature of stock prices and uses a Naïve Bayes classifier to find the optimal 'window of influence' where the effect of news to the stock price is greatest. Fung, Yu, and Lu (2005) take this idea even further by introducing complex time series segmentation methods and incorporating advanced data mining and text mining techniques in the system architecture.

Ichinose and Shimada (2016) argue that 'it is unclear whether the improvement of a classifier, such as "raising" or "dropping" of a stock price of each day, contributes to the real trading.' They are doubtful whether small improvements in the classical evaluation metrics, such as prediction accuracy, recall and/or precision rate lead to the improvement of an actual return in trading. They propose a trading simulation system that can estimate the improvement of an actual return in trading and experimentally show its effectiveness. They show that by using their system they can easily understand the effectiveness of one-day classifiers in terms of the real trading situation.

The works of Bollen, Mao, and Zeng (2001) and Chowdhury, Routh, and Chakrabarti (2014) fall into the category of papers that describe the use of sentiment in text to predict the stock market. While Bollen, Mao and Zeng (2001) use sentiment analysis on tweets, Chowdhury, Routh and Chakrabarti, (2014) try to extract sentiment from news. Many other research papers had been written on the subject of 'predicting the stock market from news information' but they can all be categorised in one or more of the above-mentioned categories (predictive power improvement, time series segmentation, trading simulation evaluation and/or sentiment analysis).

Our approach, on the other hand, tries to analyse the models themselves by looking at the words that appear in them. We try to answer the following question: 'Does a word or combination of words (from news articles) exist, such that their presence or absence tells us something about stock price movement?' While our approach is not about sentiment analysis, time series segmentation or trading simulation evaluation, we still

care about the predictive power of our models. That is why we focus on choosing appropriate machine learning algorithms to construct predictive models. Moreover, since we are interested in learning descriptive models that can be analysed in terms of the words that they contain, not all state-of-the-art machine learning algorithms are suitable for the purpose. We still chose to retain both the 'descriptive' and some state-of-the-art 'predictive' algorithms to justify the predictive power of the 'descriptive' algorithms as those are the algorithms that are finally analysed for containing word combinations.

The rest of this paper is structured as follows. The second introduces the problem by describing the data – where was it collected and how was it pre-processed. In the third section, the methodology that was used to analyse the data is presented (all the algorithms that were used to model the data and their applications are listed here). The fourth section presents the results and provides a short discussion. Finally, the fifth concludes by summarising the most important findings and giving possible directions for further work.

Data

We used data from two independent sources:

- News data: historical news headlines from Reddit WorldNews Channel (see https://www.reddit.com). They are ranked by Reddit users' votes, and only the top 25 headlines are considered for a single date.
- Stock data: Dow Jones Industrial Average (DJIA) daily index values were used (see https://finance.yahoo.com/quote/%5EDJI/history?=%5EDJI). On each date, the 'open,' 'high,' 'low,' 'close' and 'volume' values are recorded.

Data for the past eight years was collected – from 8 August 2008 to 1 July 2016. Figure 1 shows how the 'open' value of the DJIA index was changing over this period.

News and stock data were merged into a single dataset by aligning the news headlines with the trading days of the stock data – for each of the 1989 trading days all the DJIA index values for that day together with the most voted 25 news headlines for the previous day were recorded (in a previous version the news data were aligned to the stock data on the same day, but experiments showed that predicting stock value form 'yesterday's news' gives higher predictive accuracy).

Since we were only interested in predicting if the stock goes 'up' or 'down' on a particular day, only the index values of 'open' and 'close' were

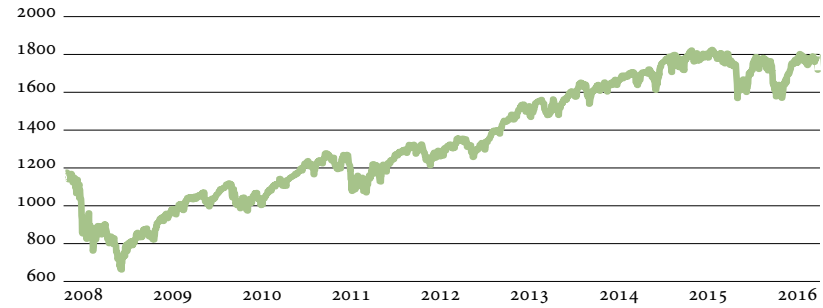

FIGURE 1 DJIA Index ('Open' Value) – Past Eight Years (8 August 2008 to 1 July 2016)

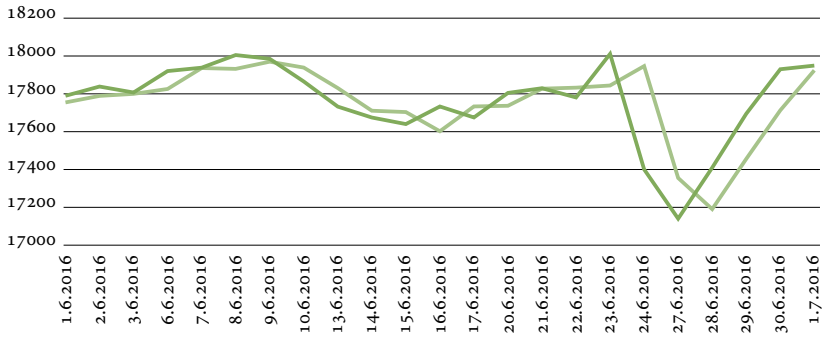

FIGURE 2 DJIA Index – Last Month (light – 'Open,' dark – 'Close' values; 1 June 2016 to 1 July 2016)

relevant. Figure 2 shows how the 'open' and 'close' values changed over the last month (of the collected data). For the sake of simplicity all the stock data was replaced by a 'label' feature in the dataset with value '0' if the DJIA index went down or stayed the same for that day (the 'close' value is the same or smaller than the 'open' value), and '1' if it went up (the 'close' value is bigger than the 'open' value). Figure 3 depicts a small subset of this transformed dataset – the first 10 trading days and just five of the 25 news headlines. The entire pre-processed dataset, shown partly on figure 3, consists of 1989 rows, representing the trading days, and 27 columns that represent the features of which 25 are the news headlines (as text), and 1 is the date of the trading day. Finally, the dependent feature 'label' reflects the rising or falling of the DJIA index.

Methodology

Our goal is to build a prediction model that will use the textual information from 'today's' Reddit top 25 news headlines to predict 'to-

	A	B	C	D	E	F	G
1	Date	Label	Top1	Top2	Top3	Top4	Top5
2	8.08.2008	0	Georgia downs two Russian warplanes as countries move to brink	BREAKING: Musharraf to be impeached.	Russia Today: Columns of troops roll into South Ossetia footage from	Russian tanks are moving towards the capital of South Ossetia, which	Afghan children raped with impunity, U.N. official says - this is sick, a three
3	11.08.2008	1	Why wont America and Nato help us? If they wont help us now, why	Bush puts foot down on Georgian conflict	Jewish Georgian minister: Thanks to Israeli training, were fending off	Georgian army flees in disarray as Russians advance - Gori abandoned	Olympic opening ceremony fireworks faked
4	12.08.2008	0	Remember that adorable 9-year-old who sang at the opening	Russia ends Georgia operation	If we had no sexual harassment we would have no children...	Al-Qaeda is losing support in Iraq because of a brutal crackdown on	Ceasefire in Georgia: Putin Outmaneuvers the West
5	13.08.2008	0	U.S. refuses Israel weapons to attack Iran: report	When the president ordered to attack Tskhinvali [the capital of South	Israel clears troops who killed Reuters cameraman	Britain\s policy of being tough on drugs is pointless, says a former civil	Body of 14 year old found in trunk Latest (ransom paid) kidnapping
6	14.08.2008	1	All the experts admit that we should legalise drugs	War in South Osetia - 89 pictures made by a Russian soldier.	Swedish wrestler Ara Abrahamian throws away medal in Olympic hissy	Russia exaggerated the death toll in South Ossetia. Now only 44 were	Missile That Killed 9 Inside Pakistan May Have Been Launched by the CIA
7	15.08.2008	1	Mom of missing gay man: Too bad hes not a 21-year-old cheerleader,	Russia: U.S. Poland Missile Deal Wont Go Unpunished	The government has been accused of creating laws that have a chilling	The Italian government has lashed out at an influential Catholic	Gorbachev: Georgia started conflict in S. Ossetia
8	18.08.2008	0	In an Afghan prison, the majority of female prisoners are serving 20-year	Little girl, youre not ugly they are	Pakistans Musharraf to Resign, Leave the Country	Tornado throws a bus in Poland, captured by one of the passengers	Britains terror laws have left me and my family shattered
9	19.08.2008	0	Man arrested and locked up for five hours after taking photo of police van	The US missile defence system is the magic pudding that will never run out	Schrder lambasted for blaming Russian conflict on Georgia	Officials: 10 French soldiers killed near Kabul	These ten laws make China a totalitarian wasteland
10	20.08.2008	1	Two elderly Chinese women have been sentenced to a year of re-	The Power of Islam: The Human Rights Council at the United Nations	We had 55 times more military soldiers in the first Gulf War than	I live here on less than a dollar a month – Obama\s brother, Kenya	Russia sends aircraft carrier to Syria.
11	21.08.2008	1	British resident held in Guantanamo Bay wins legal battle to force Foreign	Chinese may have killed 140 Tibetans this week: Dalai Lama	U.S. Navy Ships Head to Georgia	Hacker uncovers Chinese olympic fraud	If youve ever wondered what Kim Jong Il was like in grade school, here

FIGURE 3 Excerpt from the Pre-Processed Dataset Used for Further Analysis

morrow's' rise or fall of the DJIA index and to interpret this model in terms of (sets of) words that most affect the DJIA index change. We divided the work in four phases: Data transformation, Modelling technique selection, Quantitative evaluation and Qualitative evaluation. For the first three phases, we used the WEKA data mining workbench – an open-source collection of machine learning and data mining algorithms developed on the University of Waikato in New Zealand (see https://www.cs.waikato.ac.nz/ml/weka/).

DATA TRANSFORMATION PHASE

The pre-processed data in the tabular format (see table 3) is not yet suitable for further analysis, since we cannot directly link 'free text' to the categorical/binary variable 'label' (with values '0' and '1' – representing the fall and rise of the DJIA index, respectively). In order to extract valuable information from 'free text,' some natural language processing technique must be applied ('Natural Language Processing' 2017). The simplest (and often effective) such technique is the so-called Bag-of-Words model ('Bag-of-Words Model' 2017). It takes the text as input and produces a vector in which every element represents the number of appearances of some word in the text. In WEKA, the Bag-of-Words model is implemented as the StringToWordVector filter that has some additional capabilities such as converting all the words in lowercase letters and eliminating the 'stop-words' (words that typically do not carry any information and are mostly used as connecting words in a sentence). Moreover, the StringToWordVector filter does stemming (but no lemmatisation) on the text, transforming all the words in their stems (e.g. the word 'friendly' becomes 'friend'). The Lovins Stemmer (Lovins 1968) is used for this purpose. By applying this filter to our data, we transformed the 'free text'

portion of our data (the 25 text features) into 10,000 numerical features, representing the counts of the most frequent 10,000 words in our news headlines. We, furthermore, removed the 'date' feature, since our goal is not to explore the time series nature of the data, but rather to model the dependency of the DJIA index from information contained in the news headlines. Finally, we split the entire dataset in an 80% – 20% fashion, using 80% of the data to train the models (training set), and 20% to test the models in the quantitative evaluation phase (test set). Since the entire dataset is ordered in time, the 80% – 20% split must be done 'in sequence,' taking the first 80% of the examples (from 8 August 2008 to 31 December 2014; 1611 examples) for training and the last 20% of the examples (from 2 January 2015 to 1 July 2016; 378 examples) for testing. The 'sequential splitting' is important because we want to simulate the process of predicting 'new' (unknown) events from 'old' (already known) events.

MODELLING TECHNIQUE SELECTION PHASE

In this phase, we selected several machine learning algorithms implemented in WEKA to train the prediction models and compare their performance (in the quantitative evaluation phase). Particularly, we wanted to include algorithms that produce descriptive models that we can further examine in the qualitative evaluation phase. That is why we chose to include a decision tree learning algorithm – C4.5 (Quinlan 1993), and a decision rule learning algorithm – PART (Frank and Witten 1998). Nevertheless, we wanted to make sure that our chosen descriptive learning algorithms perform well, so we also included state-of-the-art machine learning algorithms to have a fair comparison. We chose the following algorithms for this purpose: Naïve Bayes (John and Langley 1995), Support Vector Machines – SVM (Hastie and Tibshirani 1998; Keerthi et al. 2001; Platt 1998), *k*-nearest neighbours – *k*NN (Aha and Kibler 1991) and random forests – RF (Breiman 2001). Finally, we included two more simple models that will serve as baseline for the comparisons – the majority class classifier that just outputs the majority value of the dependent variable, and the 'one-rule' classifier or OneR (Holte 1993) that finds the independent variable that is 'most correlated' with the 'label' variable.

QUANTITATIVE EVALUATION PHASE

In this phase, we run all the selected algorithms from the previous phase on the 80% training data. Then we test the generated models on both the same training data and on the separate 20% test set. By testing the

models on the same data, they were generated from, we verify the bias of our models. On the other hand, using a separate test set verifies the variance of the models. The metric we used to test our models is classification accuracy i.e. the percentage of correctly classified examples by the model, where 100% accuracy means a perfect model that makes no error and 50% accuracy means the model is randomly guessing the 'label.' Arguably, classification accuracy is not the most appropriate metric to measure the quality of prediction models for stock markets (Ichinose and Shimada 2016). Since prediction power of the models is not the main focus of this research, we assume that classification accuracy is a good enough measure to verify that the predictive power of selected descriptive models is comparable to the other state-of-the-art selected models. Thus, the descriptive models can be further analysed in the qualitative evaluation phase.

QUALITATIVE EVALUATION PHASE

After verifying that the prediction power of the descriptive models is high enough, this phase is used to 'look into' the models to identify the words (from news headlines) that alone or in combination with other words affect the rising or falling of the DJIA daily index. Both descriptive algorithms – PART and C4.5 – produce a model in the form of decision rules (the decision tree produced by the C4.5 algorithm can be decomposed into a set of decision rules). The decision rules are conditional clauses of the form:

$$\text{IF} < (word_1 \leq value) \,\&\, (word_2 > value) \,\&\, \ldots \\ \,\&\, (word_N \leq value) > \text{THEN} < label = 0, 1 >,$$

where the left-hand side contains the conjunction of word-values, and the right-hand side represents the decision (either label = '0,' meaning the DJIA index went down, or label = '1,' meaning the DJIA index went up – the values '0' and '1' were chosen completely arbitrarily). For example, the decision rule:

$$\text{IF } (\text{'}commander\text{'} \leq 0) \,\&\, (\text{'}mount\text{'} > 0) \text{ THEN } label = 0,$$

means that the absence of the word 'commander' and the presence of the word 'mount' (one or more times in the news headlines) is a strong indicator that the DJIA index will go down.

In this phase, we aim at identifying such strong indicative words by

TABLE 1 Quantitative Results of Prediction Model Evaluation on Training and Test Data

Algorithm	ACC-train (%)	ACC-test (%)	Learning time(s)	Description
Majority	54.19	50.79	0.95	majority = '1'
OneR	55.80	51.32	1.12	'run'
Naíve Bayes	76.35	60.79	1.11	
SVM	93.85	77.50	19.20	
*k*NN	72.01	61.41	0.66	
PART	97.70	71.21	19.27	65 rules
C4.5	96.52	70.73	12.69	209 rules
RF	100.00	79.11	27.43	100 trees

manually inspecting the decision rules generated by the description models.

Results

The results obtained in the quantitative evaluation phase are presented in table 1 (all algorithms were 'run' with default parameters). The column 'Algorithm' in this table contains the short names of the machine learning algorithms used to construct the predictive models (full names with references are presented in the third section). Columns 'ACC-train (%)' and 'ACC-test (%)' present the classification accuracies on the train and test sets (in percentage), respectively. The column 'Learning time(s)' shows the time (in seconds) taken to learn the models. The column 'Description' gives additional description where available.

It can be observed from table 1 that the learning data were evenly distributed between classes – in the training set 54.19% of the data had label '1' (the majority class); in the test set this percentage falls to 50.79.

The OneR ('one-rule') model predicted that the word 'run' has the 'highest correlation' with the label, but this model is practically useless, since its predictive accuracy doesn't really outperform the majority classifier.

Of the other six models, four had over 90% classification accuracy on the training set with Random Forests reaching the perfect 100% accuracy. This tells us that these models have very low or, in the case of Random Forests, even zero bias. On the other hand, the drop in predictive accuracy on the test set shows all our models are over-fitted to the training data, showing some amount of variance. Such behaviour of the classifi-

cation models is perfectly normal and typical in data mining. Since the task is to use the models to predict future data, the test set classification accuracies are the ones we should be looking at. Related work on predicting stock value from textual information shows us that classification accuracies of 70% and more on the test set are to be expected from 'good' prediction models (Paliyawan 2015). The algorithms SVM (Support Vector Machines), PART (decision rule learner), C4.5 (decision tree learner) and RF (Random Forests) all reached more than 70% accuracy, while Naïve Bayes and *k*NN (nearest neighbour) fell below this threshold.

The 'learning times' of the algorithms are proportional to their complexity and classification accuracy – the more complex the algorithm, higher is its accuracy and longer it takes to learn the model from data. SVN, PART, C4.5 and RT are all complex algorithms compared to Naïve Bayes, MNN and OneR.

The results in table 1 show us that our two 'descriptive' algorithms, namely, PART and C4.5 achieved a 'good' predictive accuracy compared to other state-of-the-art algorithms (SVN and RT) and to results from other researchers (Paliyawan 2015).

We now proceed to the qualitative analysis of the models learned by the algorithms PART and C4.5. PART is a decision rule learning algorithms that on our data learned a model consisting of 65 decision rules. C4.5 is a decision tree learner and (on our data) learned a decision tree with 209 leaves – that can be decomposed in 209 decision rules. To simplify the manual inspection of the generated rules we further 'pruned' the rule sets produced by both models by using a technique called post-pruning (Frank and Witten 1998; Quinlan 1993). The simplified (pruned) version of the PART rule set contained 28 rules, the C4.5 pruned tree had 15 leaves.

The use of pruning is common practice in machine learning and prevents overfitting of models to the training data and thus reduces the bias of the models.

By manually analysing all the decision rules produced by the pruned versions of PART and C4.5 we found out the following:

- The eight words that most appeared in the simplest rules generated by the two descriptive models are: *iran, gay, propaganda, map, low, massacre, web* and *reports;*
- The word *iran* alone when appearing more than three times (in the news headlines) 'tends to negatively affect' the DJIA index;

- The combination of words *propaganda* and *map* – when *map* appears, but *propaganda* doesn't appear in the text 'tends to negatively affect' the DJIA index;
- The combination of words *propaganda* and *low* – when *low* appears, but *propaganda* doesn't appear in the text 'tends to negatively affect' the DJIA index;
- The word *gay* alone when appearing more than one time 'tends to positively affect' the DJIA index;
- The word *propaganda* alone when appearing in the text 'tends to positively affect' the DJIA index;
- The word *massacre* alone when appearing in the text 'tends to positively affect' the DJIA index;
- The combination of words *web* and *reports* – when neither appears in the text 'tends to positively affect' the DJIA index;

The above findings list just the words that appeared in simple decision rules where the left-hand side consisted of only one or two conjuncts. There are many more words that appear in more complicated rules with more than two conjuncts on the left-hand side – all other words except *iran* in figure 4 are of this kind.

It should be stressed here that the phrase 'tend to negatively/positively affect' used in the above eight findings does not necessarily mean causality. It just reflects the (non)co-occurrence of specific words with either fall or rise of the DJIA index. For example, the news headline 'Iran tells Hezbollah to stop attacking Israel, turn attention to Saudi Arabia' is probably related to hostilities and indirectly to oil and those allegations could be a possible source for the fall of the DJIA index for that day.

Conclusions and Further Work

All research on the topic of 'predicting the stock market from news information' focus on either predictive power improvement, time series segmentation, trading simulation evaluation, sentiment analysis or a combination of those. Our approach, on the other hand, tries to exploit the descriptive power of the predictive models by analysing the (combination of) words that they contain and associate these to the movement of the stock price. By binarizing the stock price movement, we used the machine learning approach to learn models that are able to predict the stock rise or fall.

We have chosen six most popular machine learning algorithms (along with two base-line methods) and experimentally showed that four of

those are appropriate for prediction. We further narrowed our choice to the two descriptive algorithms – PART and C4.5 – that we analysed for 'interesting' words that 'directly' affect the stock price. By simplifying and manually inspecting the set of decision rules generated by both descriptive algorithms we found that the words *iran, gay, propaganda* and *massacre* have the biggest influence on the rising and falling of our DJIA stock market index. While the word *iran* is typically associated with DJIA falling, the other three words (*gay, propaganda* and *massacre*) are typically associated with DJIA rising. There are also other combinations of inclusion and exclusion of other words, but none as simple as the four mentioned words.

The main contribution of this work is showing that by using appropriate machine learning algorithms one can accurately predict individual words which almost always (or practically never) co-occur with the movement of a selected stock market index (DJIA in our case).

We regard this as a preliminary study of descriptive analysis of predictive models for stock markets using textual data. A first step has been done in this direction by showing that there is some relation between the words in the headlines of the daily news and the movement of the stock market price. This research concentrates on classification methods for building prediction models, which can predict only the sign of the stock movement (rise or fall). More formal ways should be studied to prove this relation and quantify its extent. A natural way to extend our approach would be to use regression instead of classification to quantify the change in stock price. In our research, we used manual inspection to identify the 'potentially interesting' (combinations of) words in decision rules. Methods for automatic extraction of such word combinations could be devised along with some corresponding evaluation measure. There is plenty of room for improvement in the natural language processing part of our study – other, more elaborate methods could be used instead of the simple Bag-of-Words for extracting word information from text. Ventura and Ferreira da Silva (2008) give an overview of the state-of-the-art methods that are in use nowadays.

Since this work presents the problem at hand (e.g. the prediction of the movement of stock market indices) mostly from a data mining perspective, there is significant room for improvement on the 'interpretation' side. As further work, we plan to show the results to a stock market expert to identify the true causes for the movement of stock market indices.

In this paper, the presented methodology is used to analyse the news headlines in English language. A natural extension of this work would be

to try to apply the methodology to other languages as well. It is however unclear if a direct approach would work better or would it be feasible to first translate the text in English and then apply our proposed methodology. There are some automatic translation systems that show promising results for similar languages (Vičič, Homola, and Kuboň 2016).

References

Aha, D., and D. Kibler. 1991. 'Instance-Based Learning Algorithms.' *Machine Learning* 6:37–66.

'Bag-of-Words Model.' 2017. *Wikipedia.* https://en.wikipedia.org/wiki/Bag-of-words_model

Bollen, J., H. Mao, and X. J. Zeng. 2001. 'Twitter Mood Predicts the Stock Market.' *Journal of Computational Science* 2 (1): 1–8.

Breiman, L. 2001. 'Random Forests.' *Machine Learning* 45(1): 5–32.

Chowdhury, S. G., S. Routh, and S. Chakrabarti. 2014. 'News Analytics and Sentiment Analysis to Predict Stock Price Trends.' *International Journal of Computer Science and Information Technologies* 5 (3): 3595–604.

Fawcett, T., and F. J. Provost. 1999. 'Activity monitoring: Noticing Interesting Changes in Behaviour.' In *Proceedings of the 5th International Conference on Knowledge Discovery and Data Mining,* 53–62. New York: Association for Computing Machinery.

Frank, E., and I. H. Witten. 1998. 'Generating Accurate Rule Sets Without Global Optimization.' In *Fifteenth International Conference on Machine Learning,* 144–51. San Francisco, CA: Morgan Kaufmann.

Fung, G. P. C., J. X. Yu, and H. Lu. 2005. 'The Predicting Power of Textual Information on Financial Markets.' *IEEE Intelligent Informatics Bulletin* 5(1), 1–10.

Gidófalvi, G. 2001. 'Using News Articles to Predict Stock Price Movements.' http://cseweb.ucsd.edu/~elkan/254spring01/gidofalvirep.pdf

Hastie, T., and R. Tibshirani. 1998. 'Classification by Pairwise Coupling.' *The Annals of Statistics* 26 (2): 451–71.

Holte, R. C., 1993. 'Very simple classification rules perform well on most commonly used datasets.' Machine Learning. 11:63–91.

Ichinose, K., and K. Shimada. 2016 'Stock Market Prediction from News on the Web and a New Evaluation Approach in Trading.' In *5th IIAI International Congress on Advanced Applied Informatics (IIAI-AAI)* 77–81. New York: IEEE Electronic Publishing.

John, G. H., and P. Langley. 1995. 'Estimating Continuous Distributions in Bayesian Classifiers.' In *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence,* 338–45. San Mateo, CA: Morgan Kaufmann.

Keerthi, S. S., S. K. Shevade, C. Bhattacharyya, and K. R. K. Murthy. 2001. 'Improvements to Platt's SMO Algorithm for SVM Classifier Design.' *Neural Computation* 13 (3): 637–49.

Lovins, J. B. 1968. 'Development of a Stemming Algorithm.' *Mechanical Translation and Computational Linguistics* 11:22–31.

'Natural Language Processing.' 2017. *Wikipedia.* https://en.wikipedia.org/wiki/Natural_language_processing

Paliyawan, P. 2015. 'Stock Market Direction Prediction Using Data Mining Classification.' *ARPN Journal of Engineering and Applied Sciences* 10 (3): 1302–10.

Permunetilleke, D., and R. K. Wong. 2002. 'Currency Exchange Rate Forecasting from News Headlines.' In *Proceedings of the 13h Australian Database Conference,* 131–9. Darlinghurst: Australian Computer Society.

Platt, J. 1998. 'Fast Training of Support Vector Machines using Sequential Minimal Optimization.' In *Advances in Kernel Methods: Support Vector Learning,* edited by B. Schoelkopf, C. Burges, and A. Smola, 41–65. Cambridge, MA: The MIT Press.

Quinlan, R. 1993. *C4.5: Programs for Machine Learning.* San Francisco, CA: Morgan Kaufmann.

Shynkevich, Y., T. M. McGinnity, S. Coleman, and A. Belatreche. 2015. 'Predicting Stock Price Movements Based on Different Categories of News Articles.' In *IEEE Symposium Series on Computational Intelligence,* 703–10. Los Alamitos, CA: IEEE Computer Society.

Thomas, J. D., and K. Sycara. 2000. 'Integrating Genetic Algorithms and Text Learning for Financial Prediction.' In *Proceedings of the Genetic and Evolutionary Computing 2000 Conference Workshop on Data Mining with Evolutionary Algorithms,* 72–5. New York: ACM Press.

Ventura, J., and J. Ferreira da Silva. 2008. 'Ranking and Extraction of Relevant Single Words in Text.' In *Brain, Vision and AI,* edited by C. Rossi, 265–84. Rijeka: InTech.

Vičič, J., P. Homola, and V. Kuboň. 2016. 'Automated Implementation Process of Machine Translation System for Related Languages.' *Computing and Informatics* 35 (2): 441–69.

Wuthrich, B., D. Permunetilleke, S. Leung, V. Cho, J. Zhang, and W. Lam. 1998. 'Daily Prediction of Major Stock Indices from Textual www Data.' In *Proceedings of the 4th International Conference on Knowledge Discovery and Data Mining,* 364–8. New York: IEEE Electronic Publishing.