IIMB Management Review (2019) 31, 157–166

available at www.sciencedirect.com

ScienceDirect

journal homepage: www.elsevier.com/locate/iimb

News-based supervised sentiment analysis for prediction of futures buying behaviour

Ritu Yadav[a,1], A. Vinay Kumar[b,*], Ashwani Kumar[c,2]

[a] *Management Information Systems Area, Indian Institute of Management Rohtak, Rohtak, Haryana, India*
[b] *Finance & Accounting Area, Indian Institute of Management Lucknow, Lucknow, Uttar Pradesh, India*
[c] *IT and Systems Area, Indian Institute of Management Lucknow, Lucknow, Uttar Pradesh, India*

Received 1 August 2016; revised form 5 April 2017; accepted 27 March 2019; Available online 2 April 2019

KEYWORDS
Sentiment analysis;
Real-time news;
Trade direction;
Vector space model;
Support vector machines;
Prediction;
Futures market;
Net buying pressure

Abstract This study examines the predictability of real-time news data on investors' buying behaviour in the futures market, using supervised sentiment analysis. Market sentiment or traders' buying behaviour is captured at the bid-ask stage of price formation using the net buying pressure (NBP). Any significant change in NBP patterns defines an "interesting market event". Real-time news headlines are automatically labelled using interesting market events, assuming a lag between the market information and its impact on buying behaviour. News was found to have an impact on the market buying behaviour of the S&P NIFTY index futures with an optimal lag of 5 minutes. Manual labelling of the news data validated this empirical finding.

Sentiment analysis can be used to determine the impact of unstructured market news on the emotions of investors, which is referred to as market sentiment. Prior studies have established the predictability of the impact of news on market sentiment in the spot market context. This study aims to capture market sentiment at the earliest price formation stage, i.e., when investors reveal their bid-ask quotes. Bid-ask quotes are outcomes based on the information available to the market participants. Investors who have superior information would reveal the same through their intentions in the form of bid-ask quotes. In this study, we perform supervised sentiment analysis using high-frequency, real-time news data and the behaviour of index futures traders at the level of bid-ask quotes.

Understanding the behaviour of futures market traders is pertinent as the futures market leads the spot markets in reacting to market events (Kumar and Jaiswal, 2013; Vipul, 2009). Net buying pressure (NBP), which is the difference between the number of buyer-initiated trades and the number of seller-initiated trades calibrated from the bid-ask quotes, proxies traders' sentiment as it reveals the direction of the trades. The objective of this study is to predict the trends in NBP in the futures market by using supervised sentiment analysis of real-time news headlines. Understanding the impact of news on market trade directions as well as the time lag between the arrival of news and the anticipated trade direction is critical to this objective.

There are two main approaches to sentiment analysis—the unsupervised dictionary-based approach and the supervised approach. In the unsupervised dictionary-based

[*] Corresponding Author. Phone: 91-522-6696645; Fax: 91-522-2734025.
E-mail addresses: ritu.yadav@iimrohtak.ac.in (R. Yadav), vinay@iiml.ac.in (A.V. Kumar), ashwani@iiml.ac.in (A. Kumar).
[1] Phone: 91-1262-228548; Fax: 91-1262-274051.
[2] Phone: 91-522-6696660; Fax: 91-522-2734025.

https://doi.org/10.1016/j.iimb.2019.03.006

approach, market sentiment is extracted from the news text directly. A dictionary of sentiment words is used to count the different sentiment-related or emotions-related textual cues (Antweiler & Frank, 2004; Das & Chen, 2007; Garcia, 2013; Tetlock, 2007). In supervised sentiment analysis, market sentiment is learnt by using historical market trends and news patterns (Wuthrich et al., 1998; Lavrenko et al., 2000; Fung et al., 2003). The efficacy of dictionary-based sentiment analysis depends on the accuracy of the sentiment dictionary that is used and the suitability of the dictionary in a specific context.

In the supervised sentiment analysis approach, training data is created by either manual labelling or automatic labelling of historical news. Although manual labelling has higher precision, this method cannot be adopted for larger data sets. Automatic labelling involves temporally aligning "interesting" or significant market trends with the news that might have caused those trends. In this study, we use the supervised sentiment analysis approach to examine high-frequency news headlines using a vector space model (VSM). We also compare the predictability of dictionary-based sentiment metrics and that of a VSM.

We assume that markets are inefficient and take time to discount market information. Hence, we consider a trend to be related to the news stories that are released within an assumed *alignment window*. The time taken to discount market information varies according to the type of market or the asset under consideration. Prior studies on spot market sentiment analysis found empirically that new market information takes around 20 minutes to reflect in the stock prices (Gidofalvi, 2001). However, the futures market leads the spot market, and bid-ask quotes reflect the investors' sentiment faster than price trends. Hence, in the context of the index futures' sentiment at the bid-ask level, the arrival of news information has a much faster impact on the quotes. This study tests this hypothesis empirically by using different values for the alignment window. The results are validated by manually labelling a small sample of the news data. The context of this study is the National Stock Exchange (NSE) futures market, with index futures as the target asset and their trade direction as the measure of investor sentiment.

The rest of this paper is organised as follows. The subsequent section presents a literature review of sentiment analysis studies conducted in a financial market setting. The third section presents the supervised sentiment analysis methodology used in this study, followed by the data analysis and the results in the fourth section. To the best of our knowledge, this is one of the first sentiment analysis studies conducted in the futures market in the Indian context. The insights gained from this study could prove useful in high-frequency algorithmic trading and in understanding buying behaviour in the futures market in general. The implications of this research for academics and financial investments are discussed in the fifth section.

Literature review

Prior studies have established the predictability of the impact of news on market sentiment in the spot market context. Supervised sentiment analysis is a text classification task where the impact of textual market information on financial markets is learnt by using labelled training data. The creation of the training data involves labelling the news instances according to their impact on the markets. The news instances can be labelled manually according to the news discourse (Davis et al., 2006; Bozic & Seese, 2011) or automatically based on the corresponding market trends (Mittermayer & Knolmayer, 2006). Labelling news instances manually is a precise but labour-intensive task; hence, this method is not suitable for high-frequency news analytics. Automatically aligning news instances with the corresponding market trends makes for one of the most challenging tasks in sentiment analysis. Yoo et al. (2005), Mittermayer & Knolmayer (2006), Nikfarjam et al. (2010), and Nassirtoussi et al. (2014) reviewed studies that have used the supervised sentiment analysis approach. A brief review of some of the major concerns that our study deals with, follows.

News-trend alignment is one of the most important parts of a supervised sentiment analysis model for testing and contextualising market efficiency. The accuracy of the news-trend alignment procedure constrains the efficacy of the supervised sentiment analysis. It results in noisy news labelling, and consequently, erroneous training data. Some of the reasons for a noisy training data set are:

1. The news stories may have been published out of time with the market trends.
2. A news story may have co-occurred with a trend by chance.
3. A news story may have contained contrary information to a market trend—for example, a positive story might have co-occurred with a negative trend.
4. A market trend may have been illusory because the market could move without the news information (Drury et al., 2012; Chan, 2003).

An efficient market requires news to be aligned with the price trends since all the information that arrives into the market is instantaneously discounted during price discovery. However, studies on sentiment analysis (Chan, 2003; Leinweber & Sisk, 2011) and behavioural finance refute the efficient market hypothesis (Baker & Wurgler, 2007). Markets frequently under-react to the post-news drifts, and often feature momentum trading (Chan, 2003; Baker & Wurgler, 2007).

Momentum trading has been associated with stale news in the extant literature (Tetlock, 2011). Long-term momentums can be filtered out to understand the impact of news on long-term market trends (Uhl et al., 2015). However, for high-frequency sentiment analysis, news headlines have been a preferred choice for analysis (Kohara et al., 1997; Bunningen, 2004; Takahashi et al., 2007; Chan, 2003). Headlines are one of the earliest forms in which information arrives into the markets (Takahashi et al., 2007). Analysing news-trend patterns using headlines alone can significantly improve classification performance compared to results using the whole news body or sentence-level analysis (Bunningen, 2004). Therefore, in this study, we conduct supervised sentiment analysis on real-time news headlines.

The time lag in news-trend alignment signifies the impact of news on the market, i.e., once news has arrived in the market, how much time does it take for the prices to reflect or discount the information? Prior studies refer to this lag period as the *window of influence* (Gidofalvi, 2001) or the *reaction time* of news (Cheng, 2010). In this paper, we refer to it as the *alignment window* as this time lag helps in aligning market trends with the probable impact of news.

For news stories that get published out of time with the corresponding trend, the time lag or a brief window of influence helps in aligning such news stories with their respective market trends. However, if the length of the window of influence is not chosen appropriately, it may introduce the risk of confounding news. Confounding news instances are instances that occur in an opportunity window of either an unrelated market trend or contradicting market trends. Such instances introduce noise while labelling the training data.

One method of handling confounding news instances is to ignore such instances altogether (Gidofalvi, 2001). Another way is to use an optimal alignment lag in aligning the news with the market trends that alleviates the confounding effect to a great extent. In the case of inefficient markets, the news needs to be aligned with the observed market trend with a certain time lag (Lavrenko et al., 2000; Gidofalvi, 2001). Lavrenko et al. (2000), one of the first studies of high-frequency news sentiment analysis, tested four different lags, namely, 0, 1, 5, and 10 hours to calibrate the impact of news on stock market trends. The predictability of the news features was found to decrease with an increase in lag, delivering significantly better prediction accuracy. Gidofalvi (2001) empirically tested different lengths of the window of influence and found that the impact of news is significantly high in [-20 minutes, 0] and [0, 20 minutes].

In tertiary duration studies, Li et al. (2010) found the optimal lag to be 15 days. While tertiary lags reflect the overall impact of news on long-term market trends, high-frequency time lags indicate the alignment lag over which the impact of the news can be profitably arbitraged (Lavrenko et al., 2000; Gidofalvi, 2001; Robertson et al., 2007). Gidofalvi's 20-minute window of influence has been the defacto standard in prior sentiment analysis studies conducted on stock markets (Gidofalvi, 2001; Schumaker & Chen, 2008). The larger the assumed opportunity window, the more profound is the presence of the confounding news. Although a large window of influence helps in accounting for more news records, resulting in a larger training data set, the data set is obtained at the cost of inducing noise in the training data. Finding an optimal window of influence is a challenging task.

Trends in the futures markets often lead the trends in the stock markets (Vipul, 2009). These trends can be best captured at the price formation stage with bid-ask quotes. This paper presents a sentiment analysis study of the futures market using real-time news headlines to represent market information and net buying pressure to represent market trading behaviour. The study has two objectives: to test the predictability of news on the sentiment in the futures market, and to find the optimal time lag in which the futures market reacts to news.

Methodology

Supervised sentiment analysis is essentially a text classification task that can be divided into the following steps:

1. *Market trend identification*, which defines and identifies the significant market patterns of a given asset.
2. *News-trend alignment* scheme, which is used to create the training data set
3. *News representation* scheme, which converts the unstructured news data into structured data
4. *Classification and evaluation*

Market trend identification

This study defined significant NBP instances as market trends and the corresponding change points as market events. For a given asset, a market nugget m_i at time t_i contains the price and trade direction event information.

$$m_i = \{p_i, e_i\} | t_{i-1} - t_i = \delta$$

where p_i is the price traded, and e_i is the trade direction trend {+b, -s, NaN} at time t_i. Here, +1 refers to the number of buyer-initiated trades in the given time interval δ, and -1 refers to the number of seller-initiated trades.

To calculate e_i, the trade direction is calculated using real-time bid-ask quotes in an event time scale with one-second precision data following Lee and Ready's (1991) algorithm. For each minute, the total number of buyer-initiated trades and the number of seller-initiated trades are calculated to get the net difference, i.e., the net buying pressure. In this one-minute NBP time series, an event is defined as significant or interesting if the NBP value and its first moment are significantly above average. In this study, an event is defined using NBP and NBP log returns time series in the following way. At time t_i, let the net buying pressure be *nbp*$_i$, and the net buying pressure return be *nbpr*$_i$, and let *z_nbp*$_i$, *zscore_nbpr*$_i$ be their respective z-scores.

$$\begin{aligned} e_i = &+1 \quad \text{if} \quad zscore\_nbp_i > 1 \quad \text{AND} \quad zscore\_nbpr_i > 1 \\ &-1 \quad \text{if} \quad zscore\_nbp_i < -1 \quad \text{AND} \quad zscore\_nbpr_i < -1 \\ &NaN \quad \text{otherwise} \end{aligned}$$

Hence, a positive trade direction event can be characterised using those time instances where the number of buyer-initiated trades significantly exceeds the number of seller-initiated trades. Using *NBP log returns* ensures that only the change points are captured and not the time instances that could have been caused by the momentum of past news.

News-trend alignment

The objective of the news-trend alignment task is to create a labelled news data set for training the sentiment classifier. To label a news story, one needs to identify the market trends that the news story is capable of triggering in the market. Given that markets are not efficient, they take time to discount market information. Hence, the news stories released within a given time lag of an interesting market trend are labelled according to that trend. The lag or the amount of time that the market takes in discounting the public information varies with the markets and the assets. The optimal lag period can be found empirically from the sentiment analysis model.

Let Δ be the optimal lag period for a given asset. A trend e_i observed at time t_i is the result of the news released during the time period between $t_{i-\Delta-1}$ and t_{i-1}. Time period $[t_{i-\Delta-1}, t_{i-1}]$ is the alignment window for trend e_i, and the set of news stories released in this alignment window is labelled with the sentiment s_i as per the trend e_i. News stories that occurred simultaneously were labelled with the same sentiment, even though they might be unrelated. This is one of the major limitations of the news trend alignment methodology. In order to handle such confounding labelling instances that were aligned with contradictory sentiments, confounding instances were removed from the analysis of the training data set.

News representation

A real-time news source typically contains two types of news: headline alerts and news stories. Headline alerts have only a headline and no textual discourse. A news story contains a headline, leading paragraphs, and the rest of the story. A market event is generally reported using headline alerts first, followed by the full-body news versions. In journalism, a full-body news story follows an inverted triangle layout, where the value of information is the maximum in the headline and decreases with the subsequent paragraphs (Van Dijk, 1988). A headline alert is a summarised abstraction of the full-body news story that reaches the markets before the main news does. News headlines have been the preferred choice in studies using supervised sentiment analysis (Kohara et al., 1997; Bunningen, 2004; Takahashi et al., 2007; Chan, 2003). This paper conducts sentiment analysis of objective information about news events using news headlines.

To capture subjectivity in the text of a news story, this study used dictionary-based sentiment metrics extracted from the full news text using Loughran and McDonald's (2011) finance sentiment dictionary. For a given news story, the number of positive, negative, and risk-related sentiment words were counted using the dictionary. These counts were aggregated and normalised to obtain the positive, negative, and uncertainty sentiment scores.

This study used a vector space model (VSM) with binary weights to convert the unstructured news headlines to a structured feature vector. One of the major challenges involved in using a VSM is the large size of the feature space. There are three ways to reduce the feature space—semantic normalisation methods such as stemming and lemmatisation, unsupervised feature selection criteria such as term frequency and TF-IDF (term frequency-inverse document frequency [Joachims, 1998]) based thresholds, and supervised feature selection methods such as chi-square, info-gain, and Gini-index. This study used lemmatisation, stop-word removal, and lowercase conversion to normalise the textual data. In addition, we normalised abbreviations to handle different variations such as "US," "U.S.," "United States," etc. using a rule base and a domain knowledge base. Further, the feature space was reduced using term frequency, chi-square, and info-gain-based selection criteria to enhance the classification task.

For a given time lag period Δ, a news data nugget N_i at time t_i is a set of news stories released during $[t_{i-\Delta-1}, t_{i-1}]$.

$$N_i = \{N_{i1}, N_{i2}, \ldots, s_i\}$$

where s_i is the sentiment labelled with the news-trend alignment, $N_{ij} = [f_{ij1}, f_{ij2}, \ldots f_{ijV}, s_i]$ or $[F_{ij}, s_{ij}]$, released at $t_{ij}| \; t_{i-\Delta-1} <= t_{ij} <= t_{i-1}$ and f_{ijk} is 1 if feature f_{ijk} is present in news N_{ij}; else, it is 0. V is the size of the feature space in the training data vocabulary.

If two market trends occur within a period Δ, their alignment windows might overlap and confound the labelling of the news stories. We removed such confounding news stories from the training data set. An optimal alignment window Δ helps to alleviate confounding news stories to an extent. We empirically tested different alignment windows to observe the optimal choice for futures buying behaviour. The training data thus formed is the set of labelled news stories N^T, where

$$N^T = \{N_1, N_2, \ldots\}$$
$$N_t = [f_{t1}, f_{t2}, \ldots f_{tV}, s_t] \text{ or } [F_t, s_t]$$

Classification and evaluation

Supervised sentiment classification is a text classification problem often characterised by a huge feature space and noisy training data. Some of the most popular choices for sentiment analysis and for resolving the text classification problem are naive Bayesian classifiers and support vector machines (SVM). A naive Bayesian classifier is a generative probabilistic classifier that assumes that features belonging to different classes are independent. Even though features belonging to different sentiment classes are far from being independent, naive Bayesian classifiers have proved to be a robust choice for text classification tasks (Zhang, 2004). Support vector machines are discriminative classifiers that translate noisy n-dimensional space to a higher dimensional plane, identifying the best hyperplanes for distinguishing the sentiment classes. They are well suited for large and noisy text classification tasks (Joachims, 1998). We explored the predictability of both naïve Bayesian classifiers and soft-margin SVMs for sentiment classification.

The sentiment analysis was conducted over a period of one year. For validation purposes, this study used (10+1) fold cross-validation data set using data from the first 11 months of the study period. The data from the last month was used as the test data set to verify the efficacy of the proposed algorithm on unseen data. For each of the sentiment classes, false positives were costlier than the false negatives; therefore, precision was more important than recall in the sentiment classification task. Detection error trade-off (DET) curves were used to compare the different sentiment analysis models to understand the impact of different parameters such as the alignment window size. The DET curve is a convenient way to visualise the trade-off between missed detection rate and false alarm rate on normal deviant scales.

Data analysis and results

This study was set in the Indian futures market, and the data sampled for the study was from the year 2009. The year 2009 was an eventful year for Indian financial markets because it was the year immediately following the 2008 global credit crisis. In 2009, Indian financial markets went through a number of troughs and crests with the Satyam bankruptcy crisis (NSE Guide, 2009), a weakening rupee, and the euphoria following the election results (Economic Times, 2009). The duration of the study was from January 1, 2009 to December 31, 2009.

The real-time trades and quotes data were collected from the NSE futures market from January 1, 2009 to December 31, 2009 for the S&P CNX NIFTY Index futures. The S&P CNX NIFTY Index, also referred to as the NIFTY 50 or simply NIFTY, includes 50 companies that cover over 22 sectors of the Indian economy. The trading hours for the NSE futures market are from 9:15 a.m. to 3:30 p.m. Trading data for the quantitative data analysis was considered from 10:00 a.m. to 3:30 p.m. to avoid early trading-hour incoherence.

The news representation module for this study was developed in Java 2.0 using Eclipse IDE. Pre-processing of the news feeds, i.e., tokenisation, lemmatisation, and part of speech (POS) tagging was done using Stanford CoreNLP (v3.2) libraries (Klein & Manning, 2003). The market quantitative data was pre-processed in Matlab R2012b. For classification, we used Matlab's naive Bayesian package and the SVMLight (Joachims, 1998) package. The SVMLight package can handle a large number of features and training records with sparse vector representations with fast optimisation algorithms. The data processing was performed using a 64-bit Intel® Core™ i7-3770 Processor, CPU @ 3.40 GHz, 4 cores, with 8 GB RAM, running on Microsoft Windows 8 Pro.

Real-time news was collected from the *Bloomberg* real-time news service. News stories related either to the NIFTY companies or to the general Indian financial, political, or economic context were extracted. As per these criteria, there were 10,000,78 news stories collected, of which 52% were company-related news stories, and 48% were socio-economic and political news stories. In the real-time news feeds, four types of news reporting formats were observed: headline alerts, full-story news feeds, summary feeds, and structured news feeds. For a given event, headline alerts are the first reports to the market, followed by full-story news feeds. A full-story news feed has three main elements—the headline, the leading paragraph that contains the most important and necessary information about the market event(s), and the rest of the story, which contains further information and analysis about the event or related past events. Structured news feeds are full-story feeds that contain formatted structures such as tables and lists. Often, such news stories are periodical summary news such as "India Stocks Preview" and "Commodities Watch."

In this data set, around 56.7% of the news feeds were headline alerts, and 37% were full-story feeds, of which 7.6% were summary news stories; 5.5% of the news feeds were formatted news stories. (Figure 1).

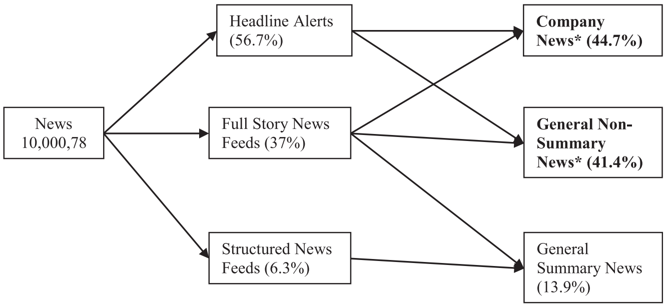
Since the aim of this study was to analyse the impact of new event information entering markets in real time, summary news stories, which includes formatted news feeds, were not considered in the analysis. After filtering out the summary news feeds and formatted news feeds, 88,821 news feeds were found relevant, i.e., around 88% of the news stories were relevant. Of these relevant news stories, around 64% were headline alerts. That is, headline alerts, which arrive in markets before their corresponding full-body news updates, accounted for more than half of the news data. Headlines from 88,821 news instances were used for the sentiment analysis of the futures market.

The news data is represented in vector form using a vector space model. To extract and count important features from the text, the following pre-processing steps were used. Sentence splitting and tokenisation were used to identify unique unigrams in the news headlines. To normalise semantic inflections, unigrams were lemmatised and converted to lower case. After removing stop words, there were 19,753 unique features in the resulting feature space. After removing features that were numeric or alphanumeric and features with document frequency less than two, the number of features left for analysis was 5545. This was significantly less than the original feature space. However, a vector of 5545 with training size of the order of 60,000–70,000 was still a challenging task for the classifier.

We empirically tested two supervised feature selection methods, chi-square and info-gain, for reducing the feature space, thereby improving the classifier's performance. The top 10 features according to TF-IDF, chi-square, and info-gain are shown in Table 1. Since the chi-square and info-gain metrics are supervised, the features ranked with respect to these two metrics hold more classification value compared

Table 1 Sample of top features ranked according to TF-IDF, chi-square, and info-gain metrics.

Top TF-IDF-based features	Top chi-square-based features	Top info-gain based features
India	*satyam*	*satyam*
Rupee	*maharashtra*	*gail*
bln	*gail*	*maharashtra*
mln	*ton*	*reduce*
rupee	*vedanta*	*jsw*
indian	*reduce*	*vedanta*
bank	*jsw*	*cipla*
share	*average*	*oilseed*
price	*oilseed*	*vijaya*
ton	*april*	*reduce*



*News used for Sentiment Analysis

Figure 1 Categories of news from real-time news source Bloomberg.

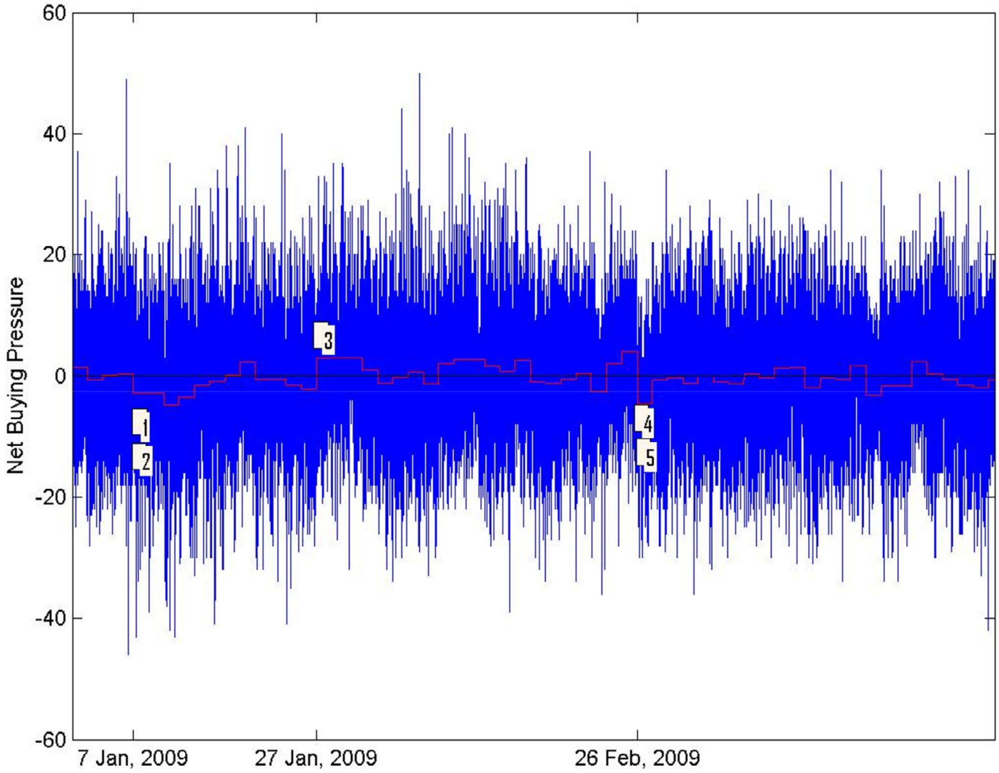


Figure 2 Net buying pressure plot for a sample of time period.
Note: Significant market events marked. Refer to Table 2 for event details.

to those ranked with the TF-IDF based metric. The info-gain metric, with its top 90 percentile words, gave better performance. Hence, the results from the info-gain configuration are reported in this study. Of the 5545 features, 387 features were selected for sentiment classification.

The trade direction, i.e., the net buying pressure, for the real-time bid-ask quotes was calculated using Lee and Ready's (1991) algorithm. There were 79,771 data points in the resulting 1-minute-frequency NBP data of one year. A significant trade point is defined as the time instances that lie above the z-score of two. Thus, the absolute NBP threshold for the market trends was found to be around 20 NBP points. With this criterion, around 2350 data points were found to have "interesting" market sentiment. The 1-minute NBP plot of a sample time period is shown in Figure 2, along with the daily aggregated NBP for better visualisation of trading trends. Some significant NBP data points along with the significant news stories released at that time are presented in Table 2. Significant NBP events marked with increased buyer-initiated or seller-initiated trades are followed by a buyer's market trend or a seller's market trend for a few days, implying significant arbitrage possibilities.

Once the significant trade points were identified, they were aligned with the news data according to the alignment window sizes [*1 mt, 5 mt, 10 mt, 15 mt, 20 mt, 30 mt*]. The news stories in each alignment window were labelled according to the aligned NBP trend (positive, negative, or neutral). The classifier classified positive news instances and negative news instances. Sentiment analysis was conducted across six different training data sets, each created assuming a varying alignment lag [*1 mt, 5 mt, 10 mt, 15 mt, 20 mt, 30 mt*], with cross-validation folds of the first 11 months (10+1) fold for each training data set. We tested two classifiers: MATLAB naive Bayes algorithm and SVMLight. A multinomial naive Bayesian classifier was used as it is best suited for VSM-based text classification. A linear kernel soft margin SVM was used. The SVM's error penalty—the trade-off between training error and margin—was set to $N/||D||$, where N is the size of the training data D. The SVMLight classifier was found to perform better than the naive Bayesian classifier on the validation data set and the test data set. Only the SVMLight results are reported in this paper.

Table 2 News released around some of the "significant" market trends.

Event number (Figure 2)	News timestamp	News headline
1	7 Jan, 2009	Emerging currencies to drop, Morgan Stanley says
2	7 Jan, 2009	SRSR holdings stake in Satyam drops to 3.6%, spokesperson says
3	27 Jan, 2009	Asian stocks climb as U.S. indicators boost exporters, banks
4	26 Feb, 2009	Gold falls for third day on speculation, economy will recover
5	26 Feb, 2009	Ranbaxy lab alleged to fake test results in drug applications

Table 3 Micro-averaged precision, recall, and F-metric of news-based sentiment analysis model on (10+1) fold cross-validation data.

Alignment lag	Precision	Recall	F-metric
1	0.321912	0.313279	0.317537
5	0.312701	0.32759	**0.319972**
10	0.297627	0.319205	0.308039
15	0.281293	0.327945	0.302833
20	0.263084	0.332933	0.293916

Table 4 Micro-averaged precision, recall, and F-metric of news-based sentiment analysis model on test data.

Alignment lag	Precision	Recall	F-metric
1	0.188772	0.300082	0.231754
5	0.262076	0.327389	**0.291114**
10	0.259737	0.308916	0.2822
15	0.268283	0.294497	0.280779
20	0.282828	0.266139	0.27423

The classifier's output and the training data quality were tested with micro-averaged weighted precision, recall, and F-metric that calibrated the efficacy of our two-class prediction model. For trading floor decisions, a good classifier should have lower false positive and false negative rates. Hence, the different classifier results were compared with the DET curves.

Table 3 and Table 4 show the micro-averaged weighted precision, recall, and F-metric of the sentiment classifier on the (10+1) fold cross-validation data and on the test data, respectively. The precision of the classifier decreased as the alignment window size increased. The precision of the classifier is an indication of the noise present in the alignment process. As the alignment window size decreases, we run the risk of missing out on important events altogether, which is evident in the decreased recall at the short lags. The F-metric provides the classifier's overall accuracy in analysing market sentiments. An alignment lag of 5 minutes was found to be optimal. The same findings were corroborated by the DET curves, as shown in Figure 3. Further, including the dictionary-based sentiment features of whole news text made no significant improvement on the predictability of the VSM features of news headlines (Figure 4).

Yadav et al. (2013) compared the precision of automatically labelled news instances and manually labelled news

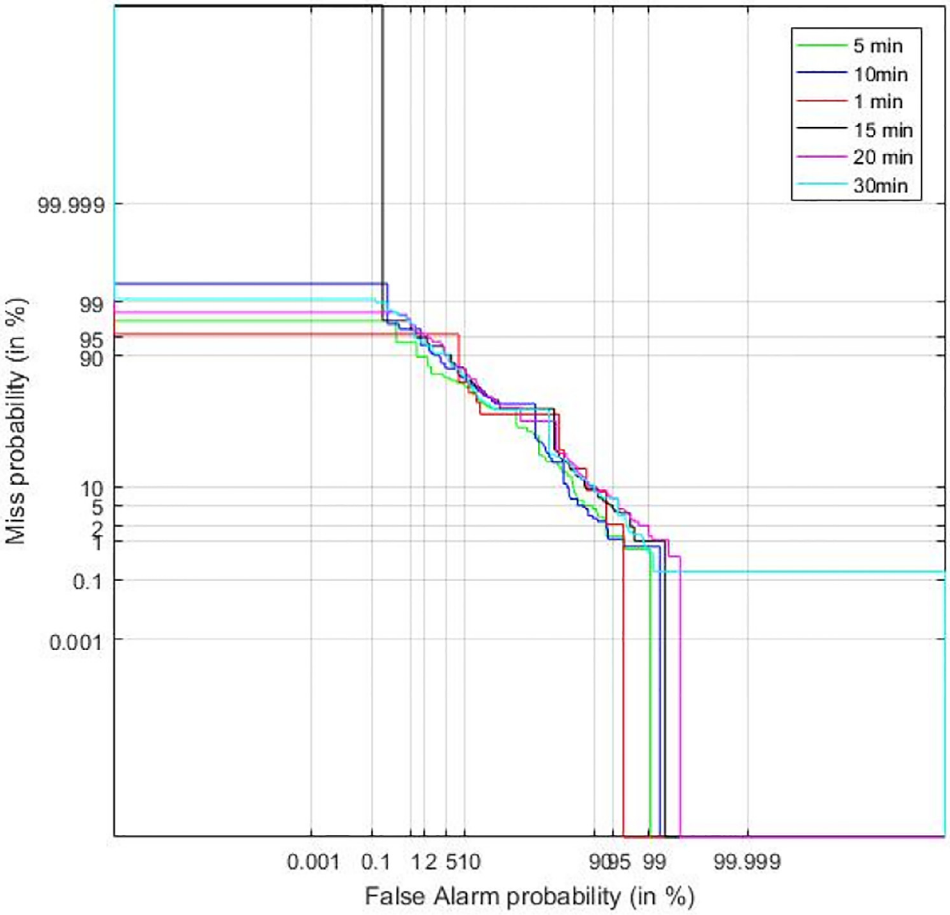


Figure 3 Detection error trade-off plot of the news-based sentiment analysis model comparing different alignment lags.

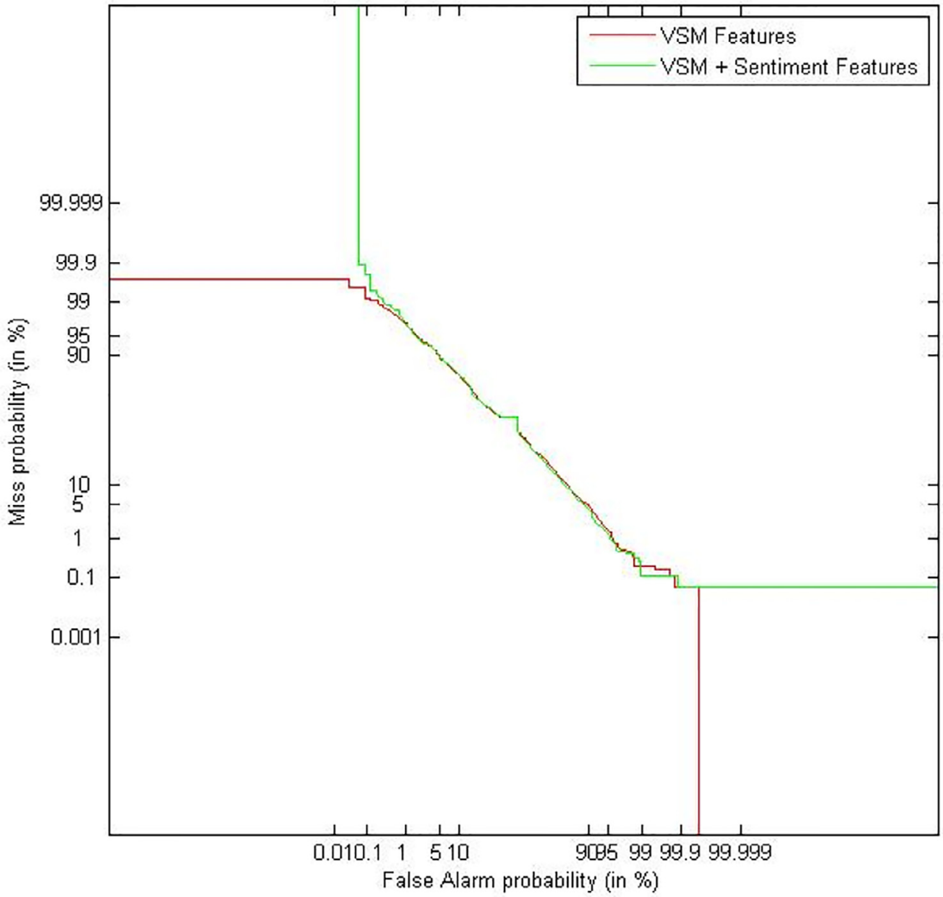


Figure 4 Detection error trade-off plot of the news-based sentiment analysis model with vector space model (VSM) features vs. dictionary-based sentiment features of headlines.

instances in the futures market, and found a window of 5 minutes to be an optimal alignment lag too. This further validates the results observed in this study. A set of significant NBP trade instances and the corresponding [*1 mt, 5 mt, 10 mt, 15 mt, 20 mt, 30 mt*] window-aligned news stories were given to the annotator. The annotator had to decide whether a news story in each alignment window was plausibly responsible for the corresponding trade direction. The results of this comparison concurred with the sentiment analysis findings. In the context of the Indian futures markets, the impact of news stories or market events is reflected in the investors' bid-ask behaviour within a time window of 5 minutes (Table 5).

A major limitation in labelling financial news automatically is the noise that is inherent in financial markets. The gold standard (i.e., the manually labelled training data) presented a curious case that highlights the inherent complexity of labelling news data. In the gold standard data, a sufficient number of news instances are manually labelled by at least two experts and refined such that the mutual agreement is above a given threshold. For instance, in Das and Chen's (2007) work, two experts annotated message board data with the ambiguity coefficient (the labelling mismatch percentage) of 27.54%, i.e. the experts themselves disagreed about the nature of impact of market news 27.54% times. This highlights the inherent complexity in understanding the impact of news on the financial markets by human experts. The inter-rater agreement scores obtained using manually labelled data can be treated as the upper bound on the accuracy expected from an automatically labelled data set. This explains the low precision and recall obtained in sentiment analysis studies.

Table 5 Precision in news trend-based manual labelling.

Alignment lag (minutes)	News-based precision
1	0.46078431
5	**0.51660517**
10	0.49862763
15	0.48341837
20	0.48396362

Conclusion

Financial markets behaviour is an outcome of the prevailing sentiment shaped by the dynamics of the arriving news. Often, markets take time to decipher the sentiment; therefore, there is a lag between the arrival of a news story and the actual trading. An investor who identifies the sentiment early would significantly profit from the anticipated direction. Therefore, understanding the alignment leads, lags, or the window of opportunity is highly crucial. Prior studies have confirmed the importance and length of the alignment lag in the context of stock markets (Gidofalvi, 2001; Robertson et al., 2007). This study analysed the optimal choice of an alignment lag that is appropriate for the index futures market in India. We observed that the sentiment classifier performed best with an alignment lag of 5 minutes in the context of the Indian futures market. This finding was further validated by the manually labelled data. Our results are consistent with the findings reported in prior studies, and indicate that the index futures buying behaviour leads the spot market sentiment. This is consistent with other studies that propound that the derivatives market leads the spot market (Kumar and Jaiswal, 2013; Vipul, 2009).

High-frequency, real-time news plays an important role in the highly competitive financial trading industry. This study captures the impact of real-time news on the futures markets right at the first interface between the market and investors, i.e., at the bid-ask stage. The findings reported here have significant implications for futures trading, especially for high-frequency algorithmic trading. With nearly efficient markets, where market information arrives with minimum delays, the timely analysis of market information is imperative for capturing an arbitrage opportunity. This study establishes that the arbitrage window of opportunity in the Indian futures markets is as short as 5 minutes, which underlines the importance of high-frequency analysis of news on the markets.

Financial markets are highly evolving institutions that have featured random walks, information efficiency, and several behavioural anomalies (Byrne and Brooks, 2008). News happens to be one of the most significant yet underutilised sources of market information; however, it is challenging to model news data. With such a stimulating research problem, we explored a new research direction where the basic unit of analysis is a news headline. There were constraints inherent with the news-based sentiment analysis approach. The foremost limitation faced in this work (and by any news-based sentiment analysis in general) is the chaotic market behaviour. The key assumption of news-based sentiment analysis is that the news moves the markets. While news does move the markets, the markets may move without news too. The second limitation was the extraction of event-related information from the news stories using the vector space model, which did not account for the word order and other semantic roles of news text.

Acknowledgements

This paper is a revised and extended version of the following conference paper: Yadav R., Kumar A., & Kumar A. V. (2013). Supervised sentiment analysis: Engineering a robust automatic training dataset. *Proceedings of the International Conference on Business Analytics and Intelligence*. Indian Institute of Management Bangalore.

References

Antweiler, W., & Frank, M.Z. (2004). Is all that talk just noise? The information content of internet stock message boards. *The Journal of Finance* 59 (3), 1259-1294.

Baker, M., & Wurgler, J. (2007). Investor sentiment in the stock market. *Journal of Economic Perspectives* 21 (2), 129-152.

Bozic, C., & Seese, D. (2011). Neural networks for sentiment detection in financial text. *Proceedings of the 14th International Business Research Conference* Retrieved from https://pdfs.semanticscholar.org/dd7f/4fa6137df5d5ec08efe97150996a548af5e7.pdf.

Bunningen, A.H. (2004). Augmented trading - From news articles to stock price predictions using syntactic analysis. (Master's thesis). University of Twente, Enschede (Netherlands).

Byrne, A., & Brooks, M. (2008). Behavioural finance: Theories and evidence. *The Research Foundation of CFA Institute Literature Review* 1-26.

Chan, W.S. (2003). Stock price reaction to news and no-news: Drift and reversal after headlines. *Journal of Financial Economics* 70 (2), 223-260.

Cheng, S.-H. (2010). Forecasting the change of intraday stock price by using text mining news of stock. In: Proceedings of the Ninth International Conference on Machine Learning and Cybernetics. IEEE, Qingdao, pp. 2605-2609.

Das, S.R., & Chen, M.Y. (2007). Yahoo! for Amazon: Sentiment extraction from small talk on the web. *Management Science* 53 (9), 1375-1388.

Davis, A. K., Piger, J. M., & Sedor, L. M. (2006). Beyond the numbers: An analysis of optimistic and pessimistic language in earnings press releases. *Working Paper 2006-005A*. Working Paper Series, Research Division Federal Reserve Bank of St. Louis.

Drury, B., Torgo, L., & Almeida, J. (2012). Classifying news stories with a constrained learning strategy to estimate the direction of a market index. *International Journal of Computer Science and Applications* 9 (1), 1-22.

Economic Times. (2009). Sensex creates history; two upper circuits in one day. https://economictimes.indiatimes.com/sensex-creates-history-two-upper-circuits-in-one-day/articleshow/4545975.cms.

Fung, G.P., Yu, J.X., & Lam, W. (2003). Stock prediction: Integrating text mining approach using real-time news. IEEE International Conference on Computational Intelligence for Financial Engineering. IEEE, Hong Kong, pp. 395-402.

Garcia, D. (2013). Sentiment during recessions. *The Journal of Finance* 68 (3), 1267-1300.

Gidofalvi, G. (2001). Using news articles to predict stock price movements. Department of Computer Science and Engineering, University of California, San Diego. Retrieved from https://people.kth.se/~gyozo/docs/financial-prediction.pdf.

Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. European Conference on Machine Learning. Springer, Berlin Heidelberg, pp. 137-142.

Klein, D., & Manning, C.D. (2003). Accurate unlexicalized parsing. In: Proceedings of the 41st Meeting of the Association for Computational Linguistics, pp. 423-430.

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.

Kumar, A., & Jaiswal, S. (2013). The information content of alternate implied volatility models: Case of Indian markets. *Journal of Emerging Market Finance* 12 (2), 293-321.

Lavrenko, V., Schmill, M., Lawrie, D., & Ogilvie, P. (2000). Mining of concurrent text and time series. In: Proceedings 6th ACM SIGKDD Int. Conference on Knowledge Discovery and Data Mining, Bostonpp. 37-44.

Lee, C.M.C., & Ready, M.J. (1991). Inferring trade direction from intraday data. *The Journal of Finance* 46 (2), 733-746.

Leinweber, D., & Sisk, J. (2011). Event-driven trading and the "new news". *The Journal of Portfolio Management* 38 (1), 110-124.

Li, X., Deng, X., Wang, F., & Dong, K. (2010). Empirical analysis: News impact on stock prices based on news density. IEEE International Conference on Data Mining Workshops, pp. 585-592.

Loughran, T., & McDonald, B. (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *The Journal of Finance* 66 (1), 35-65.

Mittermayer, M. A., & Knolmayer, G. F. (2006). Text mining systems for market response to news: A survey. *Working Paper No 184*. Bern, Switzerland: Institute of Information Systems, University of Bern.

Nassirtoussi, A.K., Aghabozorgi, S., Wah, T.Y., & Ngo, D.C. (2014). Text mining for market prediction: A systematic review. *Expert Systems with Applications* 41 (16), 7653–7670.

Nikfarjam, A., Muthaiyah, S., & Emadzadeh, E. (2010). Text mining approaches for stock market prediction. In: Proceedings of the 2nd International Conference on Computer and Automation Engineering (ICCAE 2010). IEEE, pp. 256–260. Vol. 4.

NSE Guide. (2009). Stock Market: Satyam scandal rattles stocks. http://nseguide.com/stock-views/stock-market-satyam-scandal-rattles-stocks/.

Robertson, C., Geva, S., & Wolff, R.C. (2007). Can the content of public news be used to forecast abnormal stock market behaviour? Seventh IEEE International Conference on Data Mining, pp. 637-642.

Schumaker, R., & Chen, H. (2008). Evaluating a news-aware quantitative trader: The effect of momentum and contrarian stock selection strategies. *Journal of the American Society for Information Science and Technology* 59 (2), 247–255.

Takahashi, S., Takahashi, M., Takahashi, H., & Tsuda, K. (2007). Analysis of the relation between stock price returns and headline news using text categorization. 11th International Conference, KES 2007, XVII Italian Workshop on Neural Networks Proceedings, Part II. Vietri sul Mare, Italy. Springer, pp. 1339–1345.

Tetlock, P.C. (2007). Giving content to investor sentiment: The role of media in the stock market. *The Journal of Finance* 62 (3), 1139–1168.

Tetlock, P.C. (2011). All the news that's fit to reprint: Do investors react to stale information? *Review of Financial Studies* 24 (5), 1481–1512.

Uhl, M., Pedersen, M., & Malitius, O. (2015). What's in the news? Using news sentiment momentum for tactical asset allocation. *The Journal of Portfolio Management* 41 (2), 100–112.

Van Dijk, T.A. (1988). Structures of News. In: Van Dijk, T.A. (Ed.), News as discourse. Erlbaum, Hillside, NJ, pp. 17–94. Ch 2.

Vipul (2009). Mispricing, volume, volatility and open interest. *Journal of Emerging Market Finance* 7 (3), 263–292.

Wuthrich, B., Cho, V., Leung, S., Permunetilleke, D., Sankaran, K., Zhang, J., & Lam, W. (1998). Daily stock market forecast from textual web data. 1998 IEEE International Conference on Systems, Man, and Cybernetics. IEEE, pp. 2720–2725. Vol. 3.

Yadav, R., Kumar, A., & Kumar, A.V. (2013). Supervised sentiment analysis: Engineering a robust automatic training dataset. In: Proceedings of International Conference on Business Analytics and Intelligence. Indian Institute of Management Bangalore.

Yoo, P.D., Kim, M.H., & Jan, T. (2005). Machine learning techniques and use of event information for stock market prediction: A survey and evaluation. CIMCA '05 Proceedings of the International Conference on Computational Intelligence for Modelling, Control and Automation and International Conference on Intelligent Agents, Web Technologies and Internet Commerce (CIMCA-IAWTIC'06). IEEE Computer Society, Washington, DC, pp. 835–841 . Vol. 2.

Zhang, H. (2004). The optimality of naive Bayes. *American Association for Artificial Intelligence* 1 (2), 3 Retrieved from https://www.aaai.org/Papers/FLAIRS/2004/Flairs04-097.pdf.