Expert Systems with Applications 42 (2015) 306–324

Contents lists available at ScienceDirect

Expert Systems with Applications

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

Text mining of news-headlines for FOREX market prediction: A Multi-layer Dimension Reduction Algorithm with semantics and sentiment

Arman Khadjeh Nassirtoussi [a,*], Saeed Aghabozorgi [a], Teh Ying Wah [a], David Chek Ling Ngo [b]

[a] Department of Information Science, Faculty of Computer Science & Information Technology, University of Malaya, 50603 Kuala Lumpur, Malaysia
[b] Research & Higher Degrees, Sunway University, No. 5, Jalan University, Bandar Sunway, 46150 Petaling Jaya, Selangor DE, Malaysia

ARTICLE INFO

Article history:
Available online 10 August 2014

Keywords:
News mining
News semantic analysis
Market sentiment analysis
Market prediction
FOREX prediction

ABSTRACT

In this paper a novel approach is proposed to predict intraday directional-movements of a currency-pair in the foreign exchange market based on the text of breaking financial news-headlines. The motivation behind this work is twofold: First, although market-prediction through text-mining is shown to be a promising area of work in the literature, the text-mining approaches utilized in it at this stage are not much beyond basic ones as it is still an emerging field. This work is an effort to put more emphasis on the text-mining methods and tackle some specific aspects thereof that are weak in previous works, namely: the problem of high dimensionality as well as the problem of ignoring sentiment and semantics in dealing with textual language. This research assumes that addressing these aspects of text-mining have an impact on the quality of the achieved results. The proposed system proves this assumption to be right. The second part of the motivation is to research a specific market, namely, the foreign exchange market, which seems not to have been researched in the previous works based on predictive text-mining. Therefore, results of this work also successfully demonstrate a predictive relationship between this specific market-type and the textual data of news. Besides the above two main components of the motivation, there are other specific aspects that make the setup of the proposed system and the conducted experiment unique, for example, the use of news article-headlines only and not news article-bodies, which enables usage of short pieces of text rather than long ones; or the use of general financial breaking news without any further filtration.

In order to accomplish the above, this work produces a multi-layer algorithm that tackles each of the mentioned aspects of the text-mining problem at a designated layer. The first layer is termed the Semantic Abstraction Layer and addresses the problem of co-reference in text mining that is contributing to sparsity. Co-reference occurs when two or more words in a text corpus refer to the same concept. This work produces a custom approach by the name of Heuristic-Hypernyms Feature-Selection which creates a way to recognize words with the same parent-word to be regarded as one entity. As a result, prediction accuracy increases significantly at this layer which is attributed to appropriate noise-reduction from the feature-space.

The second layer is termed Sentiment Integration Layer, which integrates sentiment analysis capability into the algorithm by proposing a sentiment weight by the name of SumScore that reflects investors' sentiment. Additionally, this layer reduces the dimensions by eliminating those that are of zero value in terms of sentiment and thereby improves prediction accuracy.

The third layer encompasses a dynamic model creation algorithm, termed Synchronous Targeted Feature Reduction (STFR). It is suitable for the challenge at hand whereby the mining of a stream of text is concerned. It updates the models with the most recent information available and, more importantly, it ensures that the dimensions are reduced to the absolute minimum.

The algorithm and each of its layers are extensively evaluated using real market data and news content across multiple years and have proven to be solid and superior to any other comparable solution. The proposed techniques implemented in the system, result in significantly high directional-accuracies of up to 83.33%.

On top of a well-rounded multifaceted algorithm, this work contributes a much needed research framework for this context with a test-bed of data that must make future research endeavors more convenient. The produced algorithm is scalable and its modular design allows improvement in each of its layers in future research. This paper provides ample details to reproduce the entire system and the conducted experiments.

* Corresponding author.
E-mail address: armankhnt@gmail.com (A. Khadjeh Nassirtoussi).

http://dx.doi.org/10.1016/j.eswa.2014.08.004
0957-4174/

1. Introduction

Today's biggest economies of the world are market-economies. With markets being the heart of economies, it is paramount to understand, utilize and predict them to the betterment of society. At the core of every market lie supply–demand equilibriums. Market participants provide supply or demand into the markets based on their perception of the world. Human perception is limited to the information available. Information is made available constantly via news channels. Hence, undeniably news content has an impact on market-movements. However, a more granular quantification of this relationship between markets and the news has been extremely challenging, because news contains unstructured information in form of language. One approach to address unstructured data and extract structured data from it is the development of specialized search engines like the financial news semantic search engine by Lupiani-Ruiz et al., 2011. However, a search engine like the above is limited to extracting the available numeric data in the texts. Deciphering language by machine constitutes the complex field of natural language processing (NLP). From this perspective this work lies at the intersection of NLP and opinion mining or sentiment analysis which are recently being increasingly researched for many emerging needs (Cambria, Schuller, Yunqing, & Havasi, 2013). There have been some early-stage efforts to make stock-market related predictions based on news-text (Hagenau, Liebmann, & Neumann, 2013; Mittermayer, 2004; Schumaker, Zhang, Huang, & Chen, 2012; Tetlock, Saar-Tsechansky, & Macskassy, 2008; Wuthrich et al., 1998) and very few Foreign-Exchange-Market (FOREX) related ones (Peramunetilleke & Wong, 2002). However, there are some similarities between the two problems. When dealing with news and market-movements the basic strategy can be to try to draw a statistical relationship between the appearance of words and the market movements. In this scenario most words are representing themselves as features in a feature-vector matrix and the technique is termed as Bag-of-Words (Mahajan, Dey, & Haque, 2008; Schumaker et al., 2012; Wuthrich et al., 1998). Bag-of-Words has been widely used in many of the related works to markets and news and its primary downside is the huge number of features that it produces, which very easily leads to the curse-of-dimensionality (Pestov, 2013). Moreover, many words may represent the same idea, concept or thing and it may make a lot more sense to somehow have them abstracted accordingly. Thereby, proposing solutions to the above two challenges, namely, the latter or feature-selection in an abstracted form and the former, a way to tackle the curse-of-dimensionality via a feature-reduction technique are center-pieces of this work.

There are two main areas of contributions made in this work. A brief summary is provided below:

A – Proposal of novel text-mining methods in 3 areas:

1. Semantic-Abstraction and Integration via a novel feature-selection technique, termed, Heuristic-Hypernyms.
2. Sentiment Integration via a novel sentiment-weighting mechanism, termed, SumScore.
3. Dimensionality Reduction via a novel feature-reduction technique, termed, Synchronous Targeted Feature-Reduction.

B – Exploration of a specific novel use-case, namely: "Short-term FOREX prediction based on news-headlines".

1. Exploration of a new market-type through predictive text-mining. The predictive text-mining of news has not been explored before in the FOREX market to the best of our knowledge.
2. Usage of news article-headlines rather than news article-bodies. News-headlines have been explored in extremely few market-predictive text-mining research works before.
3. A novel solution to enable short-term prediction of 1 to 3 h after news-release.

All of the above 6 items are contributions of this work. The hybrid of the above new techniques produces results that are significantly higher compared to scenarios without them. A directional accuracy as high as 83.33% is achieved in experiments conducted on Euro/USD currency-pair intraday-movements which is laid out and discussed in detail later in this paper.

In the rest of this paper these sections follow: 2 – Literature review; 3 – Problem description; 4 – System description; 5 – Experimental results and evaluation; 6 – Concluding remarks and future research.

2. Problem description

The specific problem that this research addresses and its requirements is briefed in the below.

The first aspect of the problem definition is a focus on a specific market-type. In general, there are multiple types of financial markets, namely: 1 – Capital markets (Stock and Bond), 2 – Commodity markets, 3 – Money markets, 4 – Derivative markets, 5 – Future markets, 6 – Insurance markets and 7 – Foreign exchange markets. As their names imply, different assets are traded in each market; therefore they demonstrate different behaviors and separate research is conducted on each of them. As it is pointed out more specifically in the literature review section of this work, most of the works in the literature concerning some kind of usage of text-mining for a predictive purpose in a financial market is mostly attending to the stock-markets and specific company stocks based on textual content about those companies. Hence, this work enters a less explored financial market namely the foreign exchange market (FOREX) which facilitates the trading of currencies.

Furthermore, this work aims to take into use uncategorized breaking news rather than categorized news based on topic or company, etc. As pointed out in the literature the usual explored path in the past works is to isolate company-specific news, for example, and make predictions for the stock of a company based on that. However, the news channel that is used for this experiment is for financial breaking news. A focus on financial breaking news rather than a source of news that has all kinds of news pieces released is assumed to provide logical relevance and avoid noise. This is inspired by what traders in financial markets actually read. But no further categorization of news is utilized.

Moreover, in terms of the length of text, subject of this research is short-texts of news-headlines. The requirement of using news article-headlines rather than news article-bodies creates a text-mining focus on short texts rather than long texts for the proposed system. Naturally, when short pieces of texts are concerned there is less repetition of words in the same document and there are also fewer irrelevant words. Therefore, in such a context the level of significance of a word in a news piece cannot be determined by its repetition within it; however, at the same time there is less noise in the space as headlines are usually concise.

In terms of prediction time-line in the financial markets, both short and long-term predictions are subjects of research. In this work, however, the short-term prediction is explored as sudden impacts of news on the market are of interest and with the passage of time the number of factors producing noise on the initial impact increases. The short-term prediction that targets market-moves within the same day is termed as intra-day market prediction. To be specific, what is predicted is the directional movement (Up or Down) of the market (price of a currency pair e.g. EUR/USD) 1-h after the end of a 2-h interval which includes the news-headlines released within it. This upwards or downwards movement at the

1-h point after the interval is determined in relation to the point at which the market was 1-h before the interval. This latter margin before the news-release interval ensures that the news release is indeed after the first point in time as different news sources may release breaking news with a slight time difference. The details of this structure are fully elaborated in a later section titled: *System Description*. However, at this point it suffices to mention that the system has a prediction time-line of 1-h after a 2-h news-interval which means the impact that is monitored and taken into consideration by the system can be 1 to 3 h away from the exact release time of a new-headline.

In terms of required accuracy for practical use of such prediction system, because a binary decision between Up and Down is concerned, any results above 50% prediction accuracy is of interest and significance from a statistical perspective. Almost all of the previous works also, as listed in the next literature review section, compare their results with the odds of chance of 50% in such a context. Practical traders agree, too, that a system that can be accurate more than half of the time can be of value to them on a day to day basis. However, accuracy results of recent comparable efforts in the same research space are also provided in this work for a better evaluation of the results achieved here.

3. Literature review

Generally when predicting financial markets is concerned there are two primary approaches: technical and fundamental. Technical approaches deal with using the historic market-data that is quantitative and make predictions based on that. There are many examples of utilizing or improving artificial intelligence algorithms to predict the foreign exchange market and recognize patterns within it based on technical data. Sermpinis, Laws, Karathanasopoulos, and Dunis (2012) study two promising classes of artificial intelligence models, the Psi Sigma Neural Network (PSI) and the Gene Expression algorithm (GEP), when applied to the task of forecasting and trading the EUR/USD exchange rate. Sermpinis, Theofilatos, Karathanasopoulos, Georgopoulos, and Dunis (2013) study forecasting foreign exchange rates with adaptive neural networks using radial-basis functions and Particle Swarm Optimization. There are many more examples of usage of neural networks for market-predictive technical analysis (Anastasakis & Mort, 2009; Ghazali, Hussain, & Liatsis, 2011; Vanstone & Finnie, 2010). Bahrepour, Akbarzadeh, Yaghoobi, and Naghibi (2011) have introduced an adaptive ordered fuzzy time series with application to the foreign exchange market (FOREX). Huang, Chuang, Wu, and Lai (2010) present a method based on Chaos-based support vector regressions for exchange rate forecasting. Premanode and Toumazou (2013) propose a new algorithm, differential Empirical Mode Decomposition (EMD) for improving prediction of exchange rates under support vector regression (SVR). Mabu, Hirasawa, Obayashi, and Kuremoto (2013) utilize rule-based genetic network programming for creating stock trading signals. However, technical analysts simply believe there are patterns in a market graph that can be detected. This principle can be challenged and may be very context specific (Yu, Nartea, Gan, & Yao, 2013).

On the other hand, in fundamental analysis, analysts look at fundamental data that is available to them from different sources from outside market-historic-data and make assumptions based on those. Such sources are composed of information about geopolitics, financial environment and business principles in general. However, it is challenging to automate fundamental analysis due to the nature of fundamental data. Fundamental data may come from structured and numeric sources like macro-economic data or regular financial reports from banks and governments. This aspect prediction based on fundamental data has been occasionally researched (Chatrath, Miao, Ramchander, & Villupuram, 2014; Fasanghari & Montazer, 2010; Khadjeh Nassirtoussi, Ying Wah, & Ngo Chek Ling, 2011). However, the vast majority of fundamental data is available in an unstructured format as textual data. A lot of such information is found in news or financial news to be more specific. Fundamental analysts consume such sources of data. Hence, how fundamental analysts react to news has an impact on the market. However, determining fundamental value in a market is heavily dependent on the analyst's perception and is studied in the context of foreign exchange market by Kaltwasser (2010). In this context studying the reactions of the foreign exchange market (FOREX) to uncategorized financial news becomes very interesting and appropriate as one is trying to relate the general sentiment to the market that is based on all sorts of news rather than studying filtered-news that is only relevant to a specific company stock.

Sentiment analysis deals with detecting the general sentiment that is available in online resources and social media to understand how people feel about a topic. It can be used in areas of research like emotion detection in suicide notes as done by Desmet and Hoste (2013); or it can be used, for example, to determine consumer sentiment towards a brand (Mostafa, 2013). Moraes, Vasconcelos, Prado, Almeida, and Gonçalves, (2013) analyze the polarity of micro-reviews in Foursquare, which is one of the currently most popular location-based social networks. Such sentiment or polarity classification provides useful tools for opinion summarization, which can help business owners as well as potential new customers to quickly obtain a predominant view of the opinions posted by users at a specific venue. Sentiment of short online textual snippets like tweets is actively studied as well (Ghiassi, Skinner, & Zimbra, 2013; Ikeda, Hattori, Ono, Asoh, & Higashino, 2013; Kontopoulos, Berberidis, Dergiades, & Bassiliades, 2013). Tweets are also studied in relation to prediction of stock markets (Bollen, Huina, & Zeng, 2010).

Sentiment analysis of news can be a good source to tap into for market prediction as it expresses the point of view and sentiment of opinion leaders, forms the public opinion to an extent and triggers public reactions. The impact of news on stock markets of different companies and regions has been the target of emerging extensive research (Hagenau et al., 2013; Huang, Liao, Yang, Chang, & Luo, 2010; Nizer & Nievola, 2012; Schumaker et al., 2012). Surprisingly, there are too few research efforts on impact of textual data in news content on the foreign exchange market for example Peramunetilleke and Wong (2002) is one of the rare examples that studies impact of news-headlines on FOREX. Evans and Lyons (2008) term a phenomenon as "news puzzle", and argue that directional effects are harder to detect in exchange rates since they are likely to be swamped by other factors. A recent work of Chatrath et al. (2014) studies the currency jumps, cojumps, however, it is based on the role of macro-news which entails structured scheduled macro-economic news announcements that reveal specific indexes like the unemployment rate of a country or its inflation rate, etc.

The core mechanics of market related prediction through text mining is only at its early stages and is far from perfection even in the stock markets related works which composes most of the literature (Groth & Muntermann, 2011; Hagenau et al., 2013; Huang et al., 2010; Nizer & Nievola, 2012; Schumaker et al., 2012; Zhang, 2011). In this field, directional classification of news impact plays a major role. Each of the following aspects of text-mining, needs to be closely studied, customized and advanced in the market-prediction context and more specifically in the FOREX-prediction context.

In text mining in general, it is verified that the feature extraction (Günal, Ergin, Gülmezoğlu, & Gerek, 2006), feature selection (Feng, Guo, Jing, & Hao, 2012; Taşcı & Güngör, 2013), classification method (Tan, Wang, & Wu, 2011) and preprocessing (Uysal & Gunal, 2014) have substantial impact on the success of text classification processes. Luo, Chen, and Xiong (2011) suggest traditional

term weighting schemes in text categorization, such as TF-IDF, only exploit the statistical information of terms in documents. They propose a novel term weighting scheme by exploiting the semantics of categories and indexing terms. Specifically, the semantics of categories are represented by senses of terms appearing in the category labels as well as the interpretation of them by WordNet (Miller, 1995). Li, Yang, and Park (2012) improve text categorization performance through a corpus-based thesaurus and WordNet, employing the k-NN algorithm and the back propagation neural network (BPNN) algorithms as the classifiers. Jiang, Pang, Wu, and Kuang (2012) improve k-NN algorithm for text categorization by combining it with a constrained one pass clustering algorithm. Naive Bayes is another algorithm that is used for text classification successfully (Chen, Huang, Tian, & Qu, 2009).

Feature selection and feature reduction are extremely important phases in classification algorithms and there have been multiple efforts to improve them in a variety of scenarios. Aghdam, Ghasem-Aghaee, and Basiri (2009) introduce a feature selection and reduction method using ant colony optimization. Shi, He, Liu, Zhang, and Song (2011) take test criteria such as frequency, dispersion and concentration indices into account and proposes an improved dimension reduction method and feature weighting method in an effort to make the selection more representative and the weighting of characteristic features more reasonable. Berka and Vajteršic (2013) make another effort in improvement of dimensionality reduction and introduce an algorithm that replaces rare terms by computing a vector which expresses their semantics in terms of common terms.

Khadjeh Nassirtoussi, Aghabozorgi, Ying Wah, and Ngo (2014) presents a comprehensive systematic review of the text mining approaches used in the past for a market predictive purpose. This work is a continuation and based on the results of that review. It concludes the general flow for a market-predictive text-mining system to be as illustrated in Fig. 1.

Khadjeh Nassirtoussi et al. (2014) puts an emphasis on the type of used dataset, the structure of pre-processing and the type of machine learning algorithm used in categorizing the available systems. Pre-processing phase includes every activity that is required to transform the raw data into a machine readable format. Within it feature-selection, -reduction, and -representation are identified as noteworthy aspects of the mechanisms of the systems. Table 1 lists the reviewed available works and points out their feature-selection, -reduction and -representation strategy (Khadjeh Nassirtoussi et al., 2014).

Another important aspect in reviewing the available works that are in this space is the type of machine-learning algorithm that is used in the systems. Table 2 groups the available works based the type of the machine learning algorithm utilized under SVM, Regression Algorithms, Naïve Bayes, Decision Rules or Trees, Combinatory Algorithms, Multi-algorithm experiments (Khadjeh Nassirtoussi et al., 2014).

Table 2 compares the available works from a further number of aspects as well, namely: the time periods used for training and testing samples, implementation of a concept termed 'sliding window' in the time-frames, availability of aspects related to integration of semantics and syntax, availability of technical market-data in addition to the textual data as well as some of the third party software pieces or algorithms used. For a detailed description of each of the above Khadjeh Nassirtoussi et al. (2014) can be referred to.

As a result of the above literature review of available systems a number of text-mining aspects are weakly constructed, if at all, in the available works. This work proves that contribution to betterment of these aspects can yield promising results. The identified text-mining aspects that are targeted in this work are:

1. Text-mining process-phases (Table 1): in the text-mining process, innovation is required in critical aspects of *feature-selection, feature-reduction and feature-representation* as most of the reviewed works do not enter deeper concepts in these areas of text-mining once they are dealing with the rather specialized field of market-predictive text-mining. From the above 3, feature-reduction is identified to be a major contributing factor to the complication and inaccuracy of the text-mining process.
2. Utilization of semantics on 2 levels: firstly, integration of it through implementation of ontologies and dictionaries rather than simply ignoring it, which is the case in many of the available works (Table 2); Secondly, tackling the problem of co-reference which is having multiple words referring to the same concept or thing.
3. Utilization of sentiment: consideration of human emotions and sentiments is a vital part of what is entailed in any form of language representation including the textual format. However, as

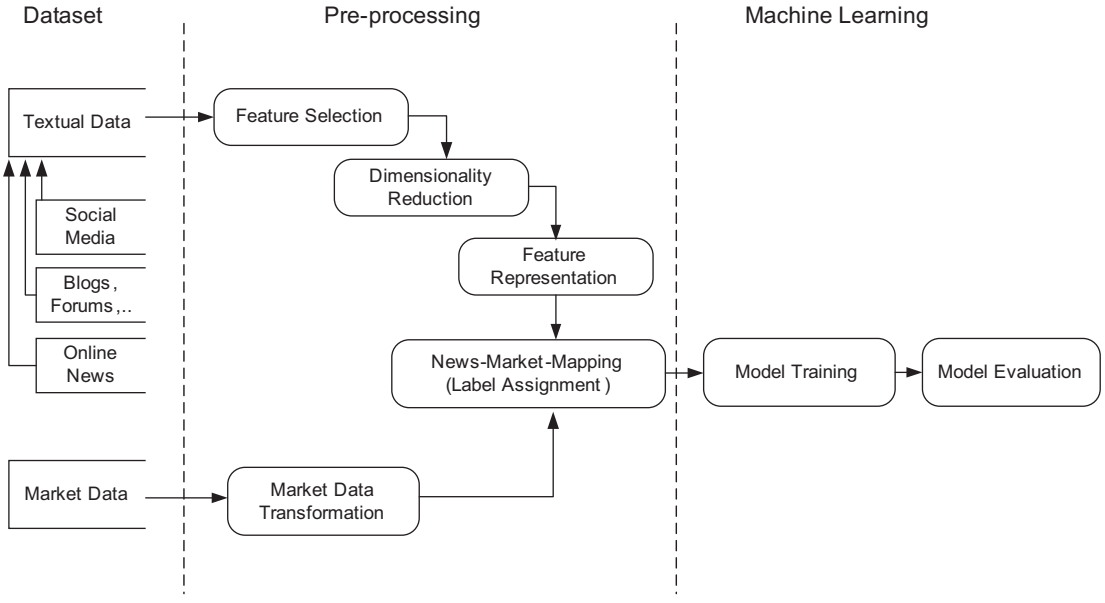

Fig. 1. Major components and flow in market-predictive text-mining systems.

Table 1
Feature-selection, -reduction and representation in the reviewed works.

References	Feature selection	Dimensionality reduction	Feature representation
Wuthrich et al. (1998)	Bag-of-words	Pre-defined dictionaries (word-sequences by an expert)	Binary
Peramunetilleke and Wong (2002)	Bag-of-words	Set of keyword records	Boolean, TF-IDF, TF-CDF
Pui Cheong Fung, Xu Yu, and Wai (2003)	Bag-of-words	Stemming, conversion to lower-case, removal of punctuation, numbers, web page addresses and stop-words	TF-IDF
Werner and Myrray (2004)	Bag-of-words	Minimum information criterion (top 1000 words)	Binary
Mittermayer (2004)	Bag-of-words	Selecting 1000 terms	TF-IDF
Das and Chen (2007)	Bag-of-words, triplets	Pre-defined dictionaries	Different discrete values for each classifier
Soni et al. (2007)	Visualization	Thesaurus made using term extraction tool of van Eck	Visual coordinates
Zhai et al. (2007)	Bag-of-words	WordNet Thesaurus (stop-word removal, POS tagging, higher level concepts via WordNet). Top 30 concepts.	Binary, TF-IDF
Rachlin, Last, Alberg, and Kandel (2007)	Bag-of-words, commonly used financial values	Most influential keywords list (Automatic extraction)	TF, Boolean, Extractor software output
Tetlock et al. (2008)	Bag-of-words for negative words	Pre-defined dictionary. Harvard-IV-4 psychosocial dictionary.	Frequency divided by total number of words
Mahajan et al. (2008)	Latent Dirichlet Allocation (LDA)	Extraction of twenty-five topics	Binary
Butler and Kešelj (2009)	Character N-grams, three readability scores, last year's performance	Minimum occurrence per document.	Frequency of the n-gram in one profile
Schumaker and Chen (2009)	Bag of Words, noun phrases, named entities	Minimum occurrence per document	Binary
Li (2010)	Bag-of-Words, tone and content	Pre-defined dictionaries	Binary, Dictionary value
Huang et al. (2010)	Simultaneous terms, ordered pairs	Synonyms replacement	Weighted based on the rise/fall ratio of index
Groth and Muntermann (2011)	Bag-of-words	Feature scoring methods using both Information Gain and Chi-Squared metrics	TF-IDF
Schumaker et al. (2012)	OpinionFinder overall tone and polarity	Minimum occurrence per document	Binary
Lugmayr and Gossen (2012)	Bag-of-words	Stemming	Sentiment value
Yu, Duan, and Cao (2013)	Bag-of-words	Not mentioned	Binary
Hagenau et al. (2013)	Bag-of-words, Noun Phrases, Word-combinations, N-grams	Frequency for news, Chi^2-approach and bi-normal separation (BNS) for exogenous-feedback-based feature selection, Dictionary.	TF-IDF
Jin et al. (2013)	Latent Dirichlet Allocation (LDA)	Topic extraction, top topic identification by manually aligning news articles with currency fluctuations	Each article's topic distribution
Chatrath et al. (2014)	Structured data	Structured data	Structured data
Bollen and Huina (2011)	By OpinionFinder	By OpinionFinder	By OpinionFinder
Vu, Chang, Ha, and Collier (2012)	Daily aggregate number of positives or negatives on Twitter Sentiment Tool (TST) and an emoticon lexicon. Daily mean of Pointwise Mutual Information (PMI) for pre-defined bullish-bearish anchor words	Pre-defined company related keywords, Named Entity Recognition based on linear Conditional Random Fields (CRF)	Real number for Daily Neg_Pos and Bullish_Bearish

the field of market-predictive text-mining is emerging, this aspect is yet to be better integrated in the process (Khadjeh Nassirtoussi et al., 2014).

Additionally, as pointed out in detail in Khadjeh Nassirtoussi et al. (2014), the context in which the market-predictive text-mining solution is conceived and implemented is crucial as the nature of text-mining solutions are highly context-specific. Therefore, it is important to explore different contexts in terms of market-types and input-text-types for such predictive systems. Hence, this work has chosen a market-type that is not explored before as the context of a predicative text-mining solution, namely: the foreign exchange market (FOREX) (Khadjeh Nassirtoussi et al., 2014). Successful performance of a predictive system in a new market-type, not only demonstrates the capability of the system but also helps establish the existence of a predictive-relationship between the content of the news and the target-market as this relationship is under ongoing investigation from an economics-research perspective as elaborated in Khadjeh Nassirtoussi et al. (2014).

This work addresses the above text-mining concerns by putting forward an innovative multi-layer text-mining algorithm in the specified context that entails components dealing with semantics and sentiment in a way that feature-selection, -reduction and -representation are enhanced. This is done by introducing a feature-selection method termed Heuristic-Hypernyms, a feature-reduction method termed Synchronous Targeted Feature-Reduction (STFR) and a feature-weighting method termed SumScore-Weighting. The proposed approach composed of these methods produces significant experimental results. A detailed system description and experimental results are provided next.

4. System description

The job of the system that is proposed in this work is to take time-stamped news headlines as input, group them together based

Table 2
Classification algorithms, semantics and syntax integration and other machine learning aspects of the reviewed works.

Reference	Algorithm type	Algorithm details	Training vs. testing volume and sampling	Sliding window	Semantics	Syntax	News & tech. data	Software
Pui Cheong Fung et al. (2003)	*SVM*	SVM-Light	First 6 consecutive months vs. the last month	No	No	No	No	Not mentioned
Mittermayer (2004)		SVM-Light	200 vs. 6,002 examples	No	No	No	No	NewsCATS
Soni et al. (2007)		SVM with standard linear kernel	80% vs. 20%	No	Yes	No	No	LibSVM package
Zhai et al. (2007)		SVM with Gaussian RBF kernel and polynomial kernel	First 12 months vs. the remaining two months	No	Yes	No	Yes	Not mentioned
Schumaker and Chen (2009)		SVM	Not mentioned	No	Yes	Yes	Yes	Arizona Text Extractor (AzTeK) & AZFin Text.
Lugmayr and Gossen (2012)		SVM	Not mentioned	No	Yes	No	Yes	SentiWordNet
Hagenau et al. (2013)	*Regression Algorithms*	SVM with a linear kernel, SVR	Not mentioned	No	Yes	Yes	Yes	Not mentioned
Schumaker et al. (2012)		SVR	Not mentioned	No	Yes	No	Yes	OpinionFinder
Jin et al. (2013)		Linear regression model	Previous day vs. a given day (2 weeks for regression)	Yes	Yes	No	No	Forex-foreteller, Loughran-McDonald financial dic., AFINN dic.
Chatrath et al. (2014)		Stepwise Multivariate Regression Model	Not applicable	No	No	No	No	Not mentioned
Tetlock et al. (2008)		OLS Regression	30 and 3 trading days prior to an earnings announcement	Yes	Yes	No	No	Harvard-IV-4 psychosocial dictionary
Yu et al. (2013)	*Naïve Bayes*	Naïve Bayes	Not mentioned	No	Yes	No	No	Open-source Natural Language Toolkit (NLTK)
Li (2010)		Naïve Bayes & dictionary-based	30,000 randomly vs. itself and the rest	No	No	No	No	Diction, General Inquirer, the Linguistic Inquiry, Word Count (LIWC).
Peramunetilleke and Wong (2002)	*Decision Rules or Trees*	Rule classifier	22 September 12:00 to 27 September 9:00 vs. 9:00 to 10:00 on 27 September	Yes	Yes	No	No	Not mentioned
Huang et al. (2010)		Weighted association rules	2005 June to 2005 October vs. 2005 November	No	Yes	Yes	No	Not mentioned
Rachlin et al. (2007)		C4.5 Decision Tree	Not mentioned	No	No	No	Yes	Extractor Software package
Vu et al. (2012)		C4.5 Decision Tree	Trained by previous day features	Yes	Yes	Yes	No	CRF++ toolkit, Firehose, TST, CMU POS Tagger, AltaVista
Das and Chen (2007)	*Combinatory Algorithms*	Combination of different classifiers	1,000 vs. the rest	No	Yes	Yes	No	General Inquirer
Mahajan et al. (2008)		Stacked classifier	August 05–December 07 vs. January 08–April 08	No	Yes	No	No	Not mentioned
Butler and Kešelj (2009)		CNG distance measure & SVM & combined	Year x vs. years x − 1 and x − 2. & all vector representations vs. particular testing year	Yes	No	No	Yes	Perl n-gram module Text::Ngrams developed by Keselj . LIBSVM
Bollen and Huina (2011)		Self-organizing fuzzy neural network (SOFNN)	28 February to 28 November vs. 1 to 19 December 2008	No	N/A	N/A	No	GPOMS, OpinionFinder
Wuthrich et al. (1998)	*Multi-algorithm experiments*	k-NN, ANNs, naïve Bayes, rule-based	Last 100 training days to forecast 1 day	Yes	Yes	No	No	Not mentioned
Werner and Myrray (2004)		Naïve Bayes, SVM	1000 messages vs. the rest	No	No	No	No	Rainbow package
Groth and Muntermann (2011)		Naïve Bayes, k-NN, ANN, SVM	Stratified cross validations	No	No	No	No	Not mentioned

on a preset interval and predict if the market is headed upwards or downwards in the next interval.

This prediction is realized by making a decision on assigning a group of news headlines in an interval to an 'Up' or 'Down' class. In other words, the job of the machine learning component of the system is a binary classification of the textual input.

In this work, the relevant market-predictive text-mining process is considered to have 3 major phases, namely:

1 – Pre-processing, 2 – Machine learning, and 3 – Evaluation as illustrated in Fig. 2.

These are also depicted in the column on the left in Fig. 3.

The system that is proposed in this work maintains the above 3 phases. A more detail view of the main components that are proposed for each of the above phases in the proposed system is to be seen in the column on the right in Fig. 3.

As it can be seen most of the proposed components belong to the Pre-processing phase. Pre-processing in this work is considered to be the component that defines text-mining and differentiates it from data-mining. It covers everything that requires to be done (processed) before a machine readable input is ready. When such input is ready, the rest of the work is very similar to data-mining. Therefore, as the focus of this work is text-mining, its main contribution is made in this phase of the proposed system.

The column on the right in Fig. 3 illustrates all phases of the proposed system. The purpose and function of each of the phases is described in separate sections that follow. The core of the

proposed system is a multi-layer algorithm termed as 'Multi-layer Dimension Reduction with Semantics and Sentiment'. In this work, it is also sometimes referred to as just the 'Multi-layer algorithm' in short. It deals with the core aspects of how features are selected for a feature-vector to be fed into the machine learning algorithm for the purpose of its training (model creation) and prediction execution. Each layer deals with a major text-mining problem that this work aims to tackle. The Multi-layer algorithm has 3 main layers, namely: 1 – Semantic Abstraction, 2 – Sentiment Integration, and 3 – Targeted Feature Reduction.

In this work, the last layer of the Multi-layer algorithm is sometimes referred to by its full name which is 'Synchronous Targeted Feature Reduction'. Each of these layers and their responsibilities and significance is explained in detail in an according section later.

The Multi-layer algorithm allows the system to reach a high accuracy by tackling some of the most fundamental text-mining challenges around in a manner that is enhanced and customized for the context of this work. In short it fulfills 3 main objectives:

1. Reduction of dimensions (features) at every layer in an incremental fashion.
2. Integration of semantics in a manner that reduces semantic redundancy (co-reference) which is usage of multiple words for the same concept or entity.
3. Integration of language sentiment in a way that the amount of emotional-charge or sentiment-load of a word is taken into consideration in weighting a feature. Language is more than words and letters; and the emotion or sentiment that each word carries matters a lot. But as pointed out in the literature review, many market-predictive text-mining systems are ignorant of this fact or are not enhanced enough to accommodate for it.

There are 8 major phases in the system flow illustrated in Fig. 4, namely: 1 – data retrieval, 2 – input-data preparation, 3 – text-tokenization and stop-words removal, 4 – Semantic Abstraction, 5 – sentiment integration, 6 – Synchronous Targeted Feature-Reduction, 7 – model creation and prediction, and 8 – evaluation which is covered in multiple evaluation sections in the experiment section at the end of this work. All of these phases are explained in more detail in the following sections.

4.1. Data retrieval

In the data retrieval phase two main datasets are retrieved. One is the news-headlines dataset for multiple years which can be retrieved from a financial news website like MarketWatch.com or others with a Really Simple Syndication (RSS) function. In order to retrieve headlines for multiple years, Google cache is accessed via the Google RSS reader API. Table 3 lists a number of retrieved news headlines as examples. Note that the news date and time of the publication is also retrieved.

The other dataset is the foreign exchange market (FOREX) historic data for the desired currency-pair which in our case is Euro/USD. There are many sources to obtain this data. In this case it is retrieved by the help of the FXCMMicro Desktop client. Table 4 lists some examples of such data. The retrieved data here is for 2-h time intervals. At each entry 8 pieces of data are present, namely the

Fig. 2. Market-predictive text-mining process phases.

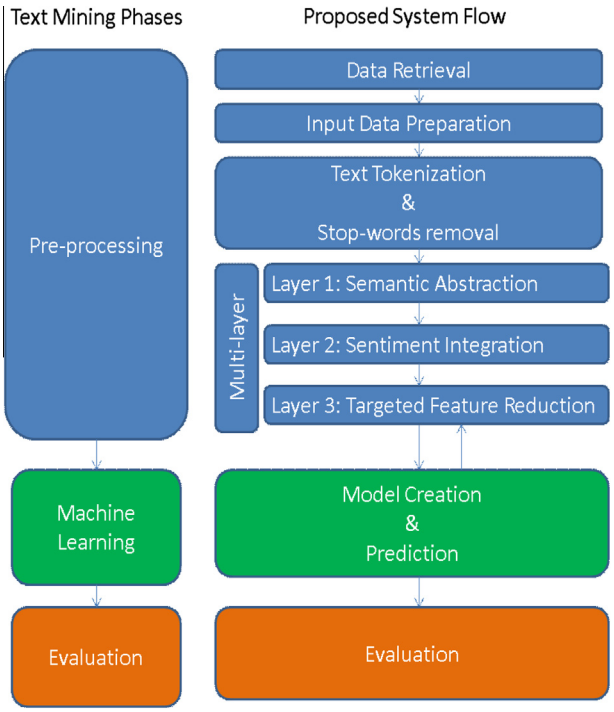

Fig. 3. Overall system flow for the proposed system.

Open, High, Low and Close for both the Ask and the Bid price. Note that the (Close, Bid) of one entry is equal to the (Open, Bid) of the next entry and so forth. Also note that Open marks the beginning of the interval and Close marks the end of an interval. Last but not least, note that 18/04/2008 13:00:00 for example is referring to the interval from 13:00 to 15:00 i.e. an interval is marked by its starting time.

4.2. Input-data preparation (News-Currency Mapping)

In this phase of the system, the input-data for the next phase is prepared by mapping the news-headlines as well as the currency data onto a time series as depicted in Fig. 5.

Fig. 5 is actually composed of 3 separate sections:

The first section that is at the top of the figure is simply a time-line that pin-points the release-time of each of our example news pieces.

The second section that is placed in the middle is a line that indicates how news-headlines are grouped together. All news-headlines that are released between 10:00 and 12:00 for example are grouped together as shown in Table 5. This means the news-headlines are grouped based on their published time; so that the system can observe if there are eventually words in an interval that can cause a reaction in the market afterwards. Then a date-time-stamp is assigned to this news-group which in this case is "18/04/2008 11:00:00". This date-time-stamp is for grouping of the news and hence is called the news-grouping date-time-stamp. Note, the time component "11:00:00" in the stamp is chosen because it is in the middle of 10:00 and 12:00 in our example. The news-grouping date-time-stamp is different from the currency-data date-time-stamp that is explained in the next part.

The third section that is at the bottom of the figure has 2 axes. The *X* axis is the time-line again. Note that on this time-line 2-h intervals are indicated differently. The 2-h interval is not e.g. 10:00 to 12:00 anymore but 11:00 to 13:00. For this interval a new time-stamp exists that looks the same as "18/04/2008 11:00:00" for example but in this new context it is referring to

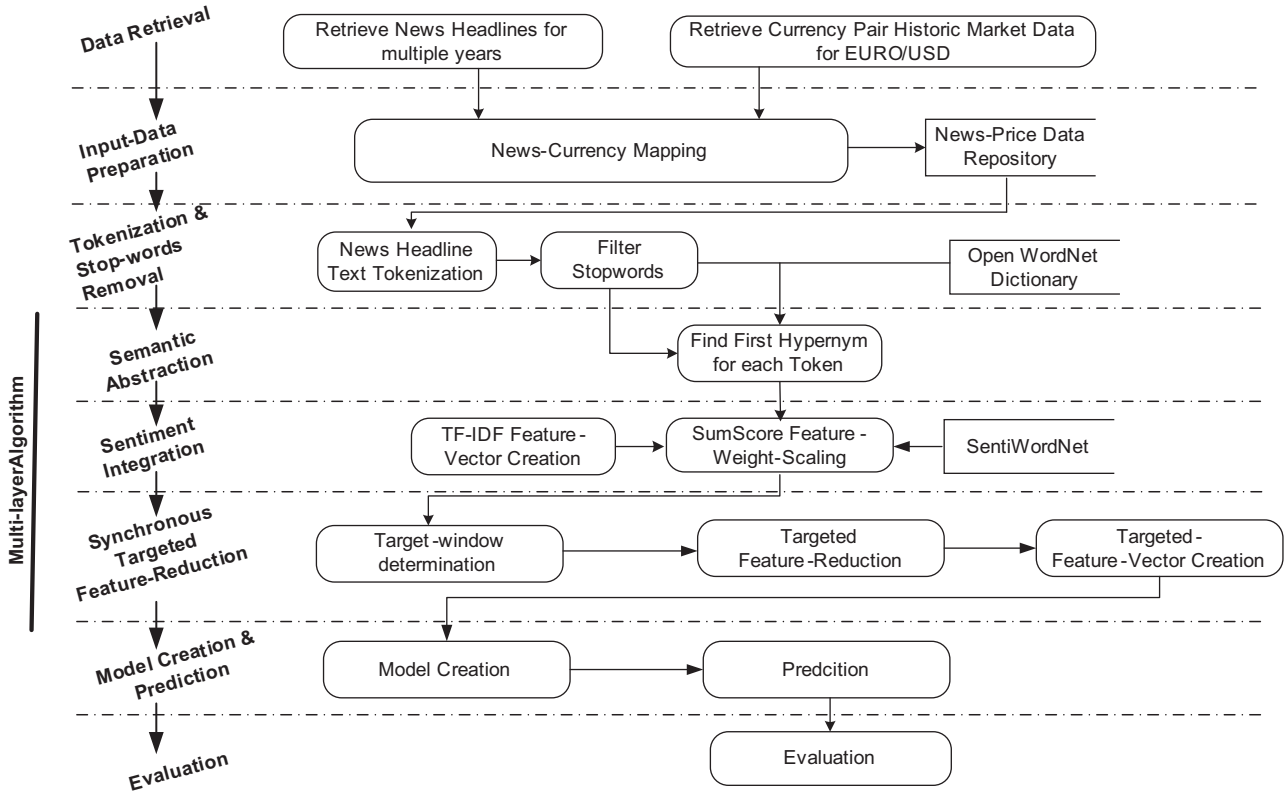

Fig. 4. Detail system design.

Table 3
Example retrieved news headlines.

News headline	News date and time (GMT)
'Strong' demand outside North America propels Caterpillar to 13% quarterly profit growth	18/4/2008 11:37:05
Dow futures stage relief rally after Citigroup results	18/4/2008 11:01:27
Citigroup swings to quarterly loss of $5 billion, revenues fall 48%	18/4/2008 10:57:23

the 11:00 to 13:00 time-interval. This is called the currency-data date-time-stamp and is different from the above news-grouping date-time-stamp in that it is 1-h in the future i.e. the example interval does not end at 12:00 but at 13:00. These intervals are marked in Fig. 5 by vertical lines across the figure.

In this way the system has 2 separate sets of time-stamps: one for the news-grouping and one for the currency-data. They look the same in terms of their composition but are lagged by 1-h. The reason for this setup is to use the news-group to predict the currency-data that is 1-h in the future. It is assumed that a time-lag is required for market impact realization of the news, having taken into consideration suggestions made by the past research like the work of Reboredo, Rivera-Castro, Miranda, and García-Rubio (2013). Past research has indicated sixty minutes to be looked at as a reasonable market convergence time to efficiency (Chordia, Goyal, Lehmann, & Saar, 2013; Chordia, Roll, & Subrahmanyam, 2005). Market convergence refers to the period during which a market reacts to the information that is made available and becomes efficient by reflecting it fully. Furthermore, this setup provides a 1-h margin before the news-release time as well, which is just to ensure that the news is really exactly between two points in time as different news sources may have somewhat different release-times and margins around news release-time increase the certainty of news-release occurrence in the desired time-window. What exactly is predicted is explained in the following.

Next, the *Y* axis is the price-point for Euro/USD (Close, Bid) at that point in time. Note, that this price-point is available at the turning point of each currency-data date-time-stamped interval. For example, the Euro/USD (Close, Bid) price-point for "18/04/2008 13:00:00" is 1.58016, the Euro/USD (Close, Bid) price-point for "18/04/2008 11:00:00" is 1.57549 and the Euro/USD (Close, Bid) price-point for "18/04/2008 09:00:00" is 1.57301 as shown in Fig. 5.

The next important piece of information that is illustrated in this part of Fig. 5 is the calculation of a label for each news-group. The ultimate objective of this system is to predict the direction of

Table 4
Example of retrieved currency-pair data.

Date and time	EUR/USD (Open, Ask)	EUR/USD (High, Ask)	EUR/USD (Low, Ask)	EUR/USD (Close, Ask)	EUR/USD (Open, Bid)	EUR/USD (High, Bid)	EUR/USD (Low, Bid)	EUR/USD (Close, Bid)
18/04/2008 15:00:00	1.58041	1.58159	1.57980	1.58103	1.58016	1.58134	1.57955	1.58082
18/04/2008 13:00:00	1.57571	1.58140	1.57422	1.58041	1.57549	1.58115	1.57397	1.58016
18/04/2008 11:00:00	1.57329	1.57581	1.57221	1.57571	1.57301	1.57556	1.57196	1.57549
18/04/2008 09:00:00	1.57404	1.57664	1.57141	1.57329	1.57382	1.57639	1.57116	1.57301
18/04/2008 07:00:00	1.58441	1.58557	1.57129	1.57404	1.58412	1.58534	1.57104	1.57382

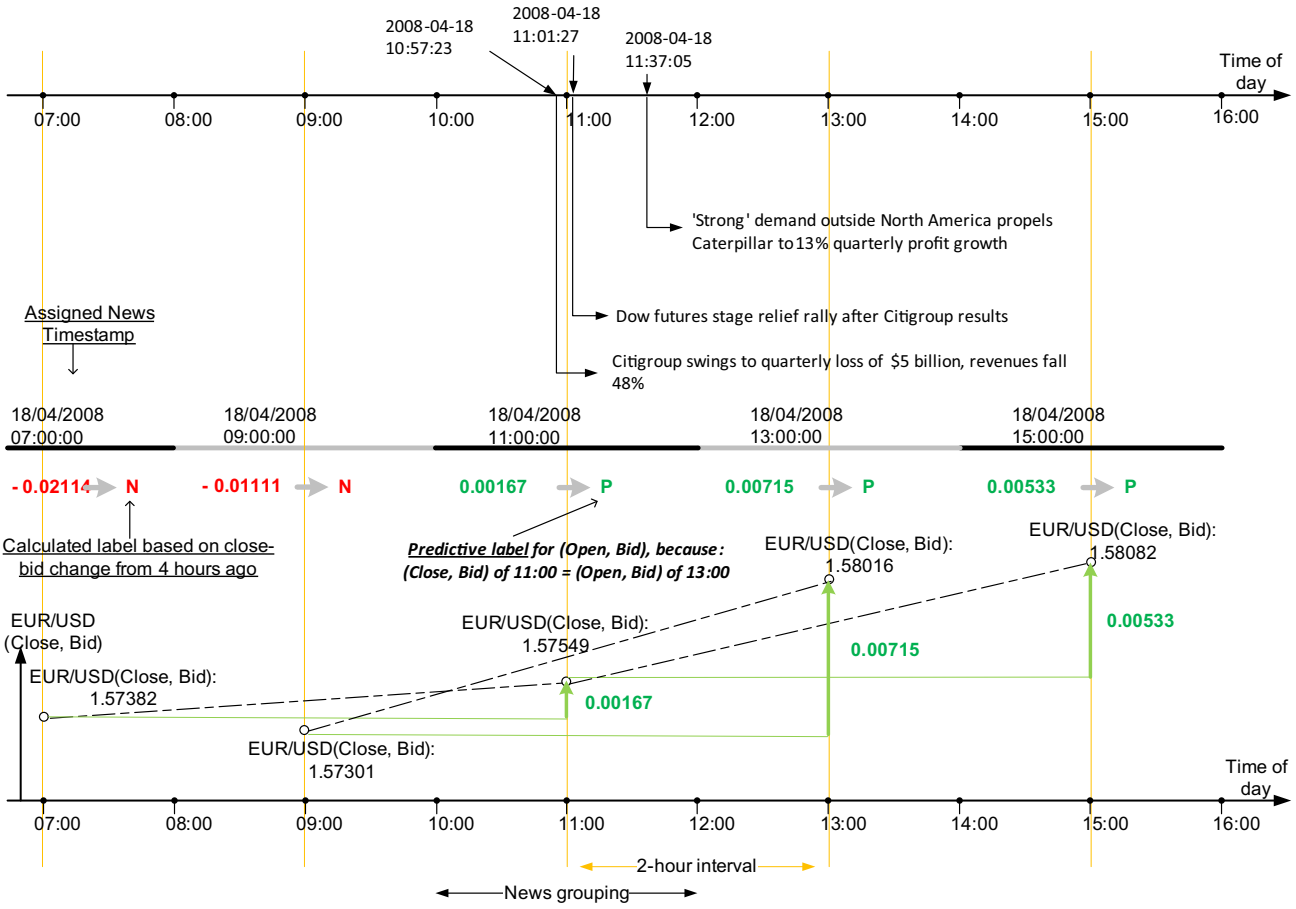

Fig. 5. News-Currency Mapping.

Table 5
Example of news-grouping.

News-grouping date-time-stamp	All news-headlines in the same news-group	Label
18/04/2008 11:00:00	'Strong' demand outside North America propels Caterpillar to 13% quarterly profit growth Dow futures stage relief rally after Citigroup results Citigroup swings to quarterly loss of $5 billion, revenues fall 48%	P

the price-movement based on a given set of news-headlines. In other words, each group of headlines must be associated with a label that indicates the upward or downward movement of the price according to those headlines i.e. that news-group. As mentioned before, the piece of price data that the system is working with is Euro/USD (Close, Bid). But Euro/USD (Close, Bid) alone is just an indication of a price-point and not a price direction. Hence, a new value is created in the system to indicate the change in Euro/USD (Close, Bid) as below:

$$C_i = Cb_i - Cb_{i-2} \tag{1}$$

$$Label: \begin{cases} IF \quad (C_i > 0) : P \\ IF \quad (C_i <= 0) : N \end{cases} \tag{2}$$

where C_i is the change in Euro/USD (Close, Bid) of the interval-turning-point i, Cb_i is the (Close, Bid) at that point and the Cb_{i-2} is the (Close, Bid) at 2 intervals ago or 4 h ago (Formula (1)). The 4-h interval is composed of 2 times 2-h market-intervals; and the 2-h news-interval is aligned at its center so that there is a 1-h margin at each end (Fig. 5). As mentioned before, the margins help ensure the news release-time is between the two points in time at the ends of the 4 h-interval. Furthermore, the 1-h margin at the right hand-side constitutes the predictive timeline which is 1-h at minimum.

If C_i is greater than 0 then label "*P*" for Positive is chosen and otherwise label "*N*" for Negative is chosen (Formula (2)).

For example, as mentioned at point 13:00 on the time-line the Euro/USD (Close, Bid) is 1.58016 and it is 1.57301 at point 09:00 which is 4 h before it. The change here would be $C_{13:00} = 0.00715$ which is greater than 0 which is construed as label "*P*".

Note that point 13:00 on the time-line is at the end of the interval between 11:00 and 13:00. This interval as explained below is currency-data date-time-stamped as "18/04/2008 11:00:00". At the same time the same date-time-stamp value is used for a news-group among the news-grouping date-time-stamps. However, this news-group is composed of the news-headlines released between 10:00 and 12:00 as explained before.

This association of a group of headlines in a news-group with a label via the above date-time-stamping system is called the News-Currency Mapping, which is the ultimate illustration objective in Fig. 5. It is at the core of the labeling mechanism in this system. The above described process flow of News-Currency Mapping is also summarized in Fig. 6.

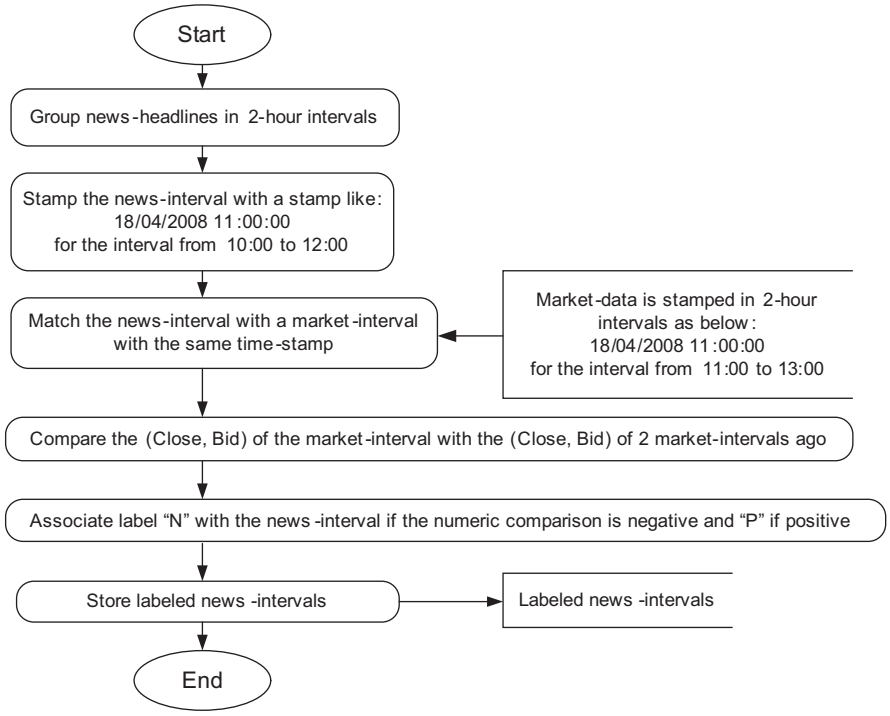

Fig. 6. News-Currency Mapping process flow.

There is one last point to note while thinking about this mechanism: In order to refer to the point 13:00 on the time-line in terms of the currency-data in the above, the interval before it i.e. 11:00 to 13:00 is selected. The reason is that a (Close, Bid) is referring to the Bid-value at the end of an interval. However, as described before the (Close, Bid) at 11:00 to 13:00 is equal to the (Open, Bid) at 13:00 to 15:00. Hence, in other words, it is exchangeable in this work to state that the system is predicting the (Close, Bid) of date-time-stamp "18/04/2008 11:00:00" or the (Open, Bid) of "18/04/2008 13:00:00". Furthermore, this prediction is made by inputting the news-headlines released from 10:00 to 12:00 into the learned-model, the creation of which is described in a separate section.

At the end of this phase, each news-group is associated with a calculated label per above description and stored in a repository. Table 5, demonstrates one entry of that repository which is based on the example headlines in Fig. 5.

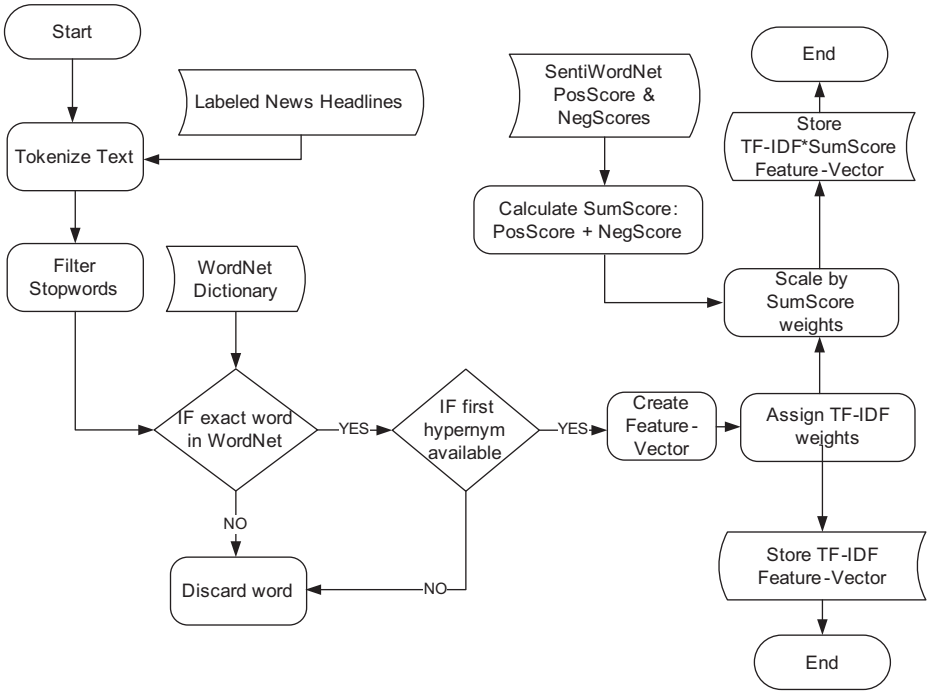

Fig. 7. Complete feature-vector preparation flow.

From this point on, the activities are focused on creation of a feature-matrix. First, a feature-vector is prepared for each record in the news-grouping repository. That means in the case of the record in Table 5 for example, the text of the news-headlines, which is shown in the second column in Table 5, is boiled down into a number of features. This process is depicted in Fig. 7, followed by a step-by-step demonstration based on the above example record in each of the following sections.

4.3. Text tokenization and stop-word removal

As seen in Fig. 7, at first the text of the grouped news-headlines is tokenized and the stop-words are removed. That means for the example record the following transformation takes place.

Original text:

'Strong' demand outside North America propels Caterpillar to 13% quarterly profit growth Dow futures stage relief rally after Citigroup results Citigroup swings to quarterly loss of $5 billion, revenues fall 48%

After tokenization and stop-word removal:

*Strong *demand *North America propels Caterpillar **quarterly profit growth Dow futures stage relief rally *Citigroup results *swings **loss **billion *revenues fall *

A '*' is placed in the above section whenever a word or a punctuation mark is removed to indicate the place of removal. Note that at the end of this transformation there are no punctuation marks left as well as no repeated words, no numbers like '13%', '$5', '48%' and no stop-words which include words like 'outside', 'to', 'after', 'of' in the above example.

4.4. Semantic Abstraction via Heuristic-Hypernym Feature-Selection

Semantic Abstraction is the name of the first layer in the multi-layer algorithm and the novel proposed technique that is used in it is termed as "Heuristic-Hypernym Feature-Selection". This is illustrated in Fig. 8.

Semantic Abstraction is a concept that is defined in this work with the following logical analysis. The most common method in the past research for the Pre-processing phase has been the so called 'Bag of Words'. In this method the news text is represented as a group of words. Each of the words is regarded as a feature. This method can be improved by creating a layer of abstraction (Schumaker & Chen, 2009). Abstraction means having every word associated with a word of a higher order or generality i.e. a word that acts as a super-category for all subordinate words. Abstraction can be perceived to have two main advantages:

1. It simplifies the feature space by reducing the number of words that are used as features.
2. It may help make similar conclusions from similar words by referring to them in the same way i.e. by referring to their category name.

Heuristic-Hypernym Feature-Selection is a novel semantic abstraction technique that is devised in this work to extract features out of the above remaining words after stop-word removal. There are two aspects to this technique: 1 – the use of hypernyms and 2 – the heuristic selection of them that is proposed in this technique.

Hypernyms have proven to be effective in increasing classification accuracy in other areas before in the literature (Jeong & Myaeng, 2013). How hypernyms are selected in this technique follows.

As illustrated in the algorithm in Fig. 7, each remaining token is looked up in WordNet Dictionary (Miller, 1995). If the exact word exists in the dictionary and it has a hypernym it is replaced with it. If there is more than 1 hypernym then the first listed hypernym is chosen. If the exact word is not available in WordNet dictionary or there are no hypernyms available then the word is simply discarded from the potential feature list.

Table 6 lists the looked up hypernyms for the example at hand.

As shown in Table 6 some words are not available in WordNet dictionary like 'Citigroup', 'Dow' and some words do not have any hypenyms like 'America' or 'Strong' and hence are marked as Not Available(N/A) in the table. Furthermore, words like 'futures', 'propels', 'results', 'revenues', 'swings' that end with an 's' are also discarded if the form with the 's' does not possess an entry in the dictionary. The same goes for words ending in 'ed', 'ing', etc. This is the heuristic aspect of this technique. The logic behind this heuristic mechanism, whereby only exact

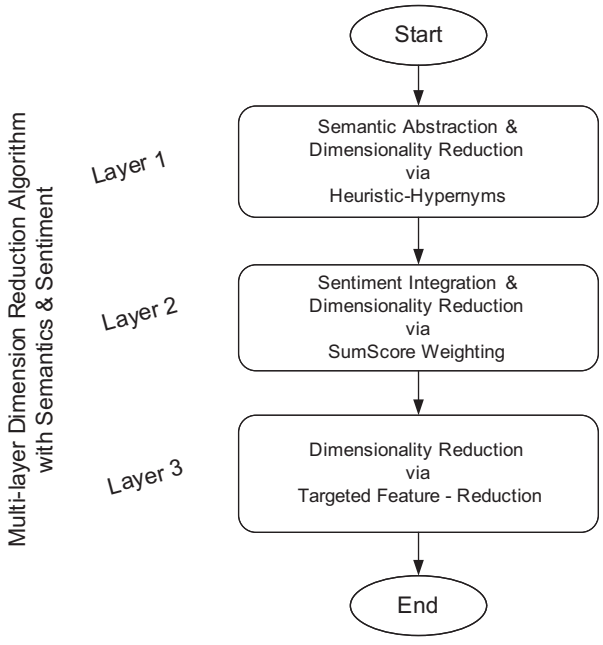

Fig. 8. Multi-layer Dimension Reduction Algorithm.

Table 6
Example for heuristic-hypernym feature-selection.

Exact words after stop-word removal	Hypernyms
America	N/A
Caterpillar	Larva
Citigroup	N/A
Dow	N/A
North	Cardinal compass point
Strong	N/A
Billion	Large integer
Demand	Request
Fall	Season
Futures	N/A
Growth	Organic process
Loss	Transferred property
Profit	Income
Propels	N/A
Quarterly	Series
Rally	Gathering
Relief	Comfort
Results	N/A
Revenues	N/A
Stage	Time period
Swings	N/A

dictionary entries are kept, is described in an according evaluation section that follows in the experiments section of this paper and discusses its impact.

Additionally, sometimes the hypernyms chosen do not make any sense in terms of meaning for example in the above 'larva' is the hypernym that is determined by the system for the word 'Caterpillar', however, in this context the word 'Caterpillar' is merely a company name and has nothing to do with the biological hypernym 'larva' in terms of meaning. It is noted that this may logically introduce some noise or at least an unreliable association, however, on the whole such a case of absolute irrelevance, as seen in the above example set, lies in the minority. Moreover, in this system the contextual meaning is less significant than a statistical place-holder i.e. the word 'larva' may very well constitute a somewhat unique representation of the word 'Caterpillar' in the context of financial news and thereby still allow the system to determine statistical significance, if any at all, towards the ultimate goal of the system that is of a statistical-pattern-recognition nature.

Another example for such "statistical place-holder" assumption comes into play in the case of a polysemous word like "fall" which could mean "drop" or be a "season". And the system as shown in Table 6 chooses the hypernym "season". If we assume that a word like "fall" in context of the financial markets predominantly means "drop"; and the current system replaces the word with "season" which is the hypernym for the other meaning of "fall"; then "season" can play the role of a statistical place-holder. In other words, the impact of the word "fall" in the system is assumed to be captured by the word "season" which is literally not the correct hypernym meaning-wise but because every instance of "fall" is replaced with it, it can be assumed to play the same statistical role. However, it should be conceded to, that an improvement on this aspect of the system may lead to even better results. This problem is referred to in the literature as "disambiguation" and is a challenging research topic. It is really not easy to choose the correct meaning for a word based on its context and therefore the proposed solution despite its imperfection is reasonable as it is in essence providing a way around "disambiguation" with the above described "place-holder" assumption. However, this aspect is yet to be enhanced in the future.

Once the choice of hypernyms per record is completed, the features of the feature-vector are actually determined. Next, they are weighted by a novel metric explained in the next section.

4.5. Sentiment integration via TF-IDF*SumScore weighting

Once features are determined in the feature space, it is crucial to realize that they have different levels of impact in terms of the sentiment that they entail. The Sentiment Integration Layer, as illustrated in Fig. 8, proposes a technique by the name of TF-IDF*-SumScore Weighting or SumScore Weighting in short that captures the sentiment in a scaled manner by TF-IDF. The scaling by TF-IDF is useful because the frequency of appearance of a word in a document and the corpus are important for a weighting scheme to remain relevant to a context. Both TF-IDF and SumScore are explained further in the next 2 subsections.

4.5.1. TF-IDF weighting

Term Frequency–Inverse Document Frequency (TF-IDF) is a standard numerical statistic that reflects how important a word is to a document in a collection or corpus. It is often used as a weighting factor in information retrieval and text mining. The TF-IDF value increases proportionally to the number of times a word appears in the document, but is offset by the frequency of the word in the corpus, which helps to control for the fact that some words are generally more common than others.

Each feature in the feature-vector of each record is assigned a TF-IDF value at this stage.

4.5.2. SentiWordNet SumScore weighting

The TF-IDF weighted features are scaled by another weighting value next, namely, the SentiWordNet SumScore that is a new score defined as a part of this work.

SentiWordNet is a dictionary of sentiment values (Baccianella & Sebastiani, 2010) that contains a Positivity Score (PosScore) and/or a Negativity Score (NegScore) between 0 and 1 for each WordNet entry i.e. synset. A further score by the name of Objectivity Score (ObjScore) is also defined as follows:

$$O_i = 1 - (P_i + N_i) \tag{3}$$

where O_i is the ObjScore of word i which is the result of subtracting the sum of P_i or PosScore of word i and N_i or NegScore of word i from 1.

However, for this system a novel measure is defined by the name of SumScore with the below definition:

$$S_i = P_i + N_i \tag{4}$$

where S_i is the SumScore of word i which is the result of summation of P_i or PosScore of word i and N_i or NegScore of word i. In other words SumScore of a word can be calculated from its ObjScore as below:

$$S_i = 1 - O_i \tag{5}$$

where O_i is the ObjScore of word i. This is important to realize that SumScore is inclusive of both positive (PosScore) and negative (NegScore) indications of emotion. It is measuring total existence of sentiment. Whereas ObjScore is intended to indicate objectivity or lack of any sentiment or emotional charge, but as it turns out it is more effective to use a measure that is indicative of existence of sentiment i.e. SumScore rather than a measure that is indicative of its absence i.e. ObjScore. And rightly so as absence or non-existence of a phenomenon is hardly measurable from a logical stand point. What is measured is existence. In the same way that in measurement of temperature what is measured is the existence of heat and not lack thereof i.e. coldness. Coldness is merely the absence of heat; hence it does not have its own existence and therefore cannot be measured.

Furthermore, it is important to note that measuring one direction of emotions alone i.e. positive or negative does not make much sense either as a market-participant may feel positive or negative about any given market direction. However, it does make sense to anticipate market activity when the total amount of emotions regardless of direction i.e. the value of SumScore tends to change.

In an according experiment section later, it is demonstrated that the proposed SumScore has significant positive impact on the accuracy of the prediction results. It outperforms ObjScore as well as NegScore and PosScore per se. SumScore definition and usage design is one of the contributions of this work.

At this stage, as seen in the flowchart in Fig. 7, a TF-IDF*SumScore weighted feature-vector is ready and stored for each record. The TF-IDF weighted feature-vector is also stored separately for an accessory purpose that is explained in the next section.

4.6. Synchronous Targeted Feature-Reduction

What is termed in this paper as the Synchronous Targeted Feature-Reduction is important to explain as it is one of the major contributions of this work. It is illustrated in Figs. 4 and 8. In short, the

common approach for training and model creation in the literature (Hagenau et al., 2013; Schumaker et al., 2012; Yu et al., 2013) is some variation of the below steps:

Step 1: Taking a feature-matrix as input.
Step 2: Usually conducting no further feature-reduction or sometimes reducing the features to a top random number, say, 100 or 200.
Step 3: Building a model based on part of the records called the training-set.
Step 4: Using the above built model to predict other records.

There are at least two main problems with the above approach:
Problem 1 – If there is no effective feature-reduction, there are too many features available in the feature-vector which leads to the curse of high dimensionality (Pestov, 2013).
Problem 2 – If there is a feature-reduction method, for example choosing the top features according to a criterion, then the reduction is somewhat random as it is in no special way optimized for the record(s) to be predicted.
"Synchronous Targeted Feature-Reduction" solves the above two issues effectively and increases the results significantly as later shown in the experimentation section of this paper. It proposes the below flow of steps:

Step 1: Taking a feature-matrix as input.
Step 2: Taking a single record whose label is to be predicted.
Step 3: Reducing all the features in the feature-vector to only those with a value in this record and create a new feature-vector thereof. In other words:
SELECT the columns of the initial feature-vector
WHERE the value of the features in the targeted-record (or the record whose value is to be predicted) is non-zero and
CREATE a new table thereof.

The word Synchronous in the name is referring to this synchronous feature-vector table creation; which is happening as the system is being run on the records to be predicted.

Step 4: Building a model based on all the other records available for training.
Step 5: Running the single record chosen in step 2 through the created model in step 4 in order to predict its label.

In the above, steps 1 to 3 summarize the proposed feature-reduction method based on the targeted-record for prediction which is termed Synchronous Targeted Feature-Reduction (STFR). STFR basically reduces all the features to those only that are available in the record that is targeted for prediction.

4.7. Model creation and prediction

In step 4 and 5 a model is created and used to predict the label for the targeted record accordingly. It is important to note that STFR presides on absolute optimization of feature-reduction by reducing the features to the minimum that is needed for the one prediction task at hand and creates a new model per prediction. In other words it creates new models synchronously and just in time as the prediction needs to happen. In the experiments section execution-time has been observed and the net-advantage of the proposed method is far greater than the extra few seconds needed for model-creation per prediction. The total of all steps from 1 to 5 summarize the entire label prediction activity that includes STFR; and can be termed as Synchronous Targeted Label Prediction or STLP. Pseudo-code 1 details STLP further as a method.

Pseudo-code 1
Pseudo-code for Synchronous Targeted Label Prediction (STLP)

```
Method: Synchronous Targeted Label Prediction (STLP)
Input: TF-IDF*SumScore Feature Matrix (M)
Intermediary Output: Reduced Matrix Based on K (R)
Overall Output: Predicted Label (LABEL)
        {INITIALIZATION}
M_i,j = TF-IDF*SumScore Feature Matrix upto K
K = Index of last record of M_i,j i.e. Prediction Target
    {RUN Synchronous Targeted Feature Reduction (STFR)}
T = M_K,J (Targeted/Kth record)
R_i,j = M_i,j
FOR T_i FROM i =1 TO i = J
  IF Ti == 0
    DROP COLUMN Column_i FROM R_K,J
  END-IF
END-FOR
      {GENERATE Synchronous Model & RUN it}
TRAINING = SELECT * FROM R WHERE ID != K
TARGET = SELECT * FROM R WHERE ID == K
MODEL = GENERATE_MODEL(TRAINING)
LABEL= RUN_MODEL(TARGET)
RETURN LABEL
```

This approach reduces the features effectively to an absolute minimum. Produces very good results in the context of this work and makes a lot of logical sense in terms of efficiency and accuracy in feature-reduction which is experimentally proven in this work and described in a later section accordingly.

As the primary input feature-vector in this work is the TF-IDF*-SumScore feature-vector, there are some rare instances whereby a selected record to be predicted is of no feature with a non-zero TF-IDF*SumScore value which is caused by the SumScore being zero for too many features in that record. Hence, in such a situation no features can be determined via the TF-IDF*SumScore values as there are none. To address this predicament, if such a record is encountered the system will dismiss the TF-IDF*SumScore feature-vector and use the TF-IDF feature vector as input for feature-reduction and model creation for that record instead. This is illustrated in the flow chart in Fig. 9 and this decision is made at the decision point where this question is asked: "*Are there any non-zero attributes?*"

Note that the input feature-vector that is used in this system is chronologically ordered and hence, the Kth record which is the record to be predicted is always the last record and the $K-1$ records before it are the records used for training.

Last but not least, the system is using a standard Support Vector Machine (SVM) algorithm (Cortes & Vapnik, 1995) for model creation and label prediction.

Support Vector Machine (SVM) is the most common machine learning algorithm in the literature for similar classification problems (Fung, Yu, & Lam, 2002; Mittermayer, 2004; Soni, van Eck, and Kaymak, 2007). SVM is a supervised learning model for classification. It generally outperforms other models like neural networks in financial time series prediction (Kim, 2003; Tay & Cao, 2001). This is attributed to its ability for structural risk minimization, where multiple local minima can be avoided (Premanode & Toumazou, 2013). Moreover, the computational complexity of SVM does not depend on the dimensionality of the input space. It is a binary classifier which means the input is classified in one of two categories as output. With an SVM model the representation of each news-group is mapped as a point in a feature-space where the examples of one category are separated by a line or a curve with as much margin as possible from the other category.

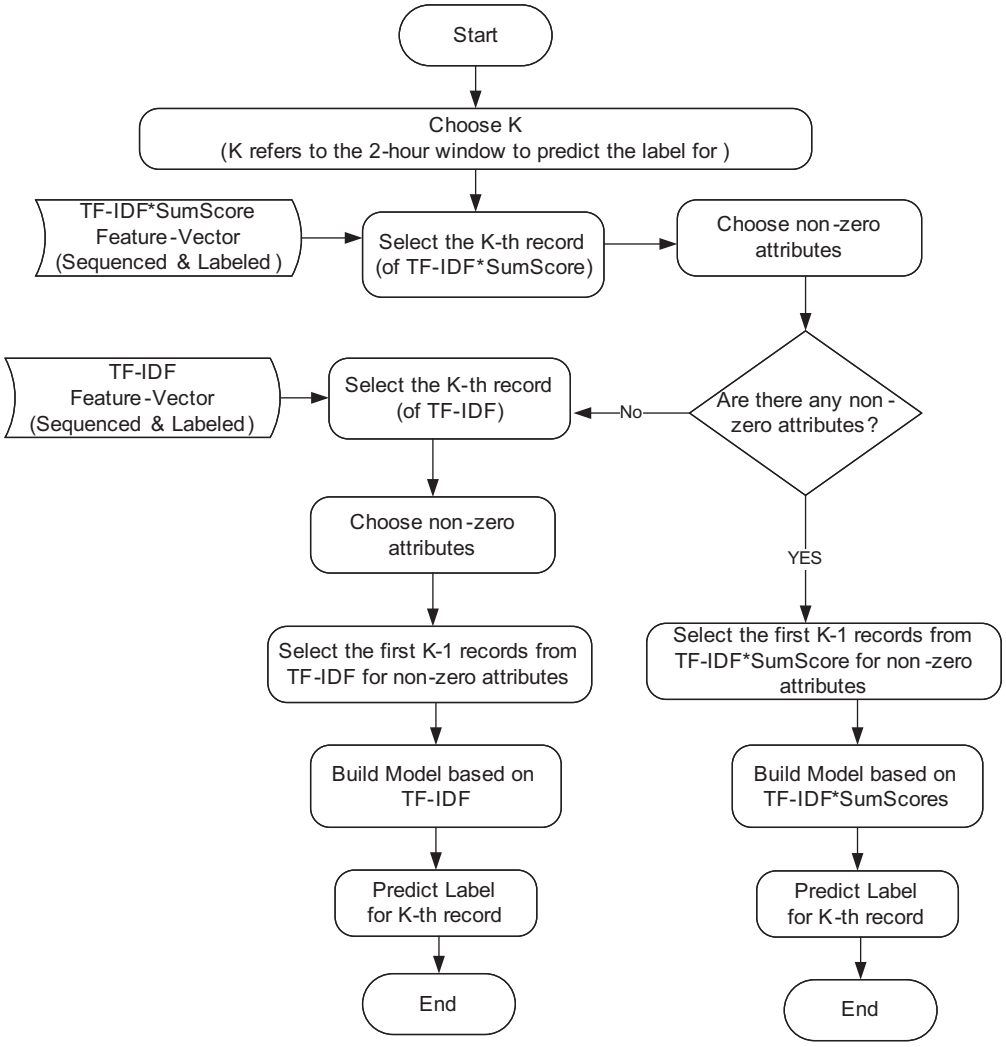

Fig. 9. Flow of Synchronous Targeted Feature-Reduction, model-creation and prediction.

5. Experimental results and evaluation

When the proposed techniques are available in the system and the experiment is run an accuracy of 83.33% is achieved on the above described testing sample. However, if all the 3 proposed techniques are removed from the system and the experiment is run the resulted accuracy is at 50.00% which coincidently is what is presumed to be the accuracy of classification by chance when 2 classes are available. In the following each of the techniques is removed from the system individually and its impact on accuracy is observed. In all cases a significant improvement in accuracy is noticed when the technique is present in the system. And this is our main evaluation approach for the results presented in this work. However, just to point out the accuracy range in other works in the literature for the sake of curious comparison: The accuracy in majority of the cases is reported in the range of 50% to 70%, while commonly arguing for better than chance results which is estimated at 50 percent (Butler & Kešelj, 2009; Li, 2010; Mahajan et al., 2008; Schumaker & Chen, 2009; Schumaker et al., 2012; Zhai, Hsu, and Halgamuge, 2007). In other words, the reported systems in the literature require different inputs and therefore are not compared with each other in the same experimental settings but most of them report their results because they have achieved better than chance results. Furthermore, generally results above 55% have been considered categorically report-worthy in similar contexts in the literature from a practical perspective (Garcke, Gerstner, & Griebel, 2013).

5.1. Dataset and execution-time details

Before describing the experiments, here is a brief account about the used datasets in them for training and testing. The total dataset spans from 2008 to 2011 and has 6906 records. The records are chronologically in order. The proposed system makes a prediction per record at a time. Hence, in order to make a prediction for the last record all records before it are used for training and for the second-last all the ones before that and so forth. Each record is associated with a 2-h time interval so that if the system is left to run for 24 h 12 records are predicted. 24 h is the time-span covered by 12 records and not the running time needed to make the relevant predictions. The total execution time for 12 predictions on a PC with total of 4 GB RAM and 3.10 GHz Intel CPU is about 2 min or 10sec per prediction. As a 24 h-coverage seems reasonable to test a system that is predicting the next hour the below tests are conducted on the last 12 records of the dataset.

5.2. Evaluation of Heuristic-Hypernym Feature-Selection

As described before in the section on Heuristic-Hypernym Feature-Selection, the mechanism that is proposed is keeping only those words with an exact WordNet entry.

Some words like 'futures', 'propels', 'results', 'revenues', 'swings' are not available in their exact form as a WordNet entry but are available as their stemmed forms which is the form without the 's' at the end in the above examples. However, these words are being discarded in this system based on the assumption that words that end in an 's' are either plural nouns or third person singular verbs. In the literature, it is indicated that in general parts of speech of noun and verb carry less sentimental or subjective value than adjectives and adverbs (Esuli & Sebastiani, 2006). This makes intuitive sense as sentimental content is mostly captured and expressed in more emotionally descriptive words namely adjectives and adverbs. The system is tested after elimination of all words of the kind 'noun' or 'verb' or both, however, best results are shown when only words like the above that are ending in an 's' or other non-exact matches like words ending in an 'ed' or 'ing' are eliminated. This makes logical sense, as blanket removal of words based on a part-of-speech may result in removal of some words unnecessarily. Words can have more than one part of speech for example the third form of a verb can be referred to as an adjective as well. For instance, the words 'weakened' or 'diminished' are adjectives but once stemmed they go to their verb form of 'weaken' and 'diminish' and can be eliminated if all verbs are removed as a part of speech category. However, they both can remain in the proposed mechanism as both 'weakened' and 'diminished' have their own entries in WordNet. Furthermore, in case of a blanket elimination of nouns, words like 'growth' and 'loss' are also discarded which are potentially valuable for the system. Hence, logically and experimentally blanket removals are not useful. Experiments show that the choice of a word by the system based on the existence of an exact match in WordNet reveals best results. In order to test these assumptions, the feature-vector is once more created but this time the words are stemmed first and then their hypernyms are looked up in WordNet. Note that in the proposed system the words are not stemmed and the exact words are passed to be looked up in WordNet. The results are shown in Table 7 and are significantly lower when the heuristic is not used.

In this experiment, the number of produced hyponyms based on stems is 2318 which is higher than the number of produced hypernyms based on exact words that is 2149. Once scaled by TF-IDF*SumScore the numbers of produced features for stems and exact words become 495 and 435 respectively as many of them are eliminated by a SumScore of 0. This indicates that a difference of 60 features is causing the accuracy to go from 83.33% to 33.33%.

A drop in accuracy from 83.33% to 33.33% indicates that when the heuristic approach is not utilized for elimination of words, more words are replaced with their hypernyms; these new words introduce noise and therefore the results turn out to be poor. This indicates that the additional hypernyms are less valuable for the predictive purpose. This is an interesting discovery that gives value to the availability of exact matches in WordNet and the proposed heuristic approach that is designed to take advantage of it.

Note that in addition to Accuracy, Precision and Recall are also reported for each of the classes (N and P) in Table 7 as well as the next evaluation tables (Tables 8–11). They provide more insights into the relevance of results and are defined as below:

Table 7
Prediction results using hypernyms of stems instead of hypernyms of exact words.

	Precision (N) (%)	Precision (P) (%)	Recall (N) (%)	Recall (P) (%)	Accuracy (%)
Hypernyms (TF-IDF*SumScore)	88.89	66.67	88.89	66.67	83.33
Hypernyms of Stems (TF-IDF*SumScore)	60.00	14.29	33.33	33.33	33.33

Precision is the fraction of predicted instances that are true for a class *C*. High precision means that an algorithm returns substantially more true results than false for a class *C*:

$$Precision(C) = \frac{True(C) \cap Predicted(C)}{Predicted(C)} \quad (6)$$

Recall is the fraction of true instances of a class *C* that are predicted. High recall means that an algorithm returns most of the true results for a class *C*:

$$Recall(C) = \frac{True(C) \cap Predicted(C)}{True(C)} \quad (7)$$

As seen in Table 7, in addition to the accuracy, both precision and recall of the two classes (N and P) are also significantly higher when the proposed Heuristic approach is in place.

To be exact, once the Heuristic approach is in place, Precision and Recall for class N are both 88.89% and for class P are both 66.67%; while they drop to 60% and 33.33% for class N and 14.29% and 33.33% for class P when the heuristic approach is not used.

This indicates that when each of the classes (N and P) are looked at individually, they both still perform better in the case where the heuristic approach is utilized. Moreover, they perform better at two levels: Firstly, in terms of the number of the correctly predicted cases from a class out of all the *available* cases in that classes (Recall). Secondly, in terms of the number of the correctly predicted cases from a class out of all of the *predicted* cases from that class (Precision). This shows that the drop in accuracy is not caused by a drop in one class only and both of the classes experience an improvement via the proposed heuristic approach.

This test proves the above assumptions and demonstrates that the existence of an exact entry for a word in WordNet is indicative of some value. Note that words that have similar endings but are not in possession of an entry in WordNet, are discarded and this heuristic seems to strike a meaningful balance in the logic of feature selection and reduction.

5.3. Evaluation of TF-IDF*SumScore

In order to determine the effectiveness of the usage of TF-IDF*SumScore weighting to scale the feature-vector in comparison with other options a number of tests are conducted on the above dataset. Table 8 illustrates the effectiveness of SumScores against other weighting possibilities for hypernyms as well as no weighting at all. TF-IDF*SumScore is clearly leading at 83.33% which is significantly high, followed by TF-IDF*NegScore at 75% and then 58.33% for both TF-IDF*PosScore and TF-IDF*ObjScore as well as the TF-IDF alone, which indicates the lack of any positive impact by PosScore and ObjScore compared to their complete absence. It also indicates the potential of NegScore. It is indicated in the literature that negative words in stories about fundamentals predict earnings and returns more effectively than negative words in other stories (Tetlock et al., 2008) and the effectiveness of negative words is observed in this experiment as well. However, the determination of Objectivity seems to be of very little value in this context.

To see if TF-IDF itself is of any impact on the results, the above tests are run again but this time on a binary feature-vector of the hypernyms instead of a TF-IDF weighted one and the results are listed in Table 9. In this context, as well, SumScore and NegScore do best but this time equally. The main objective of this set of tests is to determine if the inclusion of TF-IDF in the weighting is bringing any value to the table and the clear answer is yes as indicated by the results which are significantly lower compared to their TF-IDF counterparts in Table 8.

Table 8
SumScore compared against other scaling measures on TF-IDF represented hypernyms.

	Precision (N) (%)	Precision (P) (%)	Recall (N) (%)	Recall (P) (%)	Accuracy (%)
TF-IDF	75.00	25.00	66.67	33.33	58.33
TF-IDF*SumScore	88.89	66.67	88.89	66.67	83.33
TF-IDF*ObjScore	75.00	25.00	66.67	33.33	58.33
TF-IDF*PosScore	83.33	33.33	55.56	66.67	58.33
TF-IDF*NegScore	87.50	50.00	77.78	66.67	75.00

Table 9
SumScore compared against other scaling measures on binary represented hypernyms.

	Precision (N) (%)	Precision (P) (%)	Recall (N) (%)	Recall (P) (%)	Accuracy (%)
Binary	80.00	28.57	44.44	66.67	50.00
Binary*SumScore	77.78	33.33	77.78	33.33	66.67
Binary*ObjScore	66.67	16.67	44.44	33.33	41.67
Binary*PosScore	71.43	20.00	55.56	33.33	50.00
Binary*NegScore	85.71	40.00	66.67	66.67	66.67

Table 10
Evaluation of system without Synchronous Targeted Feature-Reduction.

	Precision (N) (%)	Precision (P) (%)	Recall (N) (%)	Recall (P) (%)	Accuracy (%)
TF-IDF*SumScores	75.00	25.00	33.33	66.67	41.67

Table 11
Results for different sizes for the testing dataset.

Testing dataset size	Precision (N) (%)	Precision (P) (%)	Recall (N) (%)	Recall (P) (%)	Accuracy (%)
6	100.00	0.00	83.33	0.00	83.33
12	88.89	66.67	88.89	66.67	83.33
18	61.54	80.00	88.89	44.44	66.67
24	52.94	85.71	90.00	42.86	62.50
30	47.62	77.78	83.33	38.89	56.67
36	45.83	66.67	73.33	38.10	52.78
42	46.15	75.00	75.00	46.15	57.14
48	54.84	70.59	77.27	46.15	60.42
54	50.00	65.00	70.83	43.33	55.56

5.4. Evaluation of Synchronous Targeted Feature-Reduction

In order to test the effectiveness of the proposed Synchronous Targeted Feature-Reduction technique, it is eliminated from the system and an experiment is conducted. In this experiment all available features are used to create the model and not only the ones relevant to the record targeted for prediction. This test results in a significantly lower accuracy as shown in Table 10.

5.5. Evaluation of machine learning algorithms

The same experiment is also conducted on the system with 3 different machine leaning algorithms, namely: SVM, K-nn and Naïve Bayes and variations of C and K for SVM and K-nn respectively. K-nn is also tested with different values of K as well as with and without weighted votes based on distance. The conclusion is that SVM with a $C = 0$ maintains the best position in terms of results at 83.33% for accuracy in all experiments. A detailed graphical breakdown of the results follows.

In Fig. 10, different accuracy levels for 11 different C values in the SVM algorithm are illustrated. $C = 0$ produces the highest result at 83.33%.

Fig. 11 compares accuracy levels for 6 different K-values in k-NN for 2 different setups: First, with normal or non-weighted votes and second with weighted votes. In general, weighted votes in this context do not seem to be of assistance. Furthermore, k-NN results with their peak at 58.33% are significantly lower than SVM results peaking at 83.33%.

In addition to SVM and k-NN another popular machine learning algorithm is experimented with, namely, Naïve Bayes. But its result is only as good as those of k-NN and is at 58.33%. The overall comparison of the 3 major algorithms is depicted in Fig. 12 with SVM at 83.33% and k-NN and Naïve Bayes at 58.33% at most on the used testing sample set.

5.6. Evaluation of sample size variations

In all of the above tests the test sample size is 12, which is way less than 1% of the total number of records that are available in our dataset. A number of tests have been conducted with other sizes for the testing-dataset and their results are accumulated in Table 11.

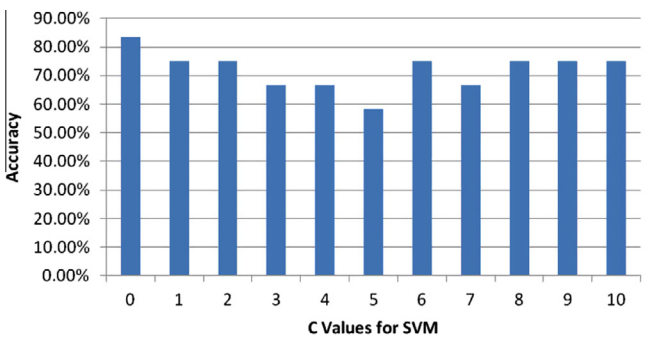

Fig. 10. Accuracy levels for different C Values in SVM.

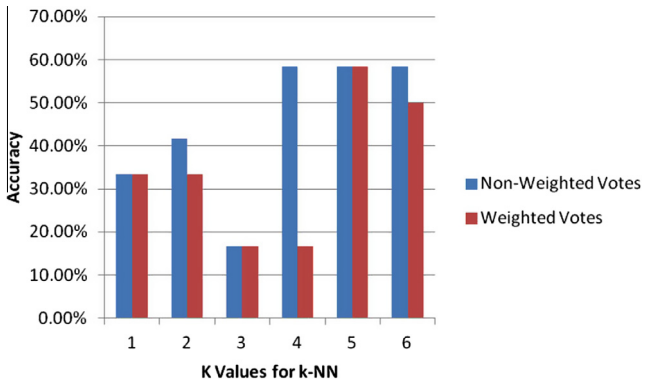

Fig. 11. Accuracy levels for different K values in k-NN for weighted and non-weighted votes.

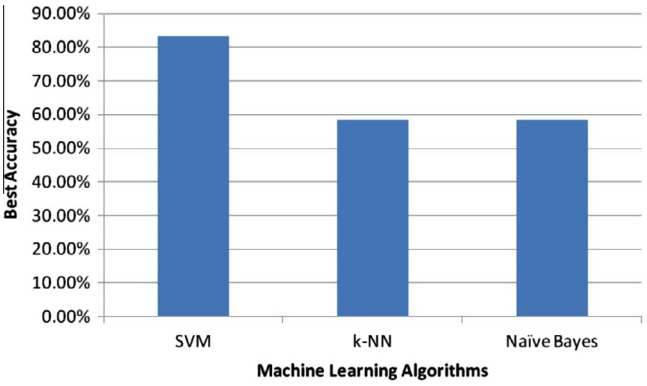

Fig. 12. Accuracy levels for different machine learning algorithms.

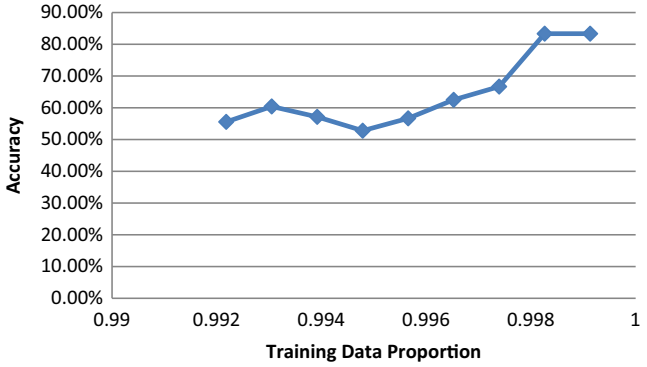

Fig. 13. Accuracy vs. training data-set size.

As illustrated in Fig. 13 below, the smaller the size of the testing dataset the bigger the size of the training dataset and therefore the more accurate the results.

6. Concluding remarks and future research

Accessing the fundamental data hidden in unstructured news text is challenging. This work addresses this challenge in a specific context by bringing together natural language processing and statistical pattern recognition as well as sentiment analysis to propose a system that predicts directional-movement of a currency-pair in the foreign exchange market based on the words used in adjacent news-headlines in the previous few hours. The system succeeds in doing so with an accuracy level of 83.33% in some cases. The proposed system is unique in at least three main aspects: One, the strategic choice of features based on hypernyms with a special heuristic; two, a novel feature-reduction algorithm named target-based feature-reduction and three a weighting mechanism based on a novel sentiment-weights termed SumScores.

Experiment-results in this work demonstrate a promising predictive relationship between news-headlines and currency-pair price-movements. This paper provides all the necessary details to have the entire system reproduced by other researchers for advancing this work. There are a number of areas that are immediate targets of future research, namely: (a) testing each of the newly proposed techniques in other contexts for example on movie reviews; (b) experimenting with other news sources and currency-pairs; (c) experimenting with other markets besides FOREX; (d) including new deep learning methods to improve conceptual abstraction.

Additionally, the main implications of this work are summarized in the below:

1. This work is among the first exploration efforts of the predictive relationship of news and the FOREX market. The promising results of this work indicate that such relationship exists and can be exploited in a predictive system like the one proposed. Therefore, the first implication of this work is that it acts as a successful feasibility test.
2. This work is an example of context-specific enhancement and specialization of text-mining methods through which promising results are achieved. This has the suggestive implication that text-mining research should be conducted in a more context-specific manner than it currently is.
3. This work emphasizes on Semantic Abstraction and Integration, Sentiment Integration and Dimensionality Reduction and produces promising results. An implication of this for future research is to also consider this strategy as one for improvement of text-mining methods in other contexts.
4. At a practical level, investment institutions and traders can benefit from the proposed market-prediction system. It can help make better financial decisions in the foreign exchange market which lead to financial returns on investments and avoidance of severe losses.
5. Financial markets are challenging to comprehend and lack of insights into them can lead to financial crises like the recent one in 2008 with negative impact on a wide range of people. A market-predictive text-mining solution may help bring about more confidence on comprehension of market-movements and its relation to human mass-psychology through text-mining of the available textual resources on the Internet.

Last but not least, this work is hoped to be advantageous to other researchers embarking on market-predictive text mining specially in the specific context of foreign exchange markets by providing grounds and a framework to build upon.

Acknowledgments

This research is supported by Program Rekan Penyelidikan UM Grant, University of Malaya, No. CG046-2013.

References

Aghdam, M. H., Ghasem-Aghaee, N., & Basiri, M. E. (2009). Text feature selection using ant colony optimization. *Expert Systems with Applications, 36*, 6843–6853.
Anastasakis, L., & Mort, N. (2009). Exchange rate forecasting using a combined parametric and nonparametric self-organising modelling approach. *Expert Systems with Applications, 36*, 12001–12011.

Baccianella, A. E. S., & Sebastiani, F. (2010). SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of the seventh conference on international language resources and evaluation LREC'10*. Valletta, Malta: European Language Resources Association (ELRA).

Bahrepour, M., Akbarzadeh, T. M.-R., Yaghoobi, M., & Naghibi, S. M.-B. (2011). An adaptive ordered fuzzy time series with application to FOREX. *Expert Systems with Applications, 38*, 475–485.

Berka, T., & Vajteršic, M. (2013). Parallel rare term vector replacement: Fast and effective dimensionality reduction for text. *Journal of Parallel and Distributed Computing, 73*, 341–351.

Bollen, J., & Huina, M. (2011). Twitter mood as a stock market predictor. *Computer, 44*, 91–94.

Bollen, J., Huina, M., & Zeng, Xiao-Jun (2010). Twitter mood predicts the stock market. *Journal of Computational Science, 2*, 1–8.

Butler, M., & Kešelj, V. (2009). Financial forecasting using character N-gram analysis and readability scores of annual reports. In Y. Gao & N. Japkowicz (Eds.). *Advances in artificial intelligence* (Vol. 5549, pp. 39–51). Berlin, Heidelberg: Springer.

Cambria, E., Schuller, B., Yunqing, X., & Havasi, C. (2013). New avenues in opinion mining and sentiment analysis. *IEEE Intelligent Systems, 28*, 15–21.

Chatrath, A., Miao, H., Ramchander, S., & Villupuram, S. (2014). Currency jumps, cojumps and the role of macro news. *Journal of International Money and Finance, 40*, 42–62.

Chen, J., Huang, H., Tian, S., & Qu, Y. (2009). Feature selection for text classification with Naïve Bayes. *Expert Systems with Applications, 36*, 5432–5435.

Chordia, T., Goyal, A., Lehmann, B. N., & Saar, G. (2013). High-frequency trading. *Journal of Financial Markets, 16*, 637–645.

Chordia, T., Roll, R., & Subrahmanyam, A. (2005). Evidence on the speed of convergence to market efficiency. *Journal of Financial Economics, 76*, 271–292.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning, 20*, 273–297.

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

Desmet, B., & Hoste, V. (2013). Emotion detection in suicide notes. *Expert Systems with Applications, 40*, 6351–6358.

Esuli, A., & Sebastiani, F. (2006). SENTIWORDNET: A publicly available lexical resource for opinion mining. In *Proceedings of the fifth conference on language resources and evaluation LREC'06* (pp. 417–422).

Evans, M. D. D., & Lyons, R. K. (2008). How is macro news transmitted to exchange rates? *Journal of Financial Economics, 88*, 26–50.

Fasanghari, M., & Montazer, G. A. (2010). Design and implementation of fuzzy expert system for Tehran stock exchange portfolio recommendation. *Expert Systems with Applications, 37*, 6138–6147.

Feng, G., Guo, J., Jing, B.-Y., & Hao, L. (2012). A Bayesian feature selection paradigm for text classification. *Information Processing & Management, 48*, 283–302.

Fung, G., Yu, J., & Lam, W. (2002). *News sensitive stock trend prediction*. In M.-S. Chen, P. Yu, & B. Liu (Eds.). *Advances in knowledge discovery and data mining* (Vol. 2336, pp. 481–493). Berlin/Heidelberg: Springer.

Garcke, J., Gerstner, T., & Griebel, M. (2013). Intraday foreign exchange rate forecasting using sparse grids. In J. Garcke & M. Griebel (Eds.). *Sparse grids and applications* (Vol. 88, pp. 81–105). Berlin, Heidelberg: Springer.

Ghazali, R., Hussain, A. J., & Liatsis, P. (2011). Dynamic ridge polynomial neural network: Forecasting the univariate non-stationary and stationary trading signals. *Expert Systems with Applications, 38*, 3765–3776.

Ghiassi, M., Skinner, J., & Zimbra, D. (2013). Twitter brand sentiment analysis: A hybrid system using n-gram analysis and dynamic artificial neural network. *Expert Systems with Applications, 40*, 6266–6282.

Groth, S. S., & Muntermann, J. (2011). An intraday market risk management approach based on textual analysis. *Decision Support Systems, 50*, 680–691.

Günal, S., Ergin, S., Gülmezoğlu, M. B, & Gerek, Ö. N. (2006). On feature extraction for spam e-mail detection. In B. Gunsel, A. Jain, A. M. Tekalp, & B. Sankur (Eds.). *Multimedia content representation, classification and security* (Vol. 4105, pp. 635–642). Berlin, Heidelberg: Springer.

Hagenau, M., Liebmann, M., & Neumann, D. (2013). Automated news reading: Stock price prediction based on financial news using context-capturing features. *Decision Support Systems, 55*, 685–697.

Huang, S.-C., Chuang, P.-J., Wu, C.-F., & Lai, H.-J. (2010). Chaos-based support vector regressions for exchange rate forecasting. *Expert Systems with Applications, 37*, 8590–8598.

Huang, C.-J., Liao, J.-J., Yang, D.-X., Chang, T.-Y., & Luo, Y.-C. (2010). Realization of a news dissemination agent based on weighted association rules and text mining techniques. *Expert Systems with Applications, 37*, 6409–6413.

Ikeda, K., Hattori, G., Ono, C., Asoh, H., & Higashino, T. (2013). Twitter user profiling based on text and community mining for market analysis. *Knowledge-Based Systems, 51*, 35–47.

Jeong, Y., & Myaeng, S.-H. (2013). Using WordNet hypernyms and dependency features for phrasal-level event recognition and type classification. In P. Serdyukov, P. Braslavski, S. Kuznetsov, J. Kamps, S. Rüger, E. Agichtein, I. Segalovich, & E. Yilmaz (Eds.). *Advances in information retrieval* (Vol. 7814, pp. 267–278). Berlin Heidelberg: Springer.

Jiang, S., Pang, G., Wu, M., & Kuang, L. (2012). An improved K-nearest-neighbor algorithm for text categorization. *Expert Systems with Applications, 39*, 1503–1509.

Jin, F., Self, N., Saraf, P., Butler, P., Wang, W., & Ramakrishnan, N. (2013). Forex-foreteller: Currency trend modeling using news articles. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 1470–1473). Chicago, IL, USA: ACM.

Kaltwasser, P. R. (2010). Uncertainty about fundamentals and herding behavior in the FOREX market. *Physica A: Statistical Mechanics and its Applications, 389*, 1215–1222.

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

Khadjeh Nassirtoussi, A., Ying Wah, T., & Ngo Chek Ling, D. (2011). A novel FOREX prediction methodology based on fundamental data. *African Journal of Business Management, 5*, 8322–8330.

Kim, K.-J. (2003). Financial time series forecasting using support vector machines. *Neurocomputing, 55*, 307–319.

Kontopoulos, E., Berberidis, C., Dergiades, T., & Bassiliades, N. (2013). Ontology-based sentiment analysis of twitter posts. *Expert Systems with Applications, 40*, 4065–4074.

Li, F. (2010). The information content of forward-looking statements in corporate filings—a Naïve Bayesian machine learning approach. *Journal of Accounting Research, 48*, 1049–1102.

Li, C. H., Yang, J. C., & Park, S. C. (2012). Text categorization algorithms using semantic approaches, corpus-based thesaurus and WordNet. *Expert Systems with Applications, 39*, 765–772.

Lugmayr, A., & Gossen, G. (2012). Evaluation of methods and techniques for language based sentiment analysis for DAX 30 stock exchange – a first concept of a "LUGO" sentiment indicator. In Lugmayr, A., Risse, T., Stockleben, B., Kaario, J., Pogorelc, B., & Serral Asensio, E. (Eds.). *SAME 2012 – fifth international workshop on semantic ambient media experience*.

Luo, Q., Chen, E., & Xiong, H. (2011). A semantic term weighting scheme for text categorization. *Expert Systems with Applications, 38*, 12708–12716.

Lupiani-Ruiz, E., García-Manotas, I., Valencia-García, R., García-Sánchez, F., Castellanos-Nieves, D., Fernández-Breis, J. T., et al. (2011). Financial news semantic search engine. *Expert Systems with Applications, 38*, 15565–15572.

Mabu, S., Hirasawa, K., Obayashi, M., & Kuremoto, T. (2013). Enhanced decision making mechanism of rule-based genetic network programming for creating stock trading signals. *Expert Systems with Applications, 40*, 6311–6320.

Mahajan, A., Dey, L., & Haque, S. M. (2008). Mining financial news for major events and their impacts on the market. In *IEEE/WIC/ACM international conference on web intelligence and intelligent agent technology, 2008. WI-IAT '08* (Vol. 1, pp. 423–426).

Miller, G. A. (1995). WordNet: A lexical database for English. *Communications of the ACM, 38*, 39–41.

Mittermayer, M. A. (2004). Forecasting intraday stock price trends with text mining techniques. In *Proceedings of the 37th annual Hawaii international conference on system sciences, 2004* (p. 10).

Moraes, F., Vasconcelos, M., Prado, P., Almeida, J., & Gonçalves, M. (2013). Polarity analysis of micro reviews in foursquare. In *Proceedings of the 19th Brazilian symposium on multimedia and the web* (pp. 113–120). Salvador, Brazil: ACM.

Mostafa, M. M. (2013). More than words: Social networks' text mining for consumer brand sentiments. *Expert Systems with Applications, 40*, 4241–4251.

Nizer, P. S. M., & Nievola, J. C. (2012). Predicting published news effect in the Brazilian stock market. *Expert Systems with Applications, 39*, 10674–10680.

Peramunetilleke, D., & Wong, R. K. (2002). Currency exchange rate forecasting from news headlines. *Australian Computer Science Communications, 24*, 131–139.

Pestov, V. (2013). Is the NN classifier in high dimensions affected by the curse of dimensionality? *Computers & Mathematics with Applications, 65*, 1427–1437.

Premanode, B., & Toumazou, C. (2013). Improving prediction of exchange rates using differential EMD. *Expert Systems with Applications, 40*, 377–384.

\Pui Cheong Fung, G., Xu Yu, J., & Wai, L. (2003). Stock prediction: Integrating text mining approach using real-time news. In *Proceedings. 2003 IEEE international conference on computational intelligence for financial engineering* (pp. 395–402).

Rachlin, G., Last, M., Alberg, D., & Kandel, A. (2007). ADMIRAL: A data mining based financial trading system. In *IEEE symposium on computational intelligence and data mining, 2007. CIDM 2007* (pp. 720–725).

Reboredo, J. C., Rivera-Castro, M. A., Miranda, J. G. V., & García-Rubio, R. (2013). How fast do stock prices adjust to market efficiency? Evidence from a detrended fluctuation analysis. *Physica A: Statistical Mechanics and its Applications, 392*, 1631–1637.

Schumaker, R. P., & Chen, H. (2009). Textual analysis of stock market prediction using breaking financial news: The AZF in text system. *ACM Transactions on Information Systems, 27*, 1–19.

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

Sermpinis, G., Laws, J., Karathanasopoulos, A., & Dunis, C. L. (2012). Forecasting and trading the EUR/USD exchange rate with gene expression and psi sigma neural networks. *Expert Systems with Applications, 39*, 8865–8877.

Sermpinis, G., Theofilatos, K., Karathanasopoulos, A., Georgopoulos, E. F., & Dunis, C. (2013). Forecasting foreign exchange rates with adaptive neural networks using radial-basis functions and particle swarm optimization. *European Journal of Operational Research, 225*, 528–540.

Shi, K., He, J., Liu, H.-T., Zhang, N.-T., & Song, W.-T. (2011). Efficient text classification method based on improved term reduction and term weighting. *The Journal of China Universities of Posts and Telecommunications, 18*(Suppl. 1), 131–135.

Soni, A., van Eck, N. J., & Kaymak, U. (2007). Prediction of stock price movements based on concept map information. In *IEEE symposium on computational intelligence in multicriteria decision making* (pp. 205–211).

Tan, S., Wang, Y., & Wu, G. (2011). Adapting centroid classifier for document categorization. *Expert Systems with Applications, 38*, 10264–10273.
Taşcı, Ş., & Güngör, T. (2013). Comparison of text feature selection policies and using an adaptive framework. *Expert Systems with Applications, 40*, 4871–4886.
Tay, F. E. H., & Cao, L. (2001). Application of support vector machines in financial time series forecasting. *Omega, 29*, 309–317.
Tetlock, P. C., Saar-Tsechansky, M., & Macskassy, S. (2008). More than words: Quantifying language to measure firms' fundamentals. *The Journal of Finance, 63*, 1437–1467.
Uysal, A. K., & Gunal, S. (2014). The impact of preprocessing on text classification. *Information Processing & Management, 50*, 104–112.
Vanstone, B., & Finnie, G. (2010). Enhancing stockmarket trading performance with ANNs. *Expert Systems with Applications, 37*, 6602–6610.
Vu, T. T., Chang, S., Ha, Q. T., & Collier, N. (2012). An experiment in integrating sentiment features for tech stock prediction in twitter. In *Proceedings of the workshop on information extraction and entity analytics on social media data* (pp. 23–38). Mumbai, India: The COLING 2012 Organizing Committee.
Werner, A., & Myrray, Z. F. (2004). Is all that talk just noise ? The information content of internet stock message boards. *Journal of Finance*, 1259–1294.
Wuthrich, B., Cho, V., Leung, S., Permunetilleke, D., Sankaran, K., & Zhang, J. (1998). Daily stock market forecast from textual web data. In *1998 IEEE international conference on systems, man, and cybernetics* (Vols. 3 and 2723, pp. 2720–2725).
Yu, Y., Duan, W., & Cao, Q. (2013). The impact of social and conventional media on firm equity value: A sentiment analysis approach. *Decision Support Systems*.
Yu, H., Nartea, G. V., Gan, C., & Yao, L. J. (2013). Predictive ability and profitability of simple technical trading rules: Recent evidence from Southeast Asian stock markets. *International Review of Economics & Finance, 25*, 356–371.
Zhai, Y., Hsu, A., & Halgamuge, S. K. (2007). Combining news and technical indicators in daily stock price trends prediction. In *Proceedings of the fourth international symposium on neural networks: advances in neural networks, Part III* (pp. 1087–1096). Nanjing, China: Springer-Verlag.
Zhang, W. (2011). *News based forecasting and modeling*. New York: State University of New York at Stony Brook.