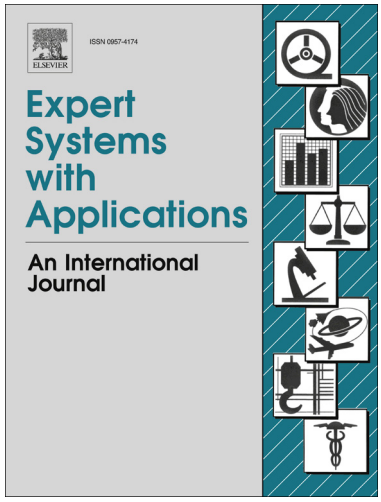

Review

Text Mining for Market Prediction: A Systematic Review

Arman Khadjeh Nassirtoussi, Teh Ying Wah, Saeed Reza Aghabozorgi, David Ngo Chek Ling

PII: S0957-4174(14)00345-5
DOI: http://dx.doi.org/10.1016/j.eswa.2014.06.009
Reference: ESWA 9376

To appear in: *Expert Systems with Applications*

Please cite this article as: Nassirtoussi, A.K., Wah, T.Y., Aghabozorgi, S.R., Ling, D.N.C., Text Mining for Market Prediction: A Systematic Review, *Expert Systems with Applications* (2014), doi: http://dx.doi.org/10.1016/j.eswa.2014.06.009

Text Mining for Market Prediction: A Systematic Review

Arman Khadjeh Nassirtoussi[1*], Teh Ying Wah[2], Saeed Reza Aghabozorgi[3] and David Ngo Chek Ling[4]

[1,2&3] Department of Information Science, Faculty of Computer Science & Information Technology, University of Malaya, 50603 Kuala Lumpur, Malaysia

[4] Research & Higher Degrees, Sunway University, No 5, Jalan University, Bandar Sunway, 46150 Petaling Jaya, Selangor DE, Malaysia

* Corresponding author. Email: armankhnt@gmail.com

Abstract

The quality of the interpretation of the sentiment in the online buzz in the social media and the online news can determine the predictability of financial markets and cause huge gains or losses. That is why a number of researchers have turned their full attention to the different aspects of this problem lately. However, there is no well-rounded theoretical and technical framework for approaching the problem to the best of our knowledge. We believe the existing lack of such clarity on the topic is due to its interdisciplinary nature that involves at its core both behavioural-economic topics as well as artificial intelligence. We dive deeper into the interdisciplinary nature and contribute to the formation of a clear frame of discussion. We review the related works that are about market prediction based on online-text-mining and produce a picture of the generic components that they all have. We, furthermore, compare each system with the rest and identify their main differentiating factors. Our comparative analysis of the systems expands onto the theoretical and technical foundations behind each. This work should help the research community to structure this emerging field and identify the exact aspects which require further research and are of special significance.

Keywords

Online Sentiment Analysis; Social Media Text Mining; News Sentiment Analysis; FOREX Market Prediction; Stock Prediction based on News

1. Introduction

The welfare of modern societies today depends on their market economies. At the heart of any market economy, lies the financial markets with their supply and demand equilibriums. Therefore, it is crucial to study markets and learn about their movements. Understanding market movements primarily facilitates one with the ability to predict future movements. Ability to predict in a market economy is equal to being able to generate wealth by avoiding financial losses and making financial gains. However, the nature of markets is as such that they are extremely difficult to predict if at all.

In general the predictive measures are divided into technical or fundamental analyses. They are differentiated based on their input data, with historic market data to be used for the former and any other kind of information or news about the country, society, company, etc for the latter. Most of the research in the past has been done on technical analysis approaches, mainly due to the availability of quantitative historic market data and the general desire among traders for technical quantitative methods. Fundamental data is more challenging to use as input specially when it is unstructured. Fundamental data may come from structured and numeric sources like macro-economic data or regular financial reports from banks and governments. Even this aspect of fundamental data has been rarely researched; but occasionally it has been demonstrated to be of predictive value as in the works of Chatrath, Miao, Ramchander, and Villupuram (2014), Khadjeh Nassirtoussi, Ying Wah, and Ngo Chek Ling (2011) and Fasanghari and Montazer (2010). Nevertheless, fundamental data available in unstructured text is the most challenging research aspect and therefore is the focus of this work. Some examples would be the fundamental data available online in the textual information in social media, news, blogs, forums, etc.

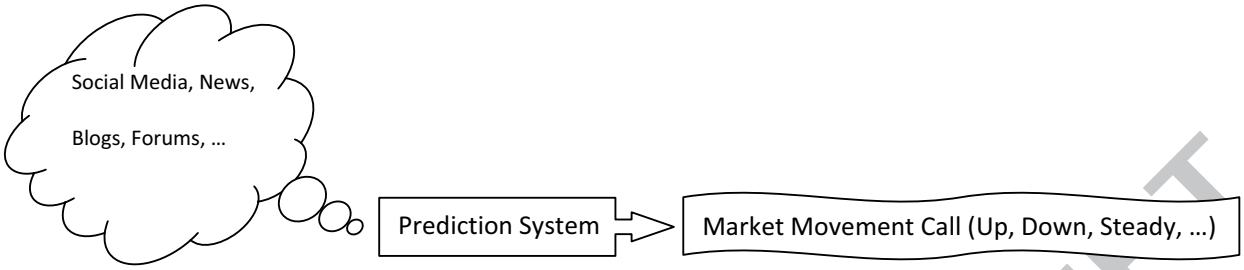

Figure 1 Online Text Sentimental Prediction System

In this work a systematic review of the past works of research with significant contribution to the topic of online-text-based market-prediction has been conducted, leading to the clarification of the today's cutting-edge research and its possible future directions. The main contributions of this work in summary are:

1- The relevant fundamental economic and computer/information science concepts are reviewed and it is clarified how they are tied into the currently proposed solutions for this research problem.

2- The most significant past literature has been reviewed with an emphasis on the cutting-edge pieces of work.

3- The main differentiating factors among the current works are identified, and used to compare and contrast the available solutions.

4- Observations are made on the areas with lack of research which can constitute possible opportunities for future work.

The rest of this paper is structured as follows. Section 2 provides insight into the interdisciplinary nature of the research problem at hand and defines the required foundational concepts for a robust comprehension of the literature. Section 3 presents the review of the main available work. Section 4 makes suggestions for future research. And Section 5 concludes this work.

2. A Review of Foundational Interdisciplinary Background Concepts

Essentially what is being targeted by this research is utilization of computational modelling and artificial intelligence in order to identify possible relationships between the textual information and the economy. That is the research problem.

In order to address this research problem adequately at least three separate fields of study must be included namely: Linguistics (to understand the nature of language), Machine-learning (to enable computational modelling and pattern recognition), Behavioral-economics (to establish economic sense).

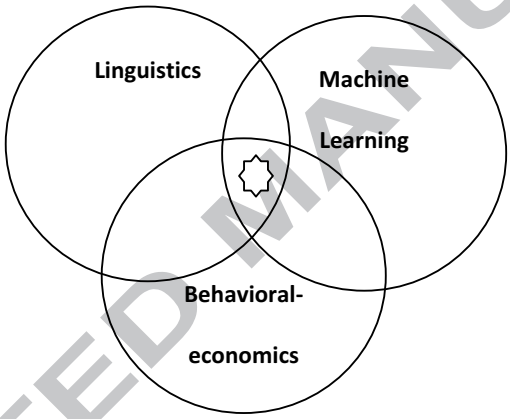

Figure 2 Interdisciplinary between linguistics, machine-learning and behavioural-economics

The main premise is aligned with recent findings of behavioural-economic principles whereby market conditions are products of human behaviours involved (Bikas, Jurevičienė, Dubinskas, & Novickytė, 2013; Tomer, 2007).

Our research has identified the below background topics as essential to develop a robust comprehension of this research problem:

2.1. Efficient Market Hypothesis (EMH)

The idea that markets are completely random and are not predictable is rooted in the efficient-market hypothesis (Fama, 1965) which asserts that financial markets are "informationally efficient"

and that consequently, one cannot consistently achieve returns in excess of average market returns on a risk-adjusted basis, given the information available at the time the investment is made. However, this absolute hypothesis is found not to be completely accurate and Fama himself revises it to include 3 levels of efficiency as strong, semi-strong and weak (Fama, 1970). This indicates that there are many markets where predictability is plausible and viable and such markets are termed as "weakly efficient". Market efficiency is correlated with information availability and a market is only "strongly efficient" when all information is completely available, which realistically is rarely the case. Hence, he conceded that his theory is stronger in certain markets where the information is openly, widely and instantly available to all participants and it gets weaker where such assumption cannot be held concretely in a market.

2.2. Behavioural-economics

Cognitive and behavioural economists look at price as a purely perceived value rather than a derivative of the production cost. The media do not report market status only, but they actively create an impact on market dynamics based on the news they release (Robertson, Geva, & Wolff, 2006; Wisniewski & Lambe, 2013).

People's interpretation of the same information varies greatly. Market participants have cognitive biases such as overconfidence, overreaction, representative bias, information bias, and various other predictable human errors in reasoning and information processing (Friesen & Weller, 2006).

Behavioural finance and investor sentiment theory have firmly established that investors' behaviour can be shaped by whether they feel optimistic (bullish) or pessimistic (bearish) about future market values (Bollen & Huina, 2011).

2.3. Adaptive Market Hypothesis (AMH)

The dilemma of market efficiency with regards to its degree and applicability to different markets is still a vibrant and ongoing topic of research with very contradictory results. For every paper

producing empirical evidence supporting the market efficiency, a contradictory paper can perhaps be found which empirically establishes market inefficiency (Majumder, 2013). Hence, a few years ago research has produced a counter-theory by the name of Adaptive Markets Hypothesis in an effort to reconcile Efficient Markets hypothesis with Behavioral Finance(Lo, 2005). (Urquhart & Hudson, 2013) have conducted a comprehensive empirical investigation on the Adaptive Markets Hypothesis (AMH) in three of the most established stock markets in the world; the US, UK and Japanese markets using very long run data. Their research provides evidence for adaptive markets, with returns going through periods of independence and dependence although the magnitude of dependence varies quite considerably. Thus the linear dependence of stock returns varies over time but nonlinear dependence is strong throughout. Their overall results suggest that the AMH provides a better description of the behaviour of stock returns than the Efficient Market Hypothesis.

2.4. Markets' Predictability

When markets are weakly efficient then it must be possible to predict their behaviour or at least determine criteria with predictive impact on them. Although, the nature of markets is as such that once such information is available, they absorb it and adjust themselves and therefore become efficient against the initial predictive criteria and thereby making them obsolete. Information absorption by markets and reaching new equilibriums constantly occur in markets and some researchers have delved into modelling its dynamics and parameters under special circumstances (García & Urošević, 2013). Nonetheless, the existence of speculative economic bubbles indicates that the market participants operate on irrational and emotional assumptions and do not pay enough attention to the actual underlying value. The research of (Potì & Siddique, 2013) indicates existence of predictability in the Foreign Exchange Market (FOREX) too. Although markets in the real world may not be absolutely efficient, research shows that some are more efficient than the others, as mentioned they can be strong, semi-strong and weak in terms of efficiency (Fama, 1970) . It has been demonstrated that markets in the emerging economies like Malaysia, Thailand, Indonesia, and

the Philippines tend to be significantly less efficient compared to the developed economies and hence predictive measures like technical trading rules seem to have more power (H. Yu, Nartea, Gan, & Yao, 2013). Additionally, the short-term variants of the technical trading rules have better predictive ability than long-term variants plus that markets become more informationally efficient over time (H. Yu, et al., 2013). Hence, there is a need to continuously revisit the level of efficiency of economically dynamic and rapidly growing emerging markets (H. Yu, et al., 2013).

2.5. Fundamental vs. Technical Analysis

Those who are convinced that markets can be at least to some extend predicted are segmented in two camps. Those who believe that historic market movements are bound to repeat themselves are known as technical analysts. They simply believe there are visual patterns in a market graph that an experienced eye can detect. Based on this belief many of the graph movements are named which forms the basis of technical analysis. At a higher level technical analysts try to detect such subtle mathematical models by the use of computation power and pattern recognition techniques. Although, technical analysis techniques are most wide spread among many of the market brokers and participants, to a scientific mind with a holistic point of view, technical analysis alone cannot seem very attractive. Specially because most technical analysis do not back any of their observations up with anything more than stating that patterns exist. They do very little if at all to find out the reason behind existence of patterns. Some of the common techniques in technical analysis are the moving average rules, relative strength rules, filter rules and the trading range breakout rules. In a recent study the effectiveness and limitations of these rules were put to test again and it was demonstrated that in many cases and contexts these rules are not of much predictive power (H. Yu, et al., 2013). Nevertheless, there are and continue to be many sophisticated financial prediction modelling efforts based on various types or combinations of machine learning algorithms like neural networks (Anastasakis & Mort, 2009; Ghazali, Hussain, & Liatsis, 2011; Sermpinis, Laws, Karathanasopoulos, & Dunis, 2012; Vanstone & Finnie, 2010), fuzzy logic (Bahrepour, Akbarzadeh-T.,

Yaghoobi, & Naghibi-S., 2011), Support Vector regression (S.-C. Huang, Chuang, Wu, & Lai, 2010; Premanode & Toumazou, 2013), rule-based genetic network programming (Mabu, Hirasawa, Obayashi, & Kuremoto, 2013).

However, there is a second school of thought known as fundamental analysis, which seems to be more promising. In fundamental analysis analysts look at fundamental data that is available to them from different sources and make assumptions based on that. We can come up with at least 5 main sources of fundamental data: 1- the financial data of a company like data in its balance sheet or financial data about a currency in the FOREX market, 2- Financial data about a market like its index, 3- Financial data about the government activities and banks, 4- Political circumstances, 5- Geographical and meteorological circumstances like natural or unnatural disasters. However, determining underlying fundamental value of any asset may be challenging and with a lot of uncertainty (Kaltwasser, 2010). Therefore, the automation of fundamental analysis is rather rare. Fasanghari and Montazer (2010) design a fuzzy expert system for stock exchange portfolio recommendation that takes some of the company fundamentals through numeric metrics as input.

Most market participants try to keep an eye on both technical and fundamental data, however, fundamental data is usually of an unstructured nature and it remains to be a challenge to make the best use of it efficiently through computing. The research challenge here is to deal with this unstructured data. One recent approach that is emerging in order to facilitate such topical unstructured data and extract structured data from it is the development of specialized search engines like this financial news semantic search engine (Lupiani-Ruiz, et al., 2011). However, it remains to be a challenge to extract meaning in a reliable fashion from text and a search engine like the above is limited to extracting the available numeric data in the relevant texts. One recent study shows the impact of US news, UK news and Dutch news on three Dutch banks during the financial crisis of 2007-2009 (Kleinnijenhuis, 2013). This specific study goes on to explore market panics from

a financial journalism perspective and communication theories specifically in an age of algorithmic and frequency trading.

2.6. Algorithmic Trading

Algorithmic trading refers to predictive mechanisms by intelligent robotic trading agents that are actively participating in every moment of market trades. The speed of such decision making has increased dramatically recently and has created the new term frequency trading. Frequency trading has been more popular in the stock market and Evans, Pappas, and Xhafa (2013) are among the first who are utilizing artificial neural networks and genetic algorithms to build an algo-trading model for intra-day foreign exchange speculation.

2.7. Sentiment and Emotional Analysis

It deals with detecting the emotional sentiment preserved in text through specialized semantic analysis for a variety of purposes for instance to gauge the quality of market reception for a new product and the overall customer feedback or to estimate the popularity of a product or brand (Ghiassi, Skinner, & Zimbra, 2013; Mostafa, 2013) among people. There is a body of research that is focused on sentiment analysis or the so called “opinion mining” (Balahur, Steinberger, Goot, Pouliquen, & Kabadjov, 2009; Cambria, Schuller, Yunqing, & Havasi, 2013; Hsinchun & Zimbra, 2010). It is mainly based on identifying positive and negative words and processing text with the purpose of classifying its emotional stance as positive or negative. An example of such sentiment analysis effort is the work of Maks and Vossen (2012) which presents a lexicon model for deep sentiment analysis and opinion mining. A similar concept can be used in other area of research like emotion detection in suicide notes as done by Desmet and Hoste (2013).

However, an analysis of emotional sentiment of news text can also be explored for market prediction. Schumaker, Zhang, Huang, and Chen (2012) have tried to evaluate the sentiment in financial news articles in relation to the stock market in his research but has not been completely

successful. One more successful recent example of this would be (L.-C. Yu, Wu, Chang, & Chu, 2013), where a contextual entropy model was proposed to expand a set of seed words by discovering similar emotion words and their corresponding intensities from online stock market news articles. This was accomplished by calculating the similarity between the seed words and candidate words from their contextual distributions using an entropy measure. Once the seed words have been expanded, both the seed words and expanded words are used to classify the sentiment of the news articles. Their experimental results show that the use of the expanded emotion words improved classification performance, which was further improved by incorporating their corresponding intensities which caused accuracy results to range from 52% to 91.5% by varying the difference of intensity levels from positive and negative classes from (-0.5 to 0.5) to >9.5 respectively. It is also interesting to note that emotional analysis of text does not have to be mere based on positivity-negativity and it can be done on other dimensions or on multi-dimensions (Ortigosa-Hernández, et al., 2012). A recent piece of research by Loia and Senatore (2014) introduces a framework for extracting the emotions and the sentiments expressed in the textual data. The sentiments are expressed by a positive or negative polarity. The emotions are based on the Minsky's conception of emotions that consists of four affective dimensions (Pleasantness, Attention, Sensitivity and Aptitude). Each dimension has six levels of activation, called sentic levels. Each level represents an emotional state of mind and can be more or less intense, depending on the position on the corresponding dimension. Another interesting development in sentiment analysis is an effort to go from sentiment of chunks of text to specific features or aspects that are related to a concept or product; Kontopoulos, Berberidis, Dergiades, and Bassiliades (2013) propose a more efficient sentiment analysis of Twitter posts, whereby posts are not simply characterized by a sentiment score but instead receive a sentiment grade for each distinct notion in them which is enabled by the help of an ontology. W. Li and Xu (2014) take on another angle by looking for features that are ''meaningful'' to emotions instead of simply choosing words with high co-occurrence degree and thereby create a text-based emotion classification using emotion cause extraction.

3. Review of the main available work

Despite the existence of multiple systems in this area of research we have not found any dedicated and comprehensive comparative analysis and review of the available systems. Nikfarjam, Emadzadeh, and Muthaiyah (2010) have made an effort in form of a conference paper that provides a rough summary under the title of "Text mining approaches for stock market prediction". Hagenau, Liebmann, and Neumann (2013) have also presented a basic comparative table in their literature review that has been referred to in some parts of this work. In this section we are filling this gap by reviewing the major systems that have been developed around the past decade.

3.1. The generic overview

All of these systems have some of the components depicted in figure 3. At one end text is fed as input to the system and at the other end some market predictive values are generated as output.

In the next sections a closer look is taken at each of the depicted components, their roles and theoretical foundation.

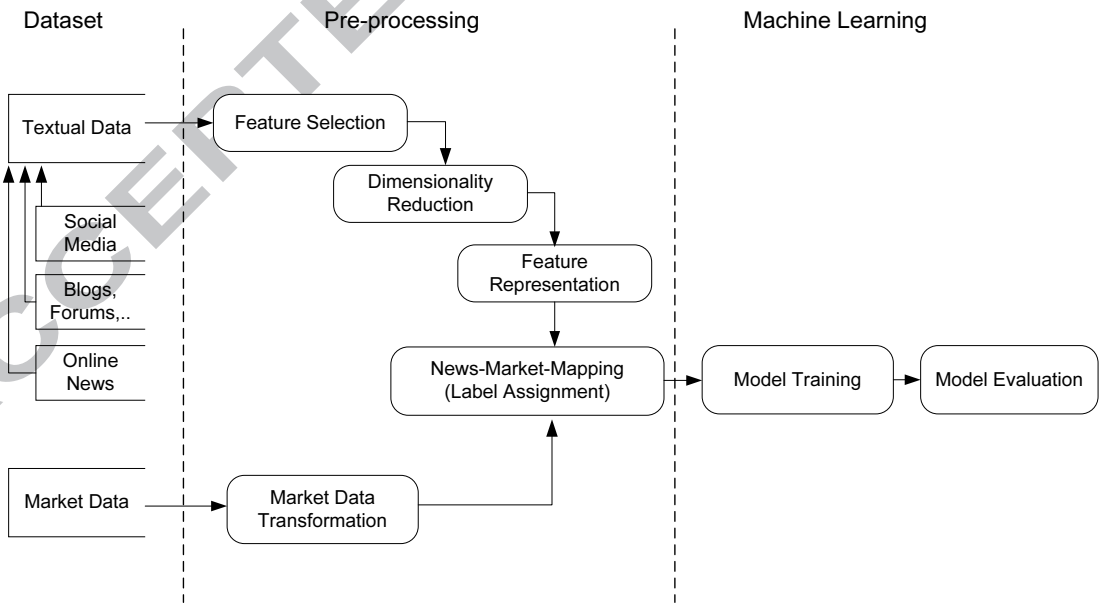

Figure 3 Generic Common System Components Diagram

3.2. Input Dataset

All systems are taking at least two sources of data as input, namely, the textual data from the online resources and the market data.

3.2.1. Textual Data

The textual input can have several sources and content types accordingly as shown in table 1. The majority of the sources used are major news websites like The Wall Street Journal (Werner & Myrray Z., 2004), Financial Times (Wuthrich, et al., 1998), Reuters (Pui Cheong Fung, Xu Yu, & Wai, 2003), Dow Jones, Bloomberg (Chatrath, et al., 2014; Jin, et al., 2013), Forbes (Rachlin, Last, Alberg, & Kandel, 2007) as well as Yahoo! Finance (Schumaker, et al., 2012). The type of the news is either general news or special financial news. The majority of the systems are using financial news as it is deemed to have less noise compared with general news. What is being extracted here is the news text or the news headline. News headlines are occasionally used and are argued to be more straight-to-the-point and, hence, of less noise caused by verbose text (C.-J. Huang, Liao, Yang, Chang, & Luo, 2010; Peramunetilleke & Wong, 2002). Fewer researchers have looked at less formal sources of textual information e.g. Das and Chen (2007) have looked at the text on online message boards in their work. And more recently Y. Yu, Duan, and Cao (2013) have looked at the textual content from the social media like twitter and blog posts. Some researchers focus solely on twitter and utilize it for market prediction and public mood analysis more efficiently (Bollen & Huina, 2011; Vu, Chang, Ha, & Collier, 2012). A third class of textual source for the systems has been the company annual reports, press releases and corporate disclosures. We observe that a difference about this class of information about companies is the nature of their timing, whereby regular reports and disclosures have prescheduled times. Pre-set timing is suspected to have an impact on prediction capability or nature which may have been caused by anticipatory reactions among market participants (Chatrath, et al., 2014; C.-J. Huang, et al., 2010). Therefore, we have included a column for this information in the table. It is also important to remember that some textual reports can have structured or semi-

structured formats like macroeconomic announcement that come from governments or central banks on the unemployment rates or the Gross Domestic Product (GDP). Chatrath, et al. (2014) have used such structured data to predict jumps in the foreign exchange market (FOREX). Furthermore, with regards to information release frequency, one interesting observation that has been made in the recent research is that increase of the news release frequency can cause a decrease in informed-trading and hence the degree of information asymmetry is lower for firms with more frequent news releases (Sankaraguruswamy, Shen, & Yamada, 2013). This raises an interesting point whereby the uninformed traders increase and thereby play a significant role in the market with too frequent news releases. Although it may feel counter-intuitive as one may expect more informed trading to occur under such circumstances.

Reference	Text Type	Text Source	No. of items	Prescheduled	Unstructured
Wuthrich, et al. (1998)	General news	The Wall Street Journal, Financial Times, Reuters, Dow Jones , Bloomberg	Not given	No	Yes
Peramunetilleke and Wong (2002)	Financial news	HFDF93 via www.olsen.ch	40 headlines per hour	No	Yes
Pui Cheong Fung, et al. (2003)	Company news	Reuters Market 3000 Extra	600,000	No	Yes
Werner and Myrray Z. (2004)	Message postings	Yahoo! Finance, Raging Bull, Wall Street Journal	1.5 million messages	No	Yes
Mittermayer (2004)	Financial news	Not mentioned	6,602	No	Yes
Das and Chen (2007)	Message postings	Message boards	145,110 messages	No	Yes
Soni, van Eck, and Kaymak (2007)	Financial news	FT Intelligence (Financial Times online service)	3493	No	Yes
Zhai, Hsu, and Halgamuge (2007)	Market-sector news	Australian Financial Review	148 direct company news and 68 indirect ones	No	Yes
Rachlin, et al. (2007)	Financial news	Forbes.com, today.reuters.com	Not mentioned	No	Yes
Paul C. Tetlock, Saar-Tsechansky, and Macskassy (2008)	Financial news	Wall Street Journal, Dow Jones News Service from Factiva news database.	350,000 stories	No	Yes
Mahajan, Dey, and Haque	Financial news	Not mentioned	700 news articles	No	Yes

(2008)					
Butler and Kešelj (2009)	Annual reports	Company websites	Not mentioned	Yes	Yes
Schumaker and Chen (2009)	Financial news	Yahoo Finance	2800	No	Yes
F. Li (2010)	Corporate filings	Management's Discussion and Analysis section of 10-K and 10-Q filings from SEC Edgar Web site	13 million forward-looking-statements in 140,000 10-Q and K filings	Yes (company annual report)	Yes
C.-J. Huang, et al. (2010)	Financial news	Leading electronic newspapers in Taiwan	12,830 headlines	No	Yes
Groth and Muntermann (2011)	Adhoc announcements	Corporate disclosures	423 disclosures	No	Yes
Schumaker, et al. (2012)	Financial news	Yahoo! Finance	2802	No	Yes
Lugmayr and Gossen (2012)	Broker newsletters	Brokers	Not available	No	Yes
Y. Yu, et al. (2013)	Daily conventional and social media	Blogs, forums, news and micro blogs (e.g., Twitter)	52,746 messages	No	Yes
Hagenau, et al. (2013)	Corporate announcements & financial news	DGAP, EuroAdhoc	10870 & 3478 respectively	No	Yes
Jin, et al. (2013)	General news	Bloomberg	361,782	No	Yes
Chatrath, et al. (2014)	Macroeconomic news	Bloomberg	Not mentioned	Yes	No
Bollen and Huina (2011)	Tweets	Twitter	9,853,498	No	Yes
Vu, et al. (2012)	Tweets	Twitter	5,001,460	No	Yes

Table 1 Comparison of the textual input for different systems

3.2.2. Market Data

The other source of input data for the systems comes from the numeric values in financial markets in form of price-points or indexes. This data is used mostly for purpose of training the machine learning algorithms and occasionally it is used for prediction purposes whereby it is fed into the machine learning algorithm as an independent variable for a feature, this topic will be discussed in a later section. In table 2, crucial details of such market data are provided. Firstly, there is a differentiation made between the stock markets and the foreign exchange market (FOREX). Past research has been mostly focused on stock market prediction, either in form of a stock market index like the Dow Jones

Industrial Average (Bollen & Huina, 2011; Werner & Myrray Z., 2004; Wuthrich, et al., 1998), the US NASDAQ Index (Rachlin, et al., 2007), Morgan Stanley High-Tech Index (MSH) (Das & Chen, 2007), the Indian Sensex Index (Mahajan, et al., 2008), S&P 500 (Schumaker & Chen, 2009) or the stock price of a specific company like BHP Billiton Ltd. (BHP.AX) (Zhai, et al., 2007) or like Apple, Google, Microsoft and Amazon in Vu, et al. (2012) or a group of companies (Hagenau, et al., 2013). The FOREX market has been only occasionally addressed in about ten percent of the reviewed works; like in the work of Peramunetilleke and Wong (2002) and more recently in the works of Chatrath, et al. (2014) and Jin, et al. (2013). It is worth considering that the efficiency levels of the FOREX markets around the world vary (Wang, Xie, & Han, 2012), and hence, it should be possible to find less efficient currency pairs that are prone to predictability.

Moreover, almost all the forecast types on any of the market measures above are categorical with discrete values like Up, Down and Steady. There are very few pieces of research that have explored an approach based on linear regression (Jin, et al., 2013; Schumaker, et al., 2012; Paul C. Tetlock, et al., 2008).

Furthermore, the predictive timeframe for each work is compared. The timeframe from the point of news release to a market impact observation can vary from seconds to days, weeks or months. The second or millisecond impact prediction is the element that fuels an entire industry by the name of micro-trading in which special equipment and vicinity to the news source as well as the market computer systems is critical; a work on trading trends has well explained such quantitative tradings (Chordia, Roll, & Subrahmanyam, 2011). Another name for the same concept is High Frequency Trading which is explored in detail by Chordia, Goyal, Lehmann, and Saar (2013). Another similar term is Low-Latency trading which is amplified in the work of Hasbrouck and Saar (2013). However, past research has indicated sixty minutes to be looked at as a reasonable market convergence time to efficiency (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. Information availability and distribution channels are critical here and the research on the market efficiency convergence time is ongoing (Reboredo, Rivera-Castro, Miranda, & García-Rubio, 2013). Most of the works compared in table 2 have a daily timeframes followed by intraday timeframes which are in the range of 5 minutes (Chatrath, et al., 2014), 15 minutes (Werner & Myrray Z., 2004), 20 minutes (Schumaker & Chen, 2009) to 1, 2, or 3 hours (Peramunetilleke & Wong, 2002). Experiment periods are also contrasted with shortest being only 5 days in case of the work of Peramunetilleke and Wong (2002) and up to multiple years with the longest at 24 years from 1980 to 2004 by P. C. Tetlock (2007) followed by 14 years from 1997 to 2011 in the work of Hagenau, et al. (2013) , 13 years from 1994 to 2007 in the work of F. Li (2010) with the latter looking at an annual timeframe and the formers at daily timeframes. The remaining majority of the works take on an experiment period with a length of multiple months as detailed in table 2.

Reference	Market	Market Index	Time-frame	Period	Forecast type
Wuthrich, et al. (1998)	Stocks	Dow Jones Industrial Average, the Nlkkei 225, the Financial Times 100, the Hang Seng, and the Singapore Straits	Daily	6 Dec 1997 to 6 Mar 1998	Categorical: Up, Steady, Down
Peramunetilleke and Wong (2002)	FOREX	Exchange rate(USD-DEM, USD-JYP)	Intraday(1, 2, or 3 hours)	22 to 27 Sept 1993	Categorical: Up, Steady, Down
Pui Cheong Fung, et al. (2003)	Stocks	33 stocks from the Hang Seng	Daily (No delay & varying lags)	1 Oct 2002 to 30 Apr 2003	Categorical: Rise, Drop
Werner and Myrray Z. (2004)	Stocks	Dow Jones Industrial Average, and the Dow Jones Internet	Intraday(15-min, 1 hour and 1 day)	Year 2000	Categorical: Buy, Sell, and Hold messages
Mittermayer (2004)	Stocks	Stock prices	Daily	1 Jan to 31 Dec 2002	Categorical: good news, bad news , no movers
Das and Chen (2007)	Stocks	24 tech-sectors in the Morgan Stanley High-Tech	Daily	July and August 2001	Aggregate sentiment index
Soni, et al. (2007)	Stocks	11 oil and gas companies	Daily	1 Jan 1995 to 15 May 2006	Categorical: Positive, Negative
Zhai, et al. (2007)	Stocks	BHP Billiton Ltd. from Australian Stock Exchange	Daily	1 Mar 2005 to 31 May 2006	Categorical: Up, Down
Rachlin, et al.	Stocks	5 stocks from US	Days	7 Feb to 7	Categorical: Up,

(2007)		NASDAQ		May 2006	Slight-Up, Expected, Slight-Down, Down
Paul C. Tetlock, et al. (2008)	Stocks	Individual S&P 500 firms and their future cash flows	Daily	1980 to 2004	Regression of a measure called neg.
Mahajan, et al. (2008)	Stocks	Sensex	Daily	5 Aug to 8 Apr	Categorical
Butler and Kešelj (2009)	Stocks	1-Year market drift	Yearly	2003 to 2008	Categorical: Over- or under-perform S&P 500 index over the coming year
Schumaker and Chen (2009)	Stocks	S&P 500 stocks	Intraday (20 min)	26 Oct to 28 Nov 2005	Categorical: Discrete numeric
F. Li (2010)	Stocks	(1) Index (2) Quarterly earnings and cash flows (3) Stock returns	Annually (Quarterly with 3 dummy quarters)	1994 to 2007	Categorical based on tone: Positive, Negative, Neutral, Uncertain
C.-J. Huang, et al. (2010)	Stocks	Taiwan Stock Exchange Financial Price	Daily	Jun to Nov 2005	Just significance degree assignment
Groth and Muntermann (2011)	Stocks	Abnormal risk exposure ARISKτ (Thomson Reuters DataScope Tick History)	Intraday (volatility during the τ=15 and 30 min)	1 Aug 2003 to 31 Jul 2005	Categorical: Positive, Negative
Schumaker, et al. (2012)	Stocks	S&P 500	Intraday (20min)	26 Oct to 28 Nov 2005	Regression
Lugmayr and Gossen (2012)	Stocks	DAX 30 Performance Index	Intraday (3 or 4 times per day)	Not available	Categorical: Sentiment [-1, 1]; Trend (Bear, Bull, Neutral); Trend strength (in %)
Y. Yu, et al. (2013)	Stocks	Abnormal returns and cumulative abnormal returns of 824 firms	Daily	1 Jul to 30 Sept 2011	Categorical: Positive or Negative
Hagenau, et al. (2013)	Stocks	Company specific	Daily	1997 to 2011	Categorical: Positive or Negative
Jin, et al. (2013)	FOREX	Exchange rate	Daily	1 Jan to 31 Dec 2012	Regression
Chatrath, et al. (2014)	FOREX	Exchange rate	Intraday(5min)	Jan 2005 to Dec 2010	Categorical: Positive or Negative jumps
Bollen and Huina (2011)	Stock	DJIA	Daily	28 Feb to 19 Dec 2008	Regression Analysis
Vu, et al. (2012)	Stock	Stock prices (at NASDAQ for AAPL, GOOG, MSFT, AMZN)	Daily	1 Apr 2011 to 31 May 2011, online test 8 Sep to 26 Sep 2012	Categorical: up and down

Table 2 The input market data, experiment timeframe, length and forecast type

3.3. Pre-processing

Once the input data is available, it must be prepared so that it can be fed into a machine learning algorithm. This for the textual data means to transform the unstructured text into a representative format that is structured and can be processed by the machine. In data mining in general and text mining specifically the pre-processing phase is of significant impact on the overall outcomes (Uysal & Gunal, 2014). There are at least three sub-processes or aspects of pre-processing which we have contrasted in the reviewed works, namely: Feature-Selection, Dimensionality-Reduction, Feature-Representation

3.3.1 Feature-Selection

The decision on features through which a piece of text is to be represented is crucial because from an incorrect representational input nothing more than a meaningless output can be expected. In table 3 the type of feature selection from text for each of the works is listed.

In most of the literature the most basic techniques have been used when dealing with text-mining-based market-prediction-problems as alluded to by Kleinnijenhuis (2013) as well consistent with our findings. The most common technique is the so called “bag-of-words” which is essentially breaking the text up into its words and considering each of the as a feature. As presented in table 3, around three quarters of the works are relying on this basic feature-selection technique in which the order and co-occurrence of words are completely ignored. Schumaker, et al. (2012) and Schumaker and Chen (2009) have explored two further techniques namely noun-phrases and named entities. In the former, first the words with a noun part-of-speech are identified with the help of a lexicon and then using syntactic rules on the surrounding parts of speech, noun-phrases are detected. In the latter, a category system is added in which the nouns or the noun phrases are organized. They used the so called MUC-7 framework of entity classification, where categories include date, location, money, organization, percentage, person and time. However, they did not have any success in improving their noun-phrase results via this further categorization in their named-entities technique. On the other hand, Vu, et al. (2012) successfully improve results by building a Named Entity Recognition

(NER) system to identify whether a Tweet contains named entities related to their targeted companies based on a linear Conditional Random Fields (CRF) model. Another less frequently used but interesting technique is the so called Latent Dirichlet Allocation (LDA) technique used by Jin, et al. (2013) as well as Mahajan, et al. (2008) in table 3 to categorize words into concepts and use the representative concepts as the selected features. Some of the other works are using a technique called n-grams (Butler & Kešelj, 2009; Hagenau, et al., 2013). An n-gram is a contiguous sequence of n items which are usually words from a given sequence of text. However, word sequence and syntactic structures could essentially be more advanced. An example for such setup could be the use of Syntactic N-grams (Sidorov, Velasquez, Stamatatos, Gelbukh, & Chanona-Hernández, 2013). However, inclusion of such features may give rise to language-dependency problems which need to be dealt with (Kim, et al., 2014). Combination techniques of lexical and semantic features for short text classification may also be enhanced (Yang, Li, Ding, & Li, 2013).

Feature selection is a standard step in the pre-processing phase of data mining and there are many other approaches that can be considered for textual feature selection, genetic algorithms being one of them as described in detail in the work of Tsai, Eberle, and Chu (2013). In another work Ant Colony Optimization has been successfully used for textual feature-selection (Aghdam, Ghasem-Aghaee, & Basiri, 2009). Feature selection for text classification can also be done based on a filter-based probabilistic approach (Uysal & Gunal, 2012). Feng, Guo, Jing, and Hao (2012) propose a generative probabilistic model, describing categories by distributions, handling the feature selection problem by introducing a binary exclusion/inclusion latent vector, which is updated via an efficient Metropolis search. Delicate pre-processing has been shown to have significant impact in similar text mining problems (Haddi, Liu, & Shi, 2013; Uysal & Gunal, 2014).

3.3.2. Dimensionality-Reduction

Having a limited number of features is extremely important as the increase in the number of features which can easily happen in feature-selection in text can make the classification or clustering problem extremely hard to solve by decreasing the efficiency of most of the learning algorithms, this situation is widely known as the curse of dimensionality (Pestov, 2013). In table 3 under Dimensionality-Reduction the approach taken by each of the works is pointed out. Zhai, et al. (2007) does this by choosing the top 30 concepts with the highest weights as the features instead of all available concepts. Mittermayer (2004) does this by filtering for the top 1000 terms out of all terms. The most common approach however is setting a minimum occurrence limit and reducing the terms by selecting the ones reaching a number of occurrences (Butler & Kešelj, 2009; Schumaker & Chen, 2009). Next common approach is using a predefined dictionary of some sort to replace them with a category name or value. Some of these dictionaries are specially put together by a market expert like the one used by Wuthrich, et al. (1998) or Peramunetilleke and Wong (2002) or they are more specific to a specific field like psychology in the case of Harvard-IV-4 which has been used in the work of Paul C. Tetlock, et al. (2008). And other times they are rather general use dictionaries like the WordNet Thesaurus used by Zhai, et al. (2007). At times a dictionary or thesaurus is created dynamically based on the text corpus using a term extraction tool (Soni, et al., 2007). Another set of activities that usually constitute the minimum of dimensionality-reduction are: features stemming, conversion to lower case letters, punctuation removal and removal of numbers, web page addresses and stop-words. These steps are almost always taken and in some works are the only steps taken as in the work by Pui Cheong Fung, et al. (2003).

Feature-reduction can be enhanced in a number of ways but the current research has yet not delved into it much as observed in the reviewed works. Berka and Vajteršic (2013) introduce a detailed method for dimensionality reduction for text, based on parallel rare term vector replacement.

3.3.3. Feature-Representation

After the minimum number of features is determined, each feature needs to be represented by a numeric value so that it can be processed by machine learning algorithms. Hence, the title "Feature-Representation" in table 3 is used for the column whereby the type of numeric value that is associated with each feature is compared for all the reviewed works. This assigned numeric value acts like a score or a weight. There are at least 5 types for it that are very popular, namely, Information Gain (IG), Chi-square Statistics (CHI), Document Frequency (DF), Accuracy Balanced (Acc2) and Term Frequency-Inverse Document Frequency (TF-IDF). A comparison of these five metrics along with proposals for some new metrics can be found in the work of Taşcı and Güngör (2013).

The most basic technique is a Boolean or a binary representation whereby two values like 0 and 1 represent the absence or presence of a feature e.g. a word in the case of a bag-of-words technique as in these works (Mahajan, et al., 2008; Schumaker, et al., 2012; Wuthrich, et al., 1998). The next most common technique is the Term Frequency–Inverse Document Frequency or TF-IDF (Groth & Muntermann, 2011; Hagenau, et al., 2013; Peramunetilleke & Wong, 2002; Pui Cheong Fung, et al., 2003). 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, to balance out the general popularity of some words. There are also other similar measures that are occasionally used like the Term Frequency-Category Discrimination (TF-CDF) metric that is derived from Category Frequency (CF) and proves to be more effective than TF-IDF in this work (Peramunetilleke & Wong, 2002).

Generally in text-mining enhanced feature-reduction(dimensionality-reduction) and feature weighting(feature-representation) can have significant impact on the eventual text-classification efficiency (Shi, He, Liu, Zhang, & Song, 2011).

Reference	Feature Selection	Dimensionality Reduction	Feature Representation
Wuthrich, et al.	Bag-of-words	Pre-defined dictionaries (word-sequences	Binary

(1998)		by an expert)	
Peramunetilleke and Wong (2002)	Bag-of-words	Set of keyword records	Boolean, TF-IDF, TF-CDF
Pui Cheong Fung, et al. (2003)	Bag-of-words	Stemming, conversion to lower-case, removal of punctuation, numbers, web page addresses and stop-words	TF-IDF
Werner and Myrray Z. (2004)	Bag-of-words	Minimum information criterion (top 1,000 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 N.J. 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, et al. (2007)	Bag-of-words, commonly used financial values	Most influential keywords list (Automatic extraction)	TF, Boolean, Extractor software output
Paul C. 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
F. Li (2010)	Bag-of-words, tone and content	Pre-defined dictionaries	Binary, Dictionary value
C.-J. 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
Y. Yu, et al. (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, et al. (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

Table 3 Pre-processing: Feature-Selection, Feature-Reduction, Feature-Representation

3.4. Machine Learning

After the pre-processing is completed and text is transformed into a number of features with a numeric representation, machine learning algorithms can be engaged. In the following a brief summary of these algorithms is presented as well as a comparison on some of the other detail technicalities in the designs of the reviewed systems.

3.4.1. Machine Learning Algorithms

In this section, it is attempted to provide a summary of the machine learning algorithms used in the reviewed works. It is noted that comparing such algorithms is not easy and full of pitfalls (Salzberg, 1997). Therefore, the main objective is to report what is used; so that it helps understand what lacks and may be possible for future research. Almost all the machine learning algorithms that have been used are classification algorithms as listed in Table 4. Basically the systems are using the input data to learn to classify an output usually in terms of the movement of the market in classes such as UP, DOWN and STEADY. However, there is also a group of works which use regression analysis for making predictions and not classification.

Table 4, categorizes the reviewed works based on their used algorithm into 6 classes:

A) Support Vector Machine (SVM)

B) Regression Algorithms

C) Naïve Bayes

D) Decision Rules or Trees

E) Combinatory Algorithms

F) Multi-algorithm experiments

A) Support Vector Machine (SVM): This section in table 4 contains the class of algorithms that the vast majority of the reviewed works are using (Mittermayer, 2004; Pui Cheong Fung, et al., 2003; Schumaker & Chen, 2009; Soni, et al., 2007; Werner & Myrray Z., 2004). SVM is a non-probabilistic binary linear classifier used for supervised learning. The main idea of SVMs is finding a hyperplane that separates two classes with a maximum margin. The training problem in SVMs can be represented as a quadratic programming optimization problem. A common implementation of SVM that is used in many works (Mittermayer, 2004; Pui Cheong Fung, et al., 2003) is SVM Light (T. Joachims, 1999). SVM Light is an implementation of an SVM learner which addresses the problem of large tasks (T. Joachims, 1999). The optimization algorithms used in SVM Light are described in T. Joachims (2002). Another commonly used implementation of SVM is LIBSVM (Chang & Lin, 2011). LIBSVM implements an SMO-type (Sequential Minimal Optimization) algorithm proposed in a paper by Fan, Chen, and Lin (2005). Soni, et al. (2007) is using this implementation for prediction of stock price movements. Aside from the implementation, the input to the SVM is rather unique in the work of Soni, et al. (2007). They take the coordinates of the news items in a visualized document-map as features. A document-map is a low-dimensional space in which each

news item is positioned on the weighted average of the coordinates of the concepts that occur in the news item. SVMs can be extended to nonlinear classifiers by applying kernel mapping (kernel trick). As a result of applying the kernel mapping, the original classification problem is transformed into a higher dimensional space. SVMs which represent linear classifiers in this high-dimensional space may correspond to nonlinear classifiers in the original feature space (Burges, 1998). The kernel function used may influence the performance of the classifier. Zhai, et al. (2007) are using SVM with a Gaussian RBF kernel and a polynomial kernel.

B) ***Regression algorithms:*** They take on different forms in the research efforts as listed in table 4. One approach is ***Support Vector Regression (SVR)*** which is a regression based variation of SVM (Drucker, Burges, Kaufman, Smola, & Vapnik, 1997). It is utilized by Hagenau, et al. (2013) and Schumaker, et al. (2012).

Hagenau, et al. (2013) use SVM primarily but they also assess the ability to predict the discrete value of the stock return using SVR. They predict returns and calculate the R2 (squared correlation coefficient) between predicted and actually observed return. The optimization behind the SVR is very similar to the SVM, but instead of a binary measure (i.e., positive or negative), it is trained on actually observed returns. While a binary measure can only be 'true' or 'false', this measure gives more weight to greater deviations between actual and predicted returns than to smaller ones. As profits or losses are higher with greater deviations, this measure better captures actual trading returns to be realized (Hagenau, et al., 2013).

Schumaker, et al. (2012) choose to implement the SVR Sequential Minimal Optimization (Platt, 1999) function through Weka (Witten & Frank, 2005). This function allows discrete numeric prediction instead of classification. They select a linear kernel and ten-fold cross-validation.

Sometimes linear regression models are directly used (Chatrath, et al., 2014; Jin, et al., 2013; Paul C. Tetlock, et al., 2008). Paul C. Tetlock, et al. (2008) use OLS (Ordinary Least Square) method for estimating the unknown parameters in the linear regression model. They use two different dependent variables (raw and abnormal next-day returns) regressed on different negative words measures. Their main result is that negative words in firm-specific news stories robustly predict slightly lower returns on the following trading day.

Chatrath, et al. (2014) use a stepwise multivariate regression in a Probit (Probability unit) model. The purpose of the model is to estimate the probability that an observation with particular characteristics will fall into a specific category. In this case it is to ascertain the probability of news releases that result in jumps. Jin, et al. (2013) apply topic clustering methods and use customized sentiment dictionaries to uncover sentiment trends by analyzing relevant sentences. A linear regression model estimates the weight for each topic and makes currency forecasts.

C) ***Naïve Bayes:*** It is the next algorithm used in a group of works in table 4. It is probably the oldest classification algorithm (Lewis, 1998). But it is still very popular and is used among many of the works (Groth & Muntermann, 2011; F. Li, 2010; Werner & Myrray Z., 2004; Wuthrich, et al., 1998). It is based on the Bayes Theorem and it is called naïve because it is based on the naïve assumption of complete independence between text features. It differentiates itself from approaches such as k-Nearest Neighbours (k-NN), Artificial Neural Networks (ANN), or Support Vector Machine (SVM) in that it builds upon probabilities (of a feature belonging to a certain category) whereas the other mentioned approaches interpret the document feature-matrix spatially.

Y. Yu, et al. (2013) apply the Naïve Bayes (NB) algorithm to conduct sentiment analysis to examine the effect of multiple sources of social media along with the effect of conventional media and to investigate their relative importance and their interrelatedness. F. Li (2010)

uses Naïve Bayes to examine the information content of the forward-looking statements in the Management Discussion and Analysis section of company filings. He uses the Niave Bayes module in the Perl programming language to conduct the computation.

D) ***Decision rules and trees:*** It is the next group of algorithms used in the literature as indicated in table 4. A few of the researchers have made an effort to create rule-based classification systems (C.-J. Huang, et al., 2010; Peramunetilleke & Wong, 2002; Vu, et al., 2012).

Peramunetilleke and Wong (2002) use a set of keywords provided by a financial domain expert . The classifier expressing the correlation between the keywords and one of the outcomes is a rule set. Each of the three rule sets (DOLLAR_UP, DOLLAR_STEADY, DOLLAR_DOWN) yields a probability saying how likely the respective event will occur in relation to available keywords.

C.-J. Huang, et al. (2010) observe that the combination of two or more keywords in a financial news headline might play a crucial role on the next trading day. They thus applied weighted association rules algorithm to detect the important compound terms in the news headlines.

Rachlin, et al. (2007) use a Decision Tree Induction algorithm, which doesn't assume attribute independence. The algorithm is C4.5 developed by Quinlan (Quinlan, 1993). This algorithm yields a set of trend predicting rules. They further show the effect of the combination between numerical and textual data. Vu, et al. (2012) also use C4.5 decision tree for the text binary classification problem for predicting the daily up and down changes in stock prices.

For decision rule categorizers, the rules are composed of words, and words have meaning, the rules themselves can be insightful. More than just attempting to assign a label, a set of decision rules may summarize how to make decisions. For example, the rules may suggest a

pattern of words found in newswires prior to the rise of a stock price. The downside of rules is that they can be less predictive if the underlying concept is complex (Weiss, Indurkhya, & Zhang, 2010). Although decision rules can be particularly satisfying solutions for text mining, the procedures for finding them are more complicated than other methods (Weiss, et al., 2010).

Decision trees are special decision rules that are organized into a tree structure. A decision tree divides the document space into non-overlapping regions at its leaves, and predictions are made at each leaf (Weiss, et al., 2010).

E) Combinatory algorithms: In table 4, it is referring to a class of algorithms which are composed of a number of machine learning algorithms stacked or grouped together. Das and Chen (2007) have combined multiple classification algorithms together by a voting system to extract investor sentiment. The algorithms are namely, Naive Classifier, Vector Distance Classifier, Discriminant-Based Classifier, Adjective-Adverb Phrase Classifier, Bayesian Classifier. Accuracy levels turn out to be similar to widely used Bayes classifiers, but false positives are lower and sentiment accuracy higher.

Mahajan, et al. (2008) identify and characterize major events that impact the market using a Latent Dirichlet Allocation (LDA) based topic extraction mechanism. Then a stacked classifier is used which is a trainable classifier that combines the predictions of multiple classifiers via a generalized voting procedure. The voting step is a separate classification problem. They use a decision tree based on information gain for handling numerical attributes in conjunction with an SVM with sigmoid kernel to design the stacked classifier. The average accuracy of the classification system is 60%.

Butler and Kešelj (2009) propose 2 methods and then combine them to achieve best performance. The first method is based on character n-gram profiles, which are generated

for each company annual report, and then labeled based on the Common N-Gram (CNG) classification. The second method combines readability scores with performance inputs and then supplies them to a support vector machine (SVM) for classification. The combined version is setup to only make decisions when the models agreed.

Bollen and Huina (2011) deploy a self-organizing fuzzy neural network (SOFNN) model to test the hypothesis that including public mood measurements can improve the accuracy of Dow Jones Industrial Average (DJIA) prediction models. A fuzzy neural network is a learning machine that finds the parameters of a fuzzy system (i.e., fuzzy sets, fuzzy rules) by exploiting approximation techniques from neural networks. Therefore it is classified as a combinatory algorithm in table 4.

F) Multi-algorithm experiments: It is another class of works in table 4, whereby the same experiments are conducted using a number of different algorithms.

Wuthrich, et al. (1998) is one of the earliest works of research in this area. They do not stack multiple algorithms together to form a bigger algorithm. However, they conduct their experiments using multiple algorithms and compare the results.

Werner and Myrray Z. (2004) also carry out all tests using two algorithms, Naïve Bayes and SVM. Furthermore, Groth and Muntermann (2011) employ Naïve Bayes, k-Nearest Neighbour (k-NN), Artificial Neural Networks (ANN), and SVM in order to detect patterns in the textual data that could explain increased risk exposure in stock markets.

In conclusion, SVM has been extensively and successfully used as a textual classification and sentiment learning approach while some other approaches like Artificial Neural Networks (ANN), K-nearest neighbours (k-NN) have rarely been considered in the text mining literature for market prediction. This is also confirmed by Moraes, Valiati, and Gavião Neto (2013). Their research presents an empirical comparison between SVM and ANN regarding document-level sentiment

analysis. Their results show that ANN can produce superior or at least comparable results to SVM's. Such results can provide grounds for looking into usage of other algorithms than the currently mostly used SVM. C. H. Li, Yang, and Park (2012) demonstrate high performance of k-NN for text categorization as well as Jiang, Pang, Wu, and Kuang (2012). Tan, Wang, and Wu (2011) claim that for document categorization, centroid classifier performs slightly better than SVM classifier and beats it in running time too. Gradojevic and Gençay (2013) present a rare piece of research that uses fuzzy logic to improve technical trading signals successfully on the EUR-USD exchange rates but fuzzy logic is rarely used in the reviewed works for marker prediction based on text mining. Only Bollen and Huina (2011) use it in combination with neural networks to devise a self-organizing fuzzy neural network (SOFNN) successfully. However, Loia and Senatore (2014) show that it can be very useful for emotion modelling in general. Exploration of such under-researched algorithms in the context of market-prediction may lead to new insights that may be of interest for future researchers.

In order to better comprehend the reviewed systems in which the machine learning algorithms have been used, an additional number of system properties have been reviewed in table 4 which are explained in the following sections.

3.4.2. Training vs. testing volume and sampling

In this column, in table 4, two aspects are summarized if the information were available. Firstly, the volume of the examples which were used for training versus the volume used for testing; around 70 or 80 percent for training vs. 30 or 20 percent for testing seems to be the norm. Secondly, if the sampling for training and testing was of a special kind; what is specially of interest here is to know if a linear sampling have been followed as in essence the samples are on a time-series. Some of the works have clearly mentioned the sampling type like Stratified (Groth & Muntermann, 2011) whereas most of the others surprisingly have not mentioned anything at all.

3.4.3. Sliding Window

The overall objective of the reviewed systems is to predict the market movement in a future time-window (prediction-window) based on the learning gained in a past time-window (training-window) where the machine learning algorithms are trained and patterns are recognized.

For example, a system may learn patterns based on the available data for multiple days (training-window) in order to predict the market movement on a new day (prediction-window). The length and location of the training-window on the timeline may have two possible formats: *fixed* or *sliding*. If the training window is fixed, the system learns based on the data available from point '*A'* to point '*B'* on the timeline and those 2 points are fixed; for instance, from date '*A'* to date *'B'*. In such a scenario the resulted learning from the training window is applied to the prediction-window regardless of where on the timeline the prediction-window is located. It may be right after the training-window or it may be farther into the future with a distance from the training-window. It is obvious that if there is a big distance between the training-window and the prediction-window the learning captured in the machine learning algorithm may not be up-to-date and therefore accurate enough because the information available in the gap is not used for training.

Hence, a second format is introduced to solve the above problem whereby the entire training-window or one side of it (the side at the end) is capable of dynamically sliding up to the point where the prediction-window starts. In other words, if the training window starts at point 'A' and ends at point 'B' and the prediction-window starts at point 'C' and ends at point 'D'. In the sliding-window format, the system always ensures that point 'B' is always right before and adjacent to point 'C'. This approach is simply referred to in this work as "sliding window". The reviewed works which do possess a *sliding window* as a property of their system design are identified and marked in table 4 under a column with the same name.

Although, it intuitively seems necessary to implement a sliding-window, there are very few of the reviewed works which actually have (Butler & Kešelj, 2009; Jin, et al., 2013; Peramunetilleke &

Wong, 2002; Paul C. Tetlock, et al., 2008; Wuthrich, et al., 1998). This seems to be an aspect that can receive more attention in the future systems.

3.4.4. Semantics and Syntax

In Natural Language Processing (NLP) two aspects of language are attentively researched: Semantics and Syntax. Simply put: *semantics* deals with the meaning of words and *syntax* deals with their order and relative positioning or grouping. In this section: Firstly, a closer look is cast at the significance of each of them and some of the recent related works of research. Secondly, it is reported if and how they have been observed in the reviewed text-mining works for market-prediction.

Tackling semantics is an important issue and research efforts are occupied with it in a number of fronts. It is important to develop specialized ontologies for specific contexts like finance; Lupiani-Ruiz, et al. (2011) present a financial news semantic search engine based on Semantic Web technologies. The search engine is accompanied by an ontology population tool that assists in keeping the financial ontology up-to-date. Furthermore, semantics can be included in design of feature-weighting schemes; Luo, Chen, and Xiong (2011) 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). Also, in their work the weight of a term is correlated to its semantic similarity with a category. WordNet (Miller, 1995) provides a semantic network that links the senses of words to each other. The main types of relations among WordNet synsets are the super-subordinate relations that are hyperonymy and hyponymy. Other relations are the meronymy and the holonymy (Loia & Senatore, 2014). It is critical to facilitate semantic relations of terms for getting a satisfactory result in the text categorization; C. H. Li, et al. (2012) show a high performance of text categorization in which semantic relations of terms drawing upon two kinds of thesauri, a corpus-based thesaurus (CBT) and WordNet (WN), were sought. When a combination of CBT and WN was used, they obtained the highest level of performance in the text categorization.

Syntax is also very important and proper observation and utilization of it along with semantics (or sometimes instead of it) can improve textual classification accuracy; Kim, et al. (2014) propose a novel kernel, called language independent semantic (LIS) kernel, which is able to effectively compute the similarity between short-text documents without using grammatical tags and lexical databases. From the experiment results on English and Korean datasets, it is shown that the LIS kernel has better performance than several existing kernels. This is essentially a syntax-based pattern extraction method. It is interesting to note that there are several approaches to such syntax-based pattern-recognition methods: In the *word occurrence method*, a pattern is considered as a word that appears in a document (Thorsten Joachims, 1998). In the *word sequence method*, it consists of a set of consecutive words that appear in a document (Lodhi, Saunders, Shawe-Taylor, Cristianini, & Watkins, 2002). In the *parse-tree method* which is based on syntactic structure, a pattern is extracted from the tree of a document by considering not only the word occurrence but also the word sequence in the document (Collins & Duffy, 2001).

Duric and Song (2012) propose a set of new feature-selection schemes that use a Content and Syntax model to automatically learn a set of features in a review document by separating the entities that are being reviewed from the subjective expressions that describe those entities in terms of polarities. The results obtained from using these features in a maximum entropy classifier are competitive with the state-of-the-art machine learning approaches (Duric & Song, 2012). Topic models such as Latent Dirichlet Allocation (LDA) are generative models that allow documents to be explained by unobserved (latent) topics. The Hidden Markov Model LDA (HMM-LDA) (Griffiths, Steyvers, Blei, & Tenenbaum, 2005) is a topic model that simultaneously models topics and syntactic structures in a collection of documents. The idea behind the model is that a typical word can play different roles. It can either be part of the content and serve in a semantic (topical) purpose or it can be used as part of the grammatical (syntactic) structure. It can also be used in both contexts (Duric & Song, 2012). HMM-LDA models this behaviour by inducing syntactic classes for each word based on how they appear together in a sentence using a Hidden Markov Model. Each word gets assigned to a

syntactic class, but one class is reserved for the semantic words. Words in this class behave as they would in a regular LDA topic model, participating in different topics and having certain probabilities of appearing in a document (Duric & Song, 2012).

In table 4, there is one column dedicated to each of these aspects, namely: Semantics and Syntax. About half of the systems are utilizing some semantic aspect into their text mining approach which is usually done by using a dictionary or thesaurus and categorizing the words based on their meaning but none is advanced in the directions pointed out above. Moreover, very few works have made an effort to include Syntax i.e. order and role of words. These basic and somewhat indirect approaches are noun-phrases (Schumaker & Chen, 2009), word-combinations and n-grams (Hagenau, et al., 2013) and simultaneous appearance of words (C.-J. Huang, et al., 2010) and the "triplets" which consist of an adjective or adverb and the two words immediately following or preceding them (Das & Chen, 2007). Some works like Vu, et al. (2012) include Part of Speech (POS) tagging as a form of attention to syntax. Loia and Senatore (2014) achieve phrase-level sentiment analysis by taking into account four syntactic categories, namely: nouns, verbs, adverbs and adjectives. The need of deep syntactic analysis for the phrase-level sentiment-analysis has been investigated by Kanayama and Nasukawa (2008).

3.4.5. Combining news and technical data or signals

It is possible to pass technical data or signals along with the text features into the classification algorithm as additional independent variables. Examples for technical data could be a price or index level at a given time. Technical signals are the outputs of technical algorithms or rules like the moving average, relative strength rules, filter rules and the trading range breakout rules. Few of the researchers have taken advantage of these additional inputs as indicated in table 4 (Butler & Kešelj, 2009; Hagenau, et al., 2013; Rachlin, et al., 2007; Schumaker & Chen, 2009; Schumaker, et al., 2012; Zhai, et al., 2007).

3.4.6. Used software

Lastly it is interesting to observe what some of the common third-party applications are that are used for the implementation of the systems. In this column of table 4 the reader can see the names of the software pieces which were used by the reviewed works as a part of their pre-processing or machine learning. They are mostly dictionaries (F. Li, 2010; Paul C. Tetlock, et al., 2008), classification algorithm implementations (Butler & Kešelj, 2009; Werner & Myrray Z., 2004), concept extraction and word combination packages (Das & Chen, 2007; Rachlin, et al., 2007) or sentiment value providers (Schumaker, et al., 2012).

Reference	Algorithm Type	Algorithm Details	Training vs. testing volume and sampling	Sliding Window	Sema -ntics	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	SentiWord Net
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	OpinionFin der
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	Not applicable	No	No	No	No	Not mentioned

		Model						
Paul C. 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
Y. Yu, et al. (2013)	***Naïve Bayes***	Naïve Bayes	Not mentioned	No	Yes	No	No	Open-source Natural Language Toolkit (NLTK)
F. 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 Sept 12:00 to 27 Sept 9:00 vs. 9:00 to 10:00 on 27 Sept	Yes	Yes	No	No	Not mentioned
C.-J. Huang, et al. (2010)		Weighted association rules	2005 Jun to 2005 Oct vs. 2005 Nov	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 - Dec 07 vs. Jan 08 – Apr 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 Feb to 28 Nov vs. 1 to 19 Dec 2008	No	N/A	N/A	No	GPOMS, OpinionFinder
Wuthrich, et al. (1998)	***Multi-algorithm***	k-NN, ANNs, naïve Bayes,	Last 100 training days	Yes	Yes	No	No	Not mentioned

	experiments	rule-based	to forecast 1 day					
Werner and Myrray Z. (2004)		Naïve Bayes, SVM	1,000 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

Table 4 Classification algorithms and other machine learning aspects

3.5. Findings of the reviewed works

In table 5, the evaluation mechanisms have been looked at as well as the new findings for each piece of research.

Most of the works are presenting a confusion matrix or parts thereof to present their results. And calculate accuracy, recall or precision and sometimes the F-measure, with accuracy being the most common. The accuracy in majority of the cases is reported in the range of 50 to 70 percent, while arguing for better than chance results which is estimated at 50 percent (Butler & Kešelj, 2009; F. Li, 2010; Mahajan, et al., 2008; Schumaker & Chen, 2009; Schumaker, et al., 2012; Zhai, et al., 2007). It is a common evaluation approach and results above 55% have been considered report-worthy in other parts of the literature as well (Garcke, Gerstner, & Griebel, 2013). However, what makes most of the results questionable is that the majority of them surprisingly have not examined or reported if their experiment data is imbalanced or not. As this is important in data-mining (Duman, Ekinci, & Tanrıverdi, 2012; Thammasiri, Delen, Meesad, & Kasap, 2014) an additional column has been placed in table 5 to check for this. Among the reviewed works only Soni, et al. (2007), Mittermayer (2004) and Peramunetilleke and Wong (2002) have paid some attention to this topic in their works. It is also crucial to note if an imbalanced dataset with imbalanced classes is encountered specially with a high dimensionality in the feature-space, devising a suitable feature-selection that can appropriately deal with both the imbalanced-data and the high dimensionality becomes critical. Feature-selection for high-dimensional imbalanced data is amplified in detail in the work of Yin, Ge, Xiao, Wang, and Quan (2013). Liu, Loh, and Sun (2009) tackle the problem of imbalanced textual data using a simple probability based term weighting scheme to better distinguish documents in minor categories.

Smales (2012) examine the relationship between order imbalance and macroeconomic news in the context of Australian interest rate futures market and identify nine major macroeconomic announcements with impact on order imbalance.

Another popular evaluation approach in addition to the above for half of the reviewed works is the assembly of a trading strategy or engine (Groth & Muntermann, 2011; Hagenau, et al., 2013; C.-J. Huang, et al., 2010; Mittermayer, 2004; Pui Cheong Fung, et al., 2003; Rachlin, et al., 2007; Schumaker & Chen, 2009; Schumaker, et al., 2012; Paul C. Tetlock, et al., 2008; Zhai, et al., 2007). Through which a trading period is simulated and profits are measured to evaluate the viability of the system.

In general researchers are using evaluation mechanisms and experimental data that widely vary and this makes an objective comparison in terms of concrete levels of effectiveness unreachable.

Reference	Findings	Trading Strategy	Balanced data
Wuthrich, et al. (1998)	Ftse 42%, Nky 47%, Dow 40% Hsi 53% and Sti 40%.	Yes	Not mentioned
Peramunetilleke and Wong (2002)	Better than chance	No	Yes
Pui Cheong Fung, et al. (2003)	The cumulative profit of monitoring multiple time series is nearly double to that of monitoring single time series.	Yes	Not mentioned
Werner and Myrray Z. (2004)	Evidence that the stock messages help predict market volatility, but not stock returns.	No	No
Mittermayer (2004)	Average profit 11% compared to average profit by random trader 0%	Yes	Yes
Das and Chen (2007)	Regression has low explanatory power	No	Not mentioned
Soni, et al. (2007)	Hit rate of classifier: 56.2% compared to 47.5% for naïve classifier and 49.1% SVM bag-of-words	No	Yes (roughly)
Zhai, et al. (2007)	Price 58.8%, Direct news 62.5%, Indirect news 50.0%, Combined news 64.7%, Price & News 70.1% Profit: for Price & News 5.1% in 2 moths and for Price and News alone around half of it each	Yes	Not mentioned
Rachlin, et al. (2007)	Cannot improve the predictive accuracy of the numeric analysis. Accuracy 82.4% for join textual and numeric analysis and 80.6% for textual analysis, and 83.3% numeric alone.	Yes	Not mentioned

Paul C. Tetlock, et al. (2008)	1) the fraction of negative words in firm-specific news stories forecasts low firm earnings; 2) firms' stock prices briefly under-react to the information embedded in negative words; and 3) the earnings and return predictability from negative words is largest for the stories that focus on fundamentals.	Yes	Not relevant
Mahajan, et al. (2008)	Accuracy 60%	No	Not mentioned
Butler and Kešelj (2009)	First method: 55% and 59% for character-grams and word-grams accuracy respectively, still superior to the benchmark portfolio. Second method: overall accuracy and over-performance precision was 62.81% and 67.80% respectively.	No	Not mentioned
Schumaker and Chen (2009)	Directional Accuracy 57.1% , Return 2.06% , Closeness 0.04261	Yes	Not mentioned
F. Li (2010)	Accuracy for tone 67% and content 63% with naïve Bayes and less than 50% with dictionary-based.	No	No
C.-J. Huang, et al. (2010)	Prediction accuracy and the recall rate up to 85.2689% and 75.3782% in average, respectively.	Yes	Not mentioned
Groth and Muntermann (2011)	Accuracy (slightly) above the 75% guessing equivalent benchmark.	Yes	No
Schumaker, et al. (2012)	Objective articles were performing poorly in Directional Accuracy versus Baseline. Neutral articles had poorer Trading Returns versus Baseline. Subjective articles performed better with 59.0% Directional Accuracy and a 3.30% Trading Return. Polarity performed poorly versus Baseline.	Yes	Not mentioned.
Lugmayr and Gossen (2012)	In progress	No	Not mentioned
Y. Yu, et al. (2013)	Polarity with 79% accuracy and 0.86 F-measure on the test set. Only total number of social media counts has a significant positive relationship with risk, but not with return. It is shown that the interaction term has a marginally negative relationship with return, but a highly negative significant relationship with risk.	No	Not mentioned
Hagenau, et al. (2013)	Feedback-based feature selection combined with 2-word combinations achieved accuracies of up to 76%	Yes	No
Jin, et al. (2013)	Precision around 0.28 on average.	No	Not mentioned
Chatrath, et al. (2014)	(a) jumps are a good proxy for news arrival in currency markets; (b) there is a systematic reaction of currency prices to economic surprises; and (c) prices respond quickly within 5-minutes of the news release	No	Not mentioned
Bollen and Huina (2011)	The mood 'Calm' had the highest Granger causality relation with the DJIA for time lags ranging from two to six days (p-values < 0.05). The other four GPOMS mood dimensions and OpinionFinder didn't have a significant correlation with stock market changes.	No	Not mentioned

Vu, et al. (2012)	Combination of Previous days's price movement, Bullish/bearish and Pos_Neg features create a superior model in all 4 companies with accuracies of: 82.93%, 80.49%, 75.61% and 75.00% and for the online test as: 76.92%, 76.92%, 69.23% and 84.62%	NO	Not mentioned

Table 5 Findings of the reviewed works, existence of a trading strategy and balanced-data

4. Suggestions for future work

Market prediction mechanisms based on online text mining are just emerging to be investigated rigorously utilizing the radical peak of computational processing power and network speed in the recent times. We foresee this trend to continue. This research helps put into perspective the role of human reactions to events in the making of markets and can lead to a better understanding of market efficiencies and convergence via information absorption. In summary, this work identifies the below as areas or aspects in need of future research and advancement:

A. **Semantics:** As discussed in section 3.4.4, advancements of techniques in semantics are crucial to the text-classification problem at hand as text-mining researchers have already shown. However, such advancements have not yet entered into the field of market-predictive text-mining. Development of specialized ontologies by creating new ones or customization of current dictionaries like WordNet requires more attention. Many of the current works of research are still too much focused on word occurrence methods and they rarely even use WordNet. Moreover, semantic relations can be researched with different objectives, from defining weighting schemes for feature-representation to semantic compression or abstraction for feature-reduction. Probably since the market-predictive text-mining itself is an emerging field, the researchers are yet to dive into semantics enhancement for the market-predictive context specifically.

B. **Syntax:** Syntactic analysis techniques has received probably even less attention than semantic ones, as also discussed in section 3.4.4. More advanced syntax-based techniques like usage of parse-trees for pattern recognition in text can improve the quality of text-

mining significantly. This aspect requires the attention of future researchers too, starting with attempting to transfer some of the learning in other areas of text-mining like reviews classification into market-predictive text-mining.

C. **Sentiment:** Sentiment and emotion analysis has gained significance prominence in the field of text-mining due to the interest of governments and multi-national companies to maintain a finger on the pulse of the public mood to win elections in the case of the former or just surprise their customers by the amount of insight about their preferences for the latter. Interestingly, market-prediction is very closely related to the mood of public or market-participants as established by behavioural-economics. However, in case of the analysis of sentiment with regards to a product the anticipation of what a piece of text entails is far more straightforward than in the case of market-prediction. There are no secrets as to whether a product-review entails positive or negative emotions about it. However, even the best traders and investors can never be completely sure what market-reaction to expect as a result of a piece of news-text. Therefore, there is a lot of room for market-predictive sentiment investigation for future research.

D. **Text-mining component, textual-source or application-market specialization:** This work has learnt that the current works of research on market-predictive text-mining are rather holistic with one-off end-to-end systems. However, in the future, the text-mining process should be broken down into its critical components like feature-selection, feature-representation and feature-reduction and each of those needs to be specifically researched for the specialized context of market-prediction; some specific suggestions for each component has been made in sections 3.3.1, 3.3.2 and 3.3.3 of this work. Furthermore, market-predictive text-mining can also become even more specialized by focusing on a specific source of text e.g. a specific social media outlet or news-source or text-role like news-headlines vs. news-body etc. Moreover, there is a need for specialized research on each type of financial markets (stocks,

bond, commodity, money, futures, derivatives, insurance, forex) or on each geographical location.

E. **Machine Learning Algorithms:** In section 3.4 it has been explained thoroughly how SVM and Naïve Bayes are heavily favoured by researchers, probably due to their straightforwardness, while many other machine learning algorithms or techniques like Artificial Neural Networks (ANN), K-Nearest Neighbours (k-NN), fuzzy-logic, etc. show seriously promising potentials for textual-classification and sentiment-analysis elsewhere in the literature but have not yet been experimented with in the context of market-predictive text-mining or are significantly under-researched at this stage.

F. **Integration of technical signals:** Despite their practical popularity with market-traders, technical signals which are the outputs of technical algorithms or rules like the moving average, relative strength rules, filter rules and the trading range breakout rules, are almost always left out of the research that is based on text-mining as pointed out in section 3.4.5 of this work. It may be because of the fact that the researchers who are for prediction based on text-mining are most probably for fundamental-analysis approaches and therefore opposed to the technical-analysis approaches as explained in section 2.5. However, it is logically conceivable that hybrid-models based on the sum of the best of the two worlds (technical and fundamental) must produce even better results and this should be considered more vigorously in future research.

G. **Relation with behavioural-economics research:** As pointed out in section 2 of this work, there exists a substantial interdisciplinary nature to this field of research specially between economics and computer-science. Deepening economics comprehension is crucial for future researchers. Currently, economics and specially behavioural economics theories are referred to in the literature just superficially and only to establish that public-mood has an impact on markets. However, a deeper study of behavioural-economics should reveal principles and

learning whose parallel implementation in text-mining by computer-scientists may lead to true breakthroughs. This direction, although somewhat immature or vague at this stage, is highly encouraged by the researchers of this work.

H. **Availability and quality of experimental datasets:** one of the major challenges observed is the unavailability of highly standardized datasets that contain mappings of text onto markets for certain periods of times that researchers can use for assimilation of their experimentation and evaluation efforts. In the available work, most researchers have attempted to accumulate their own datasets. This has naturally resulted in fragmented dataset-formats and contents and a lack of adequate observation for critical characteristics in datasets. For example, as indicated in table 5, most works have not observed if their experimental data is imbalanced in any way which may have favoured their results. Future researchers are encouraged to standardize and release datasets for experimentation in market-predictive text-mining. Currently, the predominant standard datasets circling around in text-mining works are of movie-reviews which are not appropriate for this work. Wu and Tan (2011) present an intriguing piece of research on transferring sentiment domain knowledge from one domain to another by building a framework between them that acts as a bridge between the source domain and the target domain. This may inspire some new thoughts in this area too.

I. **Evaluation methods:** much like the experimental datasets, evaluation methods as a whole are highly subjective. Most researchers are still comparing their results with the probabilities of chance and not so much with each others' works. As stated in section 3.5, researchers are generally using evaluation mechanisms that widely vary; which make an objective comparative performance-evaluation virtually impossible. Therefore, future researchers could focus on such standardization initiatives as main objectives of their research in this field as market-predictive text-mining is here to stay.

5. Conclusion

The major systems for market prediction based on online text mining have been reviewed and some of the predominant gaps that exist within them have been identified. The review was conducted on three major aspects, namely: pre-processing, machine learning and the evaluation mechanism; with each breaking down into multiple sub-discussions. It is believed to be the first effort to provide a comprehensive review from a holistic and interdisciplinary point of view. This work intended to accomplish: Firstly, facilitation of integration of research activities from different fields on the topic of market prediction based on online text mining; Secondly, provision of a study-framework to isolate the problem or different aspects of it in order to clarify the path for further improvement; Thirdly, submission of directional and theoretical suggestions for future research.

Advancements in the field of market-predictive text-mining can have the following implications in particular among many:

1- Investment banks and financial institutions as well as brokerage firms who are investing and trading in financial markets can use specialized market-trend analysis and prediction systems that are developed using the insights gained in such targeted text-mining research efforts as this work. The existence of such intelligent systems for those institutions assists with making better financial decisions which lead to considerable financial returns on their investments and avoidance of severe losses.

2- In today's global economy, even more sophisticated insights into the financial markets is needed, because the lack thereof, as we have recently witnessed during the 2008 financial crisis, can negatively impact the livelihoods of millions of people around the world. Therefore it becomes imperative to pursue research in the field of market-predictive text-mining as a viable solution that may bring about a much higher degree of confidence on comprehension of market-movements based on gaining insights into human psychology at a

macro level through text-mining of the, now widely available, textual resources on the Internet at a virtually real-time pace.

3- With the staggering amount of textual information available online, about every aspect of every conceivable topic, the necessity to develop specialized text-mining systems rapidly emerges. Such systems are highly targeted at a specific application-area and a certain type of text among too many possible alternatives. A focus on market-predictive text-mining n this research helps the formation of this emerging field as a recognizable and independent field that can be delved into vigorously and not only in shadow of general text-mining research. The formation of such independent field of research for market-predictive text-mining, distinct from product-review sentiment analysis or such, is a hopeful implication of this work.

This work is hoped to help other researchers put the various ideas in this field into perspective more conveniently and become able to make strategic decisions upfront in the design of the future systems.

Acknowledgements

The support of University of Malaya in production of this paper is appreciated.

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 Syst. Appl., 36*, 12001-12011.

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

Balahur, A., Steinberger, R., Goot, E. v. d., Pouliquen, B., & Kabadjov, M. (2009). Opinion Mining on Newspaper Quotations. In *Proceedings of the 2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology - Volume 03* (pp. 523-526): IEEE Computer Society.

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.

Bikas, E., Jurevičienė, D., Dubinskas, P., & Novickytė, L. (2013). Behavioural Finance: The Emergence and Development Trends. *Procedia - Social and Behavioral Sciences, 82*, 870-876.

Bollen, J., & Huina, M. (2011). Twitter Mood as a Stock Market Predictor. *Computer, 44*, 91-94.

Burges, C. C. (1998). A Tutorial on Support Vector Machines for Pattern Recognition. *Data Mining and Knowledge Discovery, 2*, 121-167.

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): Springer Berlin Heidelberg.

Cambria, E., Schuller, B., Yunqing, X., & Havasi, C. (2013). New Avenues in Opinion Mining and Sentiment Analysis. *Intelligent Systems, IEEE, 28*, 15-21.

Chang, C.-C., & Lin, C.-J. (2011). LIBSVM: A library for support vector machines. *ACM Trans. Intell. Syst. Technol., 2*, 1-27.

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.

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

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

Chordia, T., Roll, R., & Subrahmanyam, A. (2011). Recent trends in trading activity and market quality. *Journal of Financial Economics, 101*, 243-263.

Collins, M., & Duffy, N. (2001). Convolution Kernels for Natural Language. In (pp. 625-632): MIT Press.

Das, S. R., & Chen, M. Y. (2007). Yahoo! for Amazon: Sentiment Extraction from Small Talk on the Web. *Manage. Sci., 53*, 1375-1388.

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

Drucker, H., Burges, C. J. C., Kaufman, L., Smola, A., & Vapnik, V. (1997). Support Vector Regression Machines. In *ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS* (pp. 155-161): MIT Press.

Duman, E., Ekinci, Y., & Tanrıverdi, A. (2012). Comparing alternative classifiers for database marketing: The case of imbalanced datasets. *Expert Systems with Applications, 39*, 48-53.

Duric, A., & Song, F. (2012). Feature selection for sentiment analysis based on content and syntax models. *Decision Support Systems, 53*, 704-711.

Evans, C., Pappas, K., & Xhafa, F. (2013). Utilizing artificial neural networks and genetic algorithms to build an algo-trading model for intra-day foreign exchange speculation. *Mathematical and Computer Modelling, 58*, 1249-1266.

Fama, E. F. (1965). Random Walks in Stock Market Prices. *Financial Analysts Journal, 21*, 55-59.

Fama, E. F. (1970). Efficient Capital Markets: A Review of Theory and Empirical Work. *The Journal of Finance, 25*, 383-417.

Fan, R.-E., Chen, P.-H., & Lin, C.-J. (2005). Working Set Selection Using Second Order Information for Training Support Vector Machines. *J. Mach. Learn. Res., 6*, 1889-1918.

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.

Friesen, G., & Weller, P. A. (2006). Quantifying cognitive biases in analyst earnings forecasts. *Journal of Financial Markets, 9*, 333-365.

García, D., & Urošević, B. (2013). Noise and aggregation of information in large markets. *Journal of Financial Markets, 16*, 526-549.

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): Springer Berlin Heidelberg.

Ghazali, R., Hussain, A. J., & Liatsis, P. (2011). Dynamic Ridge Polynomial Neural Network: Forecasting the univariate non-stationary and stationary trading signals. *Expert Syst. Appl., 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.

Gradojevic, N., & Gençay, R. (2013). Fuzzy logic, trading uncertainty and technical trading. *Journal of Banking & Finance, 37*, 578-586.

Griffiths, T. L., Steyvers, M., Blei, D. M., & Tenenbaum, J. B. (2005). Integrating topics and syntax. In (pp. 537-544): MIT Press.

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

Haddi, E., Liu, X., & Shi, Y. (2013). The Role of Text Pre-processing in Sentiment Analysis. *Procedia Computer Science, 17*, 26-32.

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.

Hasbrouck, J., & Saar, G. (2013). Low-latency trading. *Journal of Financial Markets*.

Hsinchun, C., & Zimbra, D. (2010). AI and Opinion Mining. *Intelligent Systems, IEEE, 25*, 74-80.

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 Syst. Appl., 37*, 6409-6413.

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

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, Illinois, USA: ACM.

Joachims, T. (1998). Text categorization with Support Vector Machines: Learning with many relevant features. In C. Nédellec & C. Rouveirol (Eds.), *Machine Learning: ECML-98* (Vol. 1398, pp. 137-142): Springer Berlin Heidelberg.

Joachims, T. (1999). Making large-Scale {SVM} Learning Practical. In B. Schölkopf, C. Burges & A. Smola (Eds.), (pp. 169-184). Cambridge, MA: MIT Press.

Joachims, T. (2002). *Learning to Classify Text Using Support Vector Machines -- Methods, Theory, and Algorithms*: Kluwer/Springer.

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

Kanayama, H., & Nasukawa, T. (2008). Textual demand analysis: detection of users' wants and needs from opinions. In *Proceedings of the 22nd International Conference on Computational Linguistics - Volume 1* (pp. 409-416). Manchester, United Kingdom: Association for Computational Linguistics.

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., Chung, B.-s., Choi, Y., Lee, S., Jung, J.-Y., & Park, J. (2014). Language independent semantic kernels for short-text classification. *Expert Systems with Applications, 41*, 735-743.

Kleinnijenhuis, J., Schultz, F., Oegema, D. & Atteveldt, W.H. van. (2013). Financial News and Market Panics in the age of high frequency trading algorithms. *Journalism, 14*.

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

Lewis, D. (1998). Naive (Bayes) at forty: The independence assumption in information retrieval. In C. Nédellec & C. Rouveirol (Eds.), *Machine Learning: ECML-98* (Vol. 1398, pp. 4-15): Springer Berlin Heidelberg.

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.

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, W., & Xu, H. (2014). Text-based emotion classification using emotion cause extraction. *Expert Systems with Applications, 41*, 1742-1749.

Liu, Y., Loh, H. T., & Sun, A. (2009). Imbalanced text classification: A term weighting approach. *Expert Systems with Applications, 36*, 690-701.

Lo, A. W. (2005). Reconciling Efficient Markets with Behavioral Finance: The Adaptive Markets Hypothesis. *Journal of Investment Consulting*.

Lodhi, H., Saunders, C., Shawe-Taylor, J., Cristianini, N., & Watkins, C. (2002). Text classification using string kernels. *J. Mach. Learn. Res., 2*, 419-444.

Loia, V., & Senatore, S. (2014). A fuzzy-oriented sentic analysis to capture the human emotion in Web-based content. *Knowledge-Based Systems, 58*, 75-85.

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 A. Lugmayr, T. Risse, B. Stockleben, J. Kaario, B. Pogorelc & E. Serral Asensio (Eds.), *SAME 2012 – 5th 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., & Camón-Herrero, J. B. (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 *Web Intelligence and Intelligent Agent Technology, 2008. WI-IAT '08. IEEE/WIC/ACM International Conference on* (Vol. 1, pp. 423-426).

Majumder, D. (2013). Towards an efficient stock market: Empirical evidence from the Indian market. *Journal of Policy Modeling, 35*, 572-587.

Maks, I., & Vossen, P. (2012). A lexicon model for deep sentiment analysis and opinion mining applications. *Decision Support Systems, 53*, 680-688.

Miller, G. A. (1995). WordNet: a lexical database for English. *Commun. ACM, 38*, 39-41.

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

Moraes, R., Valiati, J. F., & Gavião Neto, W. P. (2013). Document-level sentiment classification: An empirical comparison between SVM and ANN. *Expert Systems with Applications, 40*, 621-633.

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

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

Ortigosa-Hernández, J., Rodríguez, J. D., Alzate, L., Lucania, M., Inza, I., & Lozano, J. A. (2012). Approaching Sentiment Analysis by using semi-supervised learning of multi-dimensional classifiers. *Neurocomputing, 92*, 98-115.

Peramunetilleke, D., & Wong, R. K. (2002). Currency exchange rate forecasting from news headlines. *Aust. Comput. Sci. Commun., 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.

Platt, J. C. (1999). Fast training of support vector machines using sequential minimal optimization. In *Advances in kernel methods* (pp. 185-208): MIT Press.

Potì, V., & Siddique, A. (2013). What drives currency predictability? *Journal of International Money and Finance, 36*, 86-106.

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 *Computational Intelligence for Financial Engineering, 2003. Proceedings. 2003 IEEE International Conference on* (pp. 395-402).

Quinlan, J. R. (1993). *C4.5: programs for machine learning*: Morgan Kaufmann Publishers Inc.

Rachlin, G., Last, M., Alberg, D., & Kandel, A. (2007). ADMIRAL: A Data Mining Based Financial Trading System. In *Computational Intelligence and Data Mining, 2007. CIDM 2007. IEEE Symposium on* (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.

Robertson, C., Geva, S., & Wolff, R. (2006). What types of events provide the strongest evidence that the stock market is affected by company specific news? In *Proceedings of the fifth Australasian conference on Data mining and analystics - Volume 61* (pp. 145-153). Sydney, Australia: Australian Computer Society, Inc.

Salzberg, S. (1997). On Comparing Classifiers: Pitfalls to Avoid and a Recommended Approach. *Data Mining and Knowledge Discovery, 1*, 317-328.

Sankaraguruswamy, S., Shen, J., & Yamada, T. (2013). The relationship between the frequency of news release and the information asymmetry: The role of uninformed trading. *Journal of Banking & Finance, 37*, 4134-4143.

Schumaker, R. P., & Chen, H. (2009). Textual analysis of stock market prediction using breaking financial news: The AZFin text system. *ACM Trans. Inf. Syst., 27*, 1-19.

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

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.

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, Supplement 1*, 131-135.

Sidorov, G., Velasquez, F., Stamatatos, E., Gelbukh, A., & Chanona-Hernández, L. (2013). Syntactic N-grams as machine learning features for natural language processing. *Expert Systems with Applications*.

Smales, L. A. (2012). Order imbalance, market returns and macroeconomic news: Evidence from the Australian interest rate futures market. *Research in International Business and Finance, 26*, 410-427.

Soni, A., van Eck, N. J., & Kaymak, U. (2007). Prediction of Stock Price Movements Based on Concept Map Information. In *Computational Intelligence in Multicriteria Decision Making, IEEE Symposium on* (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.

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

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.

Thammasiri, D., Delen, D., Meesad, P., & Kasap, N. (2014). A critical assessment of imbalanced class distribution problem: The case of predicting freshmen student attrition. *Expert Systems with Applications, 41*, 321-330.

Tomer, J. F. (2007). What is behavioral economics? *The Journal of Socio-Economics, 36*, 463-479.

Tsai, C.-F., Eberle, W., & Chu, C.-Y. (2013). Genetic algorithms in feature and instance selection. *Knowledge-Based Systems, 39*, 240-247.

Urquhart, A., & Hudson, R. (2013). Efficient or adaptive markets? Evidence from major stock markets using very long run historic data. *International Review of Financial Analysis, 28*, 130-142.

Uysal, A. K., & Gunal, S. (2012). A novel probabilistic feature selection method for text classification. *Knowledge-Based Systems, 36*, 226-235.

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.

Wang, G.-J., Xie, C., & Han, F. (2012). Multi-Scale Approximate Entropy Analysis of Foreign Exchange Markets Efficiency. *Systems Engineering Procedia, 3*, 201-208.

Weiss, S. M., Indurkhya, N., & Zhang, T. (2010). *Fundamentals of Predictive Text Mining*.

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.

Wisniewski, T. P., & Lambe, B. (2013). The role of media in the credit crunch: The case of the banking sector. *Journal of Economic Behavior & Organization, 85*, 163-175.

Witten, I. H., & Frank, E. (2005). *Data Mining: Practical Machine Learning Tools and Techniques, Second Edition (Morgan Kaufmann Series in Data Management Systems)*: Morgan Kaufmann Publishers Inc.

Wu, Q., & Tan, S. (2011). A two-stage framework for cross-domain sentiment classification. *Expert Systems with Applications, 38*, 14269-14275.

Wuthrich, B., Cho, V., Leung, S., Permunetilleke, D., Sankaran, K., & Zhang, J. (1998). Daily stock market forecast from textual web data. In *Systems, Man, and Cybernetics, 1998. 1998 IEEE International Conference on* (Vol. 3, pp. 2720-2725 vol.2723).

Yang, L., Li, C., Ding, Q., & Li, L. (2013). Combining Lexical and Semantic Features for Short Text Classification. *Procedia Computer Science, 22*, 78-86.

Yin, L., Ge, Y., Xiao, K., Wang, X., & Quan, X. (2013). Feature selection for high-dimensional imbalanced data. *Neurocomputing, 105*, 3-11.

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.

Yu, L.-C., Wu, J.-L., Chang, P.-C., & Chu, H.-S. (2013). Using a contextual entropy model to expand emotion words and their intensity for the sentiment classification of stock market news. *Knowledge-Based Systems*.

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*.

Zhai, Y., Hsu, A., & Halgamuge, S. K. (2007). Combining News and Technical Indicators in Daily Stock Price Trends Prediction. In *Proceedings of the 4th international symposium on Neural Networks: Advances in Neural Networks, Part III* (pp. 1087-1096). Nanjing, China: Springer-Verlag.

- Review of essential concepts for market prediction based on online text-mining.
- Review of the cutting-edge work in the literature.
- Identification of main differentiating factors among the available solutions.
- Observations on possible opportunities for future work.