Review

Text mining for market prediction: A systematic review

Arman Khadjeh Nassirtoussi a,⇑, Saeed Aghabozorgi a, Teh Ying Wah a, David Chek Ling Ngo b

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

Article history:
Available online 9 June 2014

Keywords:
Online sentiment analysis
Social media text mining
News sentiment analysis
FOREX market prediction
Stock prediction based on news

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 behavioral-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.

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 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. (see Fig. 1).

In this work a systematic review of the past works of research with significant contribution to the topic of online-text-mining for 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.

⇑ Corresponding author. Tel.: +49 1626230377.
E-mail address: armankhnt@gmail.com (A. Khadjeh Nassirtoussi).

http://dx.doi.org/10.1016/j.eswa.2014.06.009
0957-4174/© 2014 Elsevier Ltd. All rights reserved.
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 modeling 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 modeling and pattern recognition), behavioral-economics (to establish economic sense) (see Fig. 2).

The main premise is aligned with recent findings of behavioral-economic principles whereby market conditions are products of human behaviors 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 plausable 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. Behavioral-economics

Cognitive and behavioral 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).

Behavioral finance and investor sentiment theory have firmly established that investors’ behavior 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 and 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 behavior 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 behavior 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 modeling 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 Poti and 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 (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 (Yu, Nartea, et al., 2013). Hence, there is a need to continuously revisit the level of efficiency of economically dynamic and rapidly growing emerging markets (Yu, Nartea, 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 (Yu, Nartea, et al., 2013). Nevertheless, there are and continue to be many sophisticated financial prediction modeling efforts based on various types or combinations of machine learning algorithms like neural networks (Anastasakis & Mort, 2009; Ghazali, 2010). fuzzy logic (Bahrepour, Akbarzadeh-T, Yaghoobi, & Naghibi-S, 2011), Support Vector regression (Huang, Chuang, Wu, & Lai, 2010; Premanode & Toumazou, 2013), rule-based genetic network programming (Mabu, Hirawawa, Ohayashi, & Kuremoto, 2013).

However, there was 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 – 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 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 meaningful 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, Schultz, Oegema, & Atteweltd, 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, & Zimba, 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, & Kabadov, 2009; Cambria, Schuller, Yuning, & Havasi, 2013; Hsinchun Zimba, & Zimba, 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 (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. 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 Fig. 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.

#### 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, 2004), Financial Times (Wuthrich et al., 1998), Reuters (Pui Cheong Fung, Xu Yu, & Wai, 2003), Dow Jones, Bloomberg (Chattrath 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 (Huang, Liao, et al., 2010; Peramunetilleke & Wong, 2013; Chatrath et al., 2014; Jin et al., 2013).

![Fig. 3. Generic common system components diagram.](image-url)
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 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; Huang, Liao, et al., 2010; Huang, Chuang, 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.

### 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, 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 (Majahan, Dey, & Haque, 2008), S&P 500 (Schumaker & Chen, 2009) or the stock price of a specific company like BHP Billiton Ltd. (BHP.AX) (Zhai,
predictability. Possible to find less efficient currency pairs that are prone to trends has well explained such quantitative tradings (Chordia, Roll, & Subrahmanyam, 2005). Market efficiency (Chordia, Roll, & Subrahmanyam, 2005). Market to be looked at as a reasonable market convergence time to averaged. 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 ele-

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 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

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

<table>
<thead>
<tr>
<th>Reference</th>
<th>Market</th>
<th>Market index</th>
<th>Time-frame</th>
<th>Period</th>
<th>Forecast type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wurthrich et al. (1998)</td>
<td>Stocks</td>
<td>Dow Jones Industrial Average, the Nikkei 225, the Financial Times 100, the Hang Seng, and the Singapore Straits</td>
<td>Daily</td>
<td>6 December 1997–6 March 1998</td>
<td>Categorical: up, steady, down</td>
</tr>
<tr>
<td>Werner and Myrrat (2004)</td>
<td>Stocks</td>
<td>Dow Jones Industrial Average, and the Dow Jones Internet</td>
<td>Intraday(15-min, 1 h and 1 day)</td>
<td>Year 2000</td>
<td>Categorical: buy, sell, and hold messages</td>
</tr>
<tr>
<td>Rachlin et al. (2007)</td>
<td>Stocks</td>
<td>5 stocks from US NASDAQ</td>
<td>Days</td>
<td>7 February–7 May 2006</td>
<td>Categorical: up, slight-up, expected, slight-down, down</td>
</tr>
<tr>
<td>Butler and Kešelj (2009)</td>
<td>Stocks</td>
<td>1-Year market drift</td>
<td>Yearly</td>
<td></td>
<td>Categorical:</td>
</tr>
<tr>
<td>Li (2010)</td>
<td>Stocks</td>
<td>(1) Index (2) Quarterly earnings and cash flows (3) Stock returns</td>
<td>Annually (Quarterly with 3 dummy quarters)</td>
<td>1994–2007</td>
<td>Categorical based on tone: positive, negative, neutral, uncertain</td>
</tr>
<tr>
<td>Groot and Muntermann (2011)</td>
<td>Stocks</td>
<td>Abnormal risk exposure ARISKt (Thomson Reuters DataScope Tick History)</td>
<td>Intraday (volatility during the t = 15 and 30 min)</td>
<td>1 August 2003–31 July 2005</td>
<td>Categorical: positive, negative</td>
</tr>
<tr>
<td>Lugmayr and Gossen (2012)</td>
<td>Stocks</td>
<td>DAX 30 performance index</td>
<td>Intraday (3 or 4 times per day)</td>
<td>Not available</td>
<td>Categorical: sentiment [−1, 1]; trend (bear, bull, neutral); trend strength (in %)</td>
</tr>
<tr>
<td>Yu, Duan, et al. (2013)</td>
<td>Stocks</td>
<td>Abnormal returns and cumulative abnormal returns of 824 firms</td>
<td>Daily</td>
<td>1 July–30 September 2011</td>
<td>Categorical: positive or negative</td>
</tr>
<tr>
<td>Hagenau et al. (2013)</td>
<td>Stocks</td>
<td>Company specific</td>
<td>Daily</td>
<td>1997–2011</td>
<td>Categorical: positive or negative</td>
</tr>
<tr>
<td>Jin et al. (2013)</td>
<td>FOREX</td>
<td>Exchange rate</td>
<td>Daily</td>
<td>1 January–31 December 2012</td>
<td>Categorical: positive or negative</td>
</tr>
<tr>
<td>Chathrath et al. (2014)</td>
<td>FOREX</td>
<td>Exchange rate</td>
<td>Intraday(5 min)</td>
<td>January 2005–December 2010</td>
<td>Categorical: positive or negative jump</td>
</tr>
<tr>
<td>Bollen and Huina (2011)</td>
<td>Stock</td>
<td>DJIA</td>
<td>Daily</td>
<td>28 February–19 December 2008</td>
<td>Regression analysis</td>
</tr>
<tr>
<td>Vu et al. (2012)</td>
<td>Stock</td>
<td>Stock prices (at NASDAQ for AAPL, GOOG, MSFT, AMZN)</td>
<td>Daily</td>
<td>1 April 2011–31 May 2011, online test 8 September–26 September 2012</td>
<td>Categorical: up and down</td>
</tr>
</tbody>
</table>
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 min (Chattrath et al., 2014), 15 min (Werner & Myrray, 2004), 20 min (Schumaker & Chen, 2009) to 1, 2, or 3 h (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 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 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.

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 representa-tive format that is structured and can be processed by the machine. In data mining in general and text mining specifically the pre-pro cessing phase is of significant impact on the overall outcomes (Butler & Kešelj, 2009). Term extraction tools are used when dealing with text-mining-based market-prediction-problems as alluded to by Kleinnijenhuis et al. (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 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, 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, Stamatos, Gelbukh, & Channa-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-Aghaei, & 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 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, van Eck, & Kaymak, 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. Berk and Vajteršic (2013) introduce a detailed method for dimensionality reduction for text, based on parallel rare term vector replacement.
Table 3
Pre-processing: feature-selection, feature-reduction, feature-representation.

<table>
<thead>
<tr>
<th>Reference</th>
<th>Feature selection</th>
<th>Dimensionality reduction</th>
<th>Feature representation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wuthrich et al. (1998)</td>
<td>Bag-of-words</td>
<td>Pre-defined dictionaries (word-sequences by an expert)</td>
<td>Binary</td>
</tr>
<tr>
<td>Peramunetilleke and Wong (2002)</td>
<td>Bag-of-words</td>
<td>Set of keyword records</td>
<td>Boolean, TF-IDF, TF-CDF</td>
</tr>
<tr>
<td>Pui Cheong Fung et al. (2003)</td>
<td>Bag-of-words</td>
<td>Stemming, conversion to lower-case, removal of punctuation, numbers, web page addresses</td>
<td>TF-IDF</td>
</tr>
<tr>
<td>Werner and Myrray (2004)</td>
<td>Bag-of-words</td>
<td>Minimum information criterion (top 1000 words)</td>
<td>Binary</td>
</tr>
<tr>
<td>Das and Chen (2007)</td>
<td>Bag-of-words, Triplets</td>
<td>Pre-defined dictionaries</td>
<td>Different discrete values for</td>
</tr>
<tr>
<td>Soni et al. (2007)</td>
<td>Visualization</td>
<td>Thesaurus made using term extraction tool of N.J. van Eck</td>
<td>Visual coordinates</td>
</tr>
<tr>
<td>Zhai et al. (2007)</td>
<td>Bag-of-words</td>
<td>WordNet Thesaurus (stop-word removal, POS tagging, higher level concepts via WordNet)</td>
<td>Binary, TF-IDF</td>
</tr>
<tr>
<td>Rachlin et al. (2007)</td>
<td>Bag-of-words, commonly used financial values</td>
<td>Most influential keywords list (Automatic extraction)</td>
<td>TF, Boolean, Extractor</td>
</tr>
<tr>
<td>Tetlock et al. (2008)</td>
<td>Bag-of-words for negative words</td>
<td>Pre-defined dictionary. Harvard-IV-4 psychosocial dictionary.</td>
<td>Frequency divided by total</td>
</tr>
<tr>
<td>Mahajan et al. (2008)</td>
<td>Late Dirichlet Allocation (LDA)</td>
<td>Extraction of twenty-five topics</td>
<td>number of words</td>
</tr>
<tr>
<td>Butler and Kešelj (2009)</td>
<td>Character n-Grams, three readability scores, last year's performance</td>
<td>Minimum occurrence per document</td>
<td>Frequency of the n-gram in</td>
</tr>
<tr>
<td>Schumaker and Chen (2009)</td>
<td>Bag of words, noun phrases, named entities</td>
<td>Minimum occurrence per document</td>
<td>one profile</td>
</tr>
<tr>
<td>Li (2010)</td>
<td>Bag-of-words, tone and content</td>
<td>Pre-defined dictionaries</td>
<td>Binary</td>
</tr>
<tr>
<td>Groth and Muntermann (2011)</td>
<td>Bag-of-words</td>
<td>Feature scoring methods using both Information Gain and Chi-Squared metrics</td>
<td>Weighted based on the rise/</td>
</tr>
<tr>
<td>Schumaker et al. (2012)</td>
<td>OpinionFinder overall tone and polarity</td>
<td>Minimum occurrence per document</td>
<td>fall ratio of index</td>
</tr>
<tr>
<td>Lugmayr and Gossen (2012)</td>
<td>Bag-of-words</td>
<td>Stemming</td>
<td>Sentiment Value</td>
</tr>
<tr>
<td>Yu, Duan et al. (2013)</td>
<td>Bag-of-words</td>
<td>Not mentioned</td>
<td>Binary</td>
</tr>
<tr>
<td>Hagenau et al. (2013)</td>
<td>Bag-of-words, noun phrases, word-combinations, n-grams</td>
<td>Frequency for news, Chi²-approach and bi-normal separation (BNS) for exogenous-feedback-based feature selection, dictionary</td>
<td>TF-IDF</td>
</tr>
<tr>
<td>Jin et al. (2013)</td>
<td>Latent Dirichlet Allocation (LDA)</td>
<td>Topic extraction, top topic identification by manually aligning news articles with currency fluctuations</td>
<td>Each article’s topic distribution</td>
</tr>
<tr>
<td>Chatrath et al. (2014)</td>
<td>Structured Data</td>
<td>Structured data</td>
<td>Structured Data</td>
</tr>
<tr>
<td>Bollen and Huina (2011)</td>
<td>By OpinionFinder</td>
<td>By OpinionFinder</td>
<td>By OpinionFinder</td>
</tr>
<tr>
<td>Vu et al. (2012)</td>
<td>Daily aggregate number of positives or negatives on Twitter Sentiment Tool (TST) and an emoticon lexicon.</td>
<td>Pre-defined company related keywords, Named Entity Recognition based on linear conditional random fields (CRF)</td>
<td>Real number for Daily</td>
</tr>
</tbody>
</table>

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 Tanrıci 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).

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, 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 (Joachims, 1999). SVM Light is an implementation of an SVM learner which addresses the problem of large tasks (Joachims, 1999). The optimization algorithms used in SVM Light are described in 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; Tetlock et al., 2008). 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; Li, 2010; Werner & Myrray, 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 Neighbors (k-NN), Artificial Neural Networks (ANN), or Support Vector Machine (SVM) in that it builds on probabilities of a feature belonging to a certain category whereas the other mentioned approaches interpret the document feature-matrix spatially.

Yu, Duan, and Cao (2013c) 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. 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 Naïve Bayes module in the Perl programming language to conduct the computation.
<table>
<thead>
<tr>
<th>Reference</th>
<th>Algorithm type</th>
<th>Algorithm details</th>
<th>Training vs. testing volume and sampling</th>
<th>Sliding window</th>
<th>Semantics</th>
<th>Syntax</th>
<th>News and tech. data</th>
<th>Software</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pui Cheong Fung et al. (2003)</td>
<td>SVM</td>
<td>SVM-Light</td>
<td>First 6 consecutive months vs. the last month</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Not mentioned</td>
</tr>
<tr>
<td>Mittermayer (2004)</td>
<td>SVM-light</td>
<td>200 vs. 6002 examples</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>NewsCATS</td>
<td></td>
</tr>
<tr>
<td>Soni et al. (2007)</td>
<td>SVM with standard linear kernel</td>
<td>SVM with Gaussian RBF kernel and polynomial kernel</td>
<td>First 12 months vs. the remaining two months</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Not mentioned</td>
</tr>
<tr>
<td>Zhai et al. (2007)</td>
<td>SVM</td>
<td>Not mentioned</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Arizona Text Extractor (AzTeK) &amp; AZFin Text.</td>
<td></td>
</tr>
<tr>
<td>Schumaker and Chen (2009)</td>
<td>SVM-light</td>
<td>200 vs. 6002 examples</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>LibSVM package</td>
<td></td>
</tr>
<tr>
<td>Lugmayr and Gossen (2012)</td>
<td>Regression Algorithms</td>
<td>SVM with a linear kernel, SVR</td>
<td>Previous day vs. a given day (2 weeks for regression)</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>Forex-foreteller, Loughran–McDonald financial dic., AFINN dic.</td>
</tr>
<tr>
<td>Schumaker et al. (2012)</td>
<td>Linear regression model</td>
<td>Not mentioned</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>OpinionFinder</td>
<td></td>
</tr>
<tr>
<td>Hagenau et al. (2013)</td>
<td>Stepwise multivariate regression model</td>
<td>Not applicable</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Not mentioned</td>
<td></td>
</tr>
<tr>
<td>Jin et al. (2013)</td>
<td>Naïve Bayes</td>
<td>Naïve Bayes and dictionary-based</td>
<td>30,000 randomly vs. itself and the rest.</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Not mentioned</td>
</tr>
<tr>
<td>Chatrath et al. (2014)</td>
<td>Decision Rules or Trees</td>
<td>Rule classifier</td>
<td>22 September 12:00–27 September 9:00 vs. 9:00–10:00 on 27 September</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>Not mentioned</td>
</tr>
<tr>
<td>Tetlock et al. (2008)</td>
<td>Naïve Bayes and dictionary-based</td>
<td>1000 vs. the rest.</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Not mentioned</td>
<td></td>
</tr>
<tr>
<td>Yu, Duan, et al. (2013)</td>
<td>C4.5 decision tree</td>
<td>Trained by previous day features</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Not mentioned</td>
<td></td>
</tr>
<tr>
<td>Li (2010)</td>
<td>C4.5 Decision tree</td>
<td>2005 June–2005 October vs. 2005 November</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Not mentioned</td>
<td></td>
</tr>
<tr>
<td>Huang, Liao, et al. (2010) and Huang, Chuang, et al. (2010)</td>
<td>C4.5 decision tree</td>
<td>Trained by previous day features</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Not mentioned</td>
<td></td>
</tr>
<tr>
<td>Rachlin et al. (2007)</td>
<td>C4.5 decision tree</td>
<td>28 February–28 November vs. 1–19 December 2008</td>
<td>No</td>
<td>N/A</td>
<td>N/A</td>
<td>No</td>
<td>Not mentioned</td>
<td></td>
</tr>
<tr>
<td>Bollen and Huina (2011)</td>
<td>Stacked classifier</td>
<td>Perl n-gram module Text::Ngrams developed by Keselj, LIBSVM</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Not mentioned</td>
<td></td>
</tr>
</tbody>
</table>
Table 4 (continued)

<table>
<thead>
<tr>
<th>Reference</th>
<th>Algorithm type</th>
<th>Algorithm details</th>
<th>Training vs. testing volume and sampling</th>
<th>Sliding window</th>
<th>Semantics</th>
<th>Syntax</th>
<th>News and tech. data</th>
<th>Software</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wuthrich et al. (1998)</td>
<td>Multi-algorithm experiments</td>
<td>k-NN, ANNs, naive Bayes, rule-based</td>
<td>Last 100 training days to forecast 1 day</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>Not mentioned</td>
</tr>
<tr>
<td>Werner and Myrray (2004)</td>
<td>Naive Bayes, SVM</td>
<td></td>
<td>1000 messages vs. the rest</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Rainbow</td>
</tr>
<tr>
<td>Groth and Muntermann (2011)</td>
<td>Naive Bayes, k-NN, ANN, SVM</td>
<td></td>
<td>Stratified cross validations</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Not mentioned</td>
</tr>
</tbody>
</table>

(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 (Huang, Chuang, et al., 2010; Huang, Liao, 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. Huang, Liao, et al. (2010) and Huang, Chuang, 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 does not assume attribute independence. The algorithm is C4.5 developed by 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 (2004) also carry out all tests using two algorithms, Naive Bayes and SVM. Furthermore, Groth and Muntermann (2011) employ Naive Bayes, k-Nearest Neighbor (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 Neighbors (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. 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 Gencay (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 modeling 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 vs. 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; 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; 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 (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
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 behavior 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).

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 (Huang, Chuang, et al., 2010; Huang, Liao, 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 part of their pre-processing or machine learning. They are mostly dictionaries (Li, 2010; Tetlock et al., 2008), classification algorithm implementations (Butler & Kešelj, 2009; Werner & Myrray, 2004), concept extraction and word combination packages (Das & Chen, 2007; Rachlin et al., 2007) or sentiment value providers (Schumaker et al., 2012).

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 50–70 percent, while arguing for better than chance results which is estimated at 50 percent (Butler & Kešelj, 2009; Li, 2010; Mahajan et al., 2008; Schumaker & Chen, 2009; Schumaker et al., 2012; Zhai 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, & Tanriverdi, 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; Huang, Chuang, et al., 2010; Huang, Liao, et al., 2010b; Mittermayer, 2004; Pui Cheong Fung et al., 2003; Rachlin et al., 2007; Schumaker & Chen, 2009; Schumaker et al., 2012; 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.

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

<table>
<thead>
<tr>
<th>Reference</th>
<th>Findings</th>
<th>Trading strategy</th>
<th>Balanced data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wuthrich et al. (1998)</td>
<td>Ftse 42%, Nky 47%, Dow 40% Hsi 53% and Sti 40%.</td>
<td>Yes</td>
<td>Not mentioned</td>
</tr>
<tr>
<td>Peramnetlleike and Wong (2002)</td>
<td>Better than chance</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Pui Cheong Fung et al. (2003)</td>
<td>The cumulative profit of monitoring multiple time series is nearly double to that of monitoring single time series.</td>
<td>Yes</td>
<td>Not mentioned</td>
</tr>
<tr>
<td>Werner and Myrayr (2004)</td>
<td>Evidence that the stock messages help predict market volatility, but not stock returns</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Mittermayer (2004)</td>
<td>Average profit 11% compared to average profit by random trader 0%</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Das and Chen (2007)</td>
<td>Regression has low explanatory power</td>
<td>No</td>
<td>Not mentioned</td>
</tr>
<tr>
<td>Soni et al. (2007)</td>
<td>Hit rate of classifier: 56.2% compared to 47.5% for naive classifier and 49.1% SVM bag-of-words</td>
<td>No</td>
<td>Yes (roughly)</td>
</tr>
<tr>
<td>Zhai et al. (2007)</td>
<td>Price 58.8%, direct news 62.5%, indirect news 50.0%, combined news 64.7%, price and news 70.1%</td>
<td>Yes</td>
<td>Not mentioned</td>
</tr>
<tr>
<td>Rachlin et al. (2007)</td>
<td>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</td>
<td>Yes</td>
<td>Not mentioned</td>
</tr>
<tr>
<td>Tetlock et al. (2008)</td>
<td>(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</td>
<td>Yes</td>
<td>Not relevant</td>
</tr>
<tr>
<td>Mahajan et al. (2008)</td>
<td>Accuracy 60%</td>
<td>No</td>
<td>Not mentioned</td>
</tr>
<tr>
<td>Butler and Kešelj (2009)</td>
<td>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</td>
<td>No</td>
<td>Not mentioned</td>
</tr>
<tr>
<td>Schumaker and Chen (2009)</td>
<td>Directional accuracy 57.1%, return 2.06%, closeness 0.04261</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Li (2010)</td>
<td>Accuracy for tone 67% and content 63% with naive Bayes and less than 50% with dictionary-based</td>
<td>No</td>
<td>Not mentioned</td>
</tr>
<tr>
<td>Huang, Liao, et al. (2010) and Huang, Chuang, et al. (2010)</td>
<td>Prediction accuracy and the recall rate up to 85.2689% and 75.3782% in average, respectively</td>
<td>Yes</td>
<td>Not mentioned</td>
</tr>
<tr>
<td>Groth and Muntermann (2011)</td>
<td>Accuracy (slightly) above the 75% guessing equivalent benchmark Objective articles were performing poorly in directional accuracy vs. baseline. Subjective articles performed better with 59.0% directional accuracy and a 3.30% trading return. Polarity performed poorly vs. baseline</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Schumaker et al. (2012)</td>
<td>Neutral articles had poorer trading returns vs. baseline. Subjective articles performed better with 59.0% directional accuracy and a 3.30% trading return. Polarity performed poorly vs. baseline</td>
<td>No</td>
<td>Not mentioned</td>
</tr>
<tr>
<td>Lugnayr and Gossen (2012)</td>
<td>In progress</td>
<td>No</td>
<td>Not mentioned</td>
</tr>
<tr>
<td>Yu, Duan, et al. (2013a)</td>
<td>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 significant negative relationship with risk</td>
<td>No</td>
<td>Not mentioned</td>
</tr>
<tr>
<td>Hagenau et al. (2013)</td>
<td>Feedback-based feature selection combined with 2-word combinations achieved accuracies of up to 76%</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Jin et al. (2013)</td>
<td>Precision around 0.28 on average</td>
<td>No</td>
<td>Not mentioned</td>
</tr>
<tr>
<td>Chattrath et al. (2014)</td>
<td>(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-min of the news release</td>
<td>No</td>
<td>Not mentioned</td>
</tr>
<tr>
<td>Bollen and Huina (2011)</td>
<td>The mood ‘Calm’ had the highest granger causality relation with the DJIA for time lags ranging from two to six days (p-values &lt; 0.05). The other four GPMFs mood dimensions and OpinionFinder did not have a significant correlation with stock market changes</td>
<td>No</td>
<td>Not mentioned</td>
</tr>
<tr>
<td>Vu et al. (2012)</td>
<td>Combination of previous day’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%</td>
<td>No</td>
<td>Not mentioned</td>
</tr>
</tbody>
</table>

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 behavioral-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 3.3.1–3.3.3 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 favored by researchers, probably due to their straightforwardness, while many other machine learning algorithms or techniques like Artificial Neural Networks (ANN), K-Nearest Neighbors (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 behavioral-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 behavioral 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 behavioral-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 favored 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.

1. **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 in 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.

Acknowledgement

This work is supported by University of Malaya High Impact Research Grant no vote UM.C/625/HIR/MOHE/SC/13/2 from Ministry of Education Malaysia.

References


