Big data: web-crawling and analysing financial news using RapidMiner

Jesse Lane and Hak J. Kim*

Hofstra University,
Hempstead, NY 11590, USA
Email: jlane27@gmail.com
Email: hak.j.kim@hofstra.edu
*Corresponding author

Abstract: Big data today is one of the hottest topics in the ICT field, but still there are many questions around what it is, what it really means, and how it can be used. This paper presents the notion of big data and then attempts to analyse it using a typical analytics tool, which is called RapidMiner. We use actual real-world social media data as an empirical test. Our preliminary result shows that social media data does not provide valuable meaning to predict the future stock market. However, we believe that the analysis of big data is meaningful if more sophisticated methodology and data collection procedures are used.

Keywords: big data; social media; financial news; RapidMiner.

Reference to this paper should be made as follows: Lane, J. and Kim, H.J. (2015) 'Big data: web-crawling and analysing financial news using RapidMiner', *Int. J. Business Information Systems*, Vol. 19, No. 1, pp.41–57.

Biographical notes: Jesse Lane works as Senior Financial Analyst at CA Technologies. While working, he quickly realised the power of adding an IT perspective to his experience in the finance world. He obtained his MS in Information Technology from Hofstra University. He graduated with a Bachelor of Accountancy from The George Washington University. He enjoys learning new ways to analyse data and is passionate about his startup, Fantasy Press Conference, LLC.

Hak J. Kim is an Associate Professor in the Zarb School of Business at Hofstra University. He received his PhD in Information Science from the University of Pittsburgh. His research interests include mobile location-based services with mobile app, big data analytics in social media, cyber security in cloud computing, and digital forensics.

1 Introduction

Modern digital data mostly came from computers and stored in database as a structured format (i.e., field names). They are relatively small amount, single type (i.e., text), and non-real time data. However, today's data show different shapes, such as huge amount, diverse types, and real-time data. This phenomenon is called 'big data'. Big data are generated on a daily bases from social networks, sensors, model simulations, and many other sources. With their complex and unstructured characteristics, big data are becoming

challenge for developing new applications and systems to manage, discover, access, and process the big data (Manyika et al., 2011).

With the development of mobile cloud computing technologies (Hayes, 2008; Milojicic, 2008; Sotomayor et al., 2009; Fernando et al., 2013), the wide adoption of smartphones generates more data through social media applications, such as Facebook (Blitz, 2012) and Twitter (Bulearca and Bulearca, 2010). Social media today is popularised as new paradigm of communication (Violino, 2011). People communicate with their colleagues using social media tools (i.e., Twitter and Facebook) instead of traditional communication tools (i.e., phone and e-mail).

The emerging of cloud computing (Cusumano, 2011) as a new generation computing infrastructure provides potential computing solutions to the management, discovery, access, and processing of the big data. Cloud computing provides solutions for big data with computing infrastructure capability to process and obtain unprecedented information different from the traditional internet environment (Werbach, 1997). At the same time, big data pose grand challenges as opportunities to advance cloud computing.

Organisations have increased their understanding of what big data is and what value it can potentially deliver to the business. Through our customer interactions, questions have shifted from "What is big data?" and "Why should I care?" (IBM, 2013) to "What is the ROI?" and "What are the organizational changes and skills required?" Yet many organisations are still in the early stages of experimentation and few have thought through a strategy or realised the profound impact that big data will have on organisations and information infrastructure.

This paper explores financial news in the web as a type of big data and compares it with financial data (i.e., stock price index). In this study, we attempt to show how social media data can be used to predict financial stock market and then to build a simple model for predicting stock market change.

2 Big data: an overview

2.1 Notion of big data

Big data today is one of the hottest topics in IT area. Mckinsey's latest report (Manyika et al., 2011) says that big data is the next frontier for innovation, competition, and productivity. Big data is also chosen one of the big five IT trends of the next half decade by ZDNet (Hinchcliffe, 2011) and identified one of the Top 10 Strategic Technologies for 2012 by Gartner (2011). According to IBM (2013), 2.5 quintillion bytes of data are created every day and 90% of the data in the world today has been created in the last two years alone. Although everybody is talking about big data, there are still many questions around what it is, what it means, and how it can use.

Big data also affects business intelligence (BI) (Chen et al., 2012; Chaudhuri et al., 2011) and data mining (Pabreja and Datta, 2012; Davenport, 2006). In BI, the traditional BI is focused on searching and collecting meaningful data and implicitly ignores their analysis. However, the explosive growth of data makes impossible simply to find useful data (Chiang et al., 2012). With the popular use of web and social media, new techniques are emerged, such as web-crawling (Pavalam et al., 2011; Olston and Najork, 2010; Cothey, 2004) and social media analysis (Blitz, 2012; Bulearca and Bulearca, 2010). Bhatnagar (2013) introduces the layered framework for analysing big data efficiently.

Another approach for big data analysis is probabilistic topic models (Blei, 2012) which is techniques to analyse data using probability. Karpf (2009) suggests a mixed-methods approach to rapidly changing systems.

LaValle et al. (2011) shows how to convert big data insights to value through analysing big data. For example, in financial sector, company financial data like prices and costs can be analysed its business using some models. For example, what was last year's top selling product in the northeast area? Then, we can generate some graphs about it. In the big data era, knowing that is not enough to produce a meaningful strategy to remain competitive. Along with sales data, we now have an abundance of other meaningful data. Increasingly, data has been auto-generated from networks, websites, supply chains, sensors, markets, system logs etc. Reviews, shares, and sentiment from social networking sites, information from partners and suppliers, as well as their social and operational data all come into play. So if you are a strategic retailer trying to gauge the northeast region's top selling product for next year, you now need to make use of all the information relevant to your business. Failing to do so, could mean missing anticipated trends, and consequently, next year's sales target.

Social media are becoming more prevalent and emerged as a new way of life to the people (Hathi, 2009). According to IDC (2012), more than two billion internet users and 4.6 billon mobile phones are in the world. Facebook (Foster et al., 2010) has more than 500 million users and created 30 billion pieces of content every month. And about 340 millions of data every day in Twitter are exchanged. As a result, we are living the age of big data.

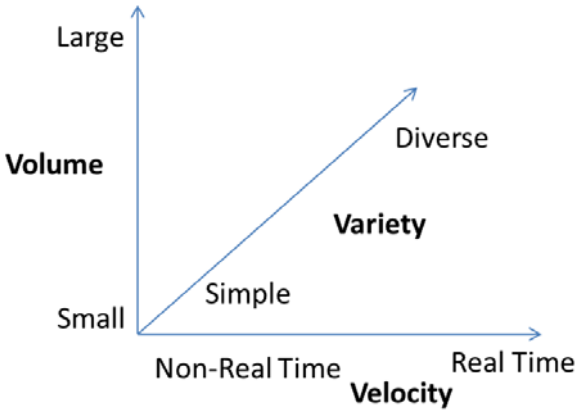
Nowadays big data terminology is popularly used in the business world. Then, what is big data? There is no single definition of big data until now, but broadly speaking it is the tidal wave of data, not only volume but also velocity and variety, from the cloud computing and social media. Narrowly, it can be defined as datasets whose size is beyond the ability of typical database software tools to capture, store, manage, analyse, and visualise.

2.2 Characteristics

As shown Figure 1, big data can be characterised by volume, velocity, and variety. They provide a helpful lens to view and understand the nature of big data; big volume, high velocity, and wide variety. First, data volume is exploding due to the increase of social media, online data and location data. The pace of business activity and competitive pressure increases as companies begin to use data occurring on a more frequent basis, including streaming data. Many companies have large amounts of archived data, perhaps in the form of logs, but not the capacity to process it. Big data does not mean absolute amount of data, but relatively large amount of data. The size of big data is varied by industry sectors from few dozen tera bytes to multiple peta bytes. Second, the internet and mobile era means that the way we deliver and consume products and services is increasingly instrumented, generating a data flow back to the provider. Online retailers are able to compile large histories of customers' every click and interaction. The smartphone era increases again the rate of data inflow, as consumers carry with them a streaming source of geo-located imagery and audio data. Third, a common theme in big data systems is that the source data is diverse, and doesn't fall into neat relational structures. It could be text from social networks, image data, a raw feed directly from a

sensor source. Even on the web, different browsers send different data, users withhold information. They use differing software versions or vendors to communicate with you.

Figure 1 Characteristics of big data (see online version for colours)



2.3 Data types: structured vs. unstructured

There are two types of data; structured and unstructured. Structured data refers to data with high degree of organisation in a structure so that it is identifiable, such as data in database. While unstructured data is the opposite. It is simply the lack of structure. The typical types of unstructured data include video clips, weblogs, social media feeds, etc. For example, e-mail is a type of unstructured data because it does not generally write about precisely one subject and even the format. Data in spreadsheets, on the other hand, is an example of structured data because it can be arranged in a database system. In reality, about 80% of the world's data in the business world is unstructured (IDC, 2012). It may be data we've been aggregating before, but could not process with current data mining tools.

2.4 Platform

Big data tends to rely on a hotchpotch of multiple software applications, hardware and services rather than single, unified platforms.

A rough and simplified view of a common big-data model depicts:

1 information acquired or ingested from many different sources, including relational (structured) and non-relational (unstructured)

2 this is then passed on to distributed file systems

3 processing engines based on Apache's Hadoop software framework, say, for integration and organisation

4 it is aggregated in data warehouses or other storage repositories

5 finally it is pushed out to analytics or BI applications.

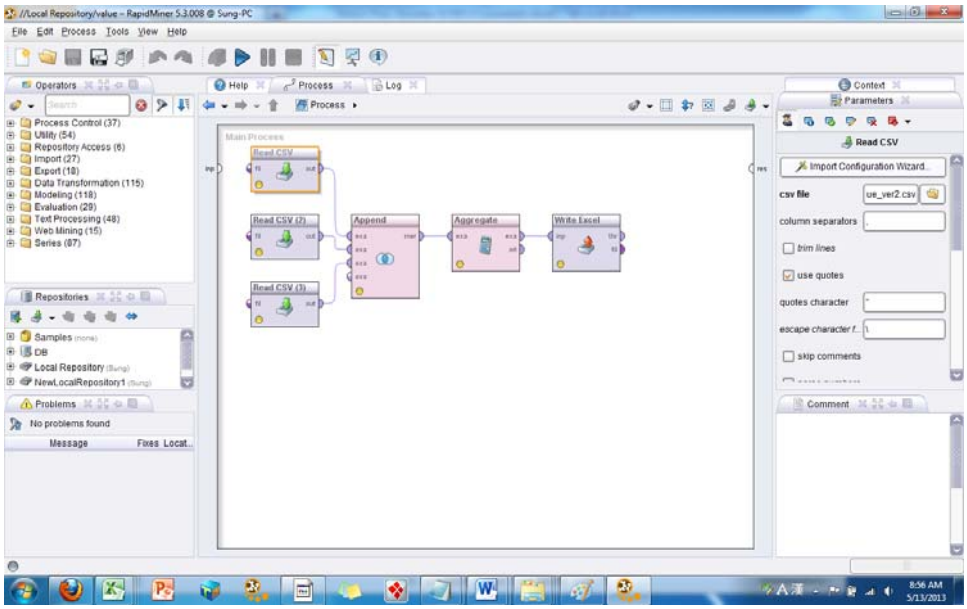
These in turn present the results of specific queries in some sort of graphical report format which are easily understood by end-users.

Hadoop comes in: filtering and using MapReduce so that it can be decided just where the resulting data should reside – in a SQL database or in a schema-less, NoSQL equivalent. Once all; this has been done, then big-data analytics can be brought to bear. To put it another way, big-data methodologies are not an end in themselves, but rather the starting point of a longer, even more resource-intensive process.

3 RapidMiner as big data analytic tool

One of the popular big data analytics software is RapidMiner (Rapid-I, 2012). It is an open-source system for data mining used by businesses, in academics, programmers and anyone else interested in data mining. It was written in the Java programming language and is available to be downloaded for free from its website (http://www.rapid-i.com). In addition to providing a GUI interface for its users, developers also have the opportunity to integrate RapidMiner into their own product. It could take in data sources like Excel, Access, Oracle… etc. to perform data analytics and produce results such as evaluations and visualisations. In this analysis, Microsoft Excel 2010 was used mainly for data preparation and transformation so that RapidMiner could analyse the data using several statistical techniques. Figure 2 shows a main screen of Rapid Miner.

Figure 2 Main screen of RapidMiner (see online version for colours)



RapidMiner was created under its former name of yet another learning environment (YALE) in 2001 by Ralf Klinkenberg, Info Mierswa, and Simon Fischer at the University of Dortmund, Germany. We envisioned a data mining tool that was more flexible and by far more powerful than the tools available in the market (Mierswa and Klinkenberg, 2010). In other words, RapidMiner was created to be this more powerful tool to perform

data mining. The company Rapid-I GmbH acquired the software in 2007 and has been developing it since. In today's world data drives decisions, and in order to choose the best option, many hours are spent analysing this data. RapidMiner can take this data and assist in finding the optimal solutions.

Figure 3 Operators menu (see online version for colours)

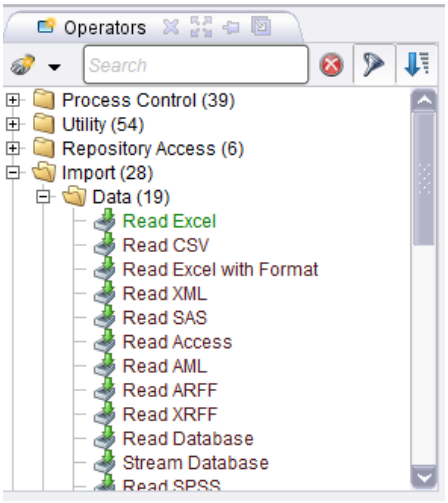
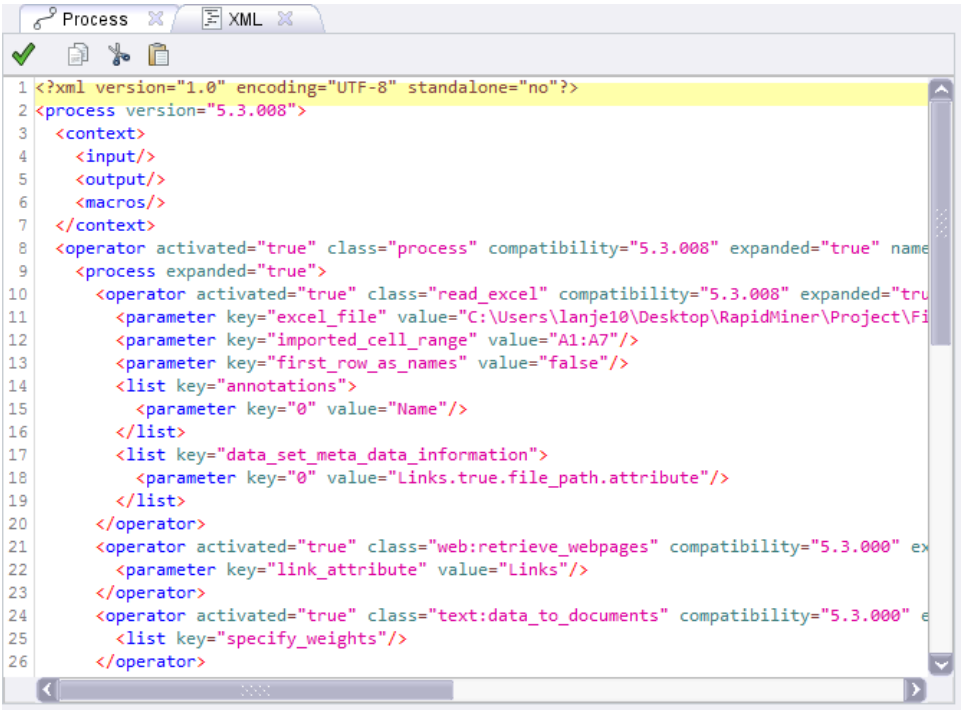


Figure 4 Description of operator by XML (see online version for colours)



There is a tremendous amount of functionality within RapidMiner. RapidMiner contains more than 500 operators altogether for all tasks of professional data analysis, i.e., operators for input and output as well as data processing (ETL), modelling and other aspects of data mining. Furthermore, there are additional extensions available to incorporate text mining, web mining, the statistical R language, and a few others to help manage the data. RapidMiner is open-sourced software that allows users to analyse large amounts of data and return a desired set of information. There are numerous operators that can be included in the process flow. Figure 3 illustrates a sample of the operators, as shown Figure 3.

Furthermore, besides the GUI interface of the process flow, it can also be illustrated in XML notation (Figure 4). There is a tab at the top of the screen that identifies XML and by clicking it, the following is an example of what is displayed.

The RapidMiner GUI interface uses drag and drop technology for each of its operators. Drag processes and operators in order to create a process flow. Next, certain operators require parameters to be set. By clicking on the operator in the process flow, a list of the parameters will be displayed on the right side of the screen. Certain operators require nested operators, processes within the main process. To get to the nested process, double click the parent operator and the sub will be displayed. There are also arrows at the top of the software that allow for easy navigation between parent/sub processes.

To connect each operator, click the output/input area of the process and drag this to the next operator in line. The interface that RapidMiner provides makes adding and connecting processes very simple. Once the entire process flow is created, simply press the play button at the top of the screen to get the results. RapidMiner will display numerous statistics about the process that was run, such as time to complete. After it is run, the desired results will be displayed, whether it is a similarity measure, word list, or any of the other many options that can be setup in the process. Further, there are different graphical views that RapidMiner offers, if the user chooses to include these in the process flow. To save the results, simply click the save disk icon and it can be placed within the repository that the entire process flow sits in. A user only needs to click on the saved results to have them displayed.

The RapidMiner interface has six major components; operators, repositories, process, problems, toolbar, and parameters. All these components can be moved around in the interface according to the user's preference. The 'operators' component, which default location is top left, is the collection of tasks to be included in the process. Users can simply drag-and-drop, or double-click to add functions to the process. The operators are grouped by their functions. For example, the process control group contains all operators that control the process flow, such as loops or conditional branches. If an extension is installed, additional group(s) of operators would be added.

The 'repositories' component, which location is on the middle left, is the collection of data, files, projects and directories that help the organisation in the structured manner. Through this component, users could simply click the available data to open, look or incorporate in the process. Furthermore, the repositories help the structured management of data, processes, results, and reports so that the navigation becomes easier and simpler for the user. The repository can be located in a local or shared file system, or even in an external RapidMiner analysis server called RapidAnalytics.

The 'process' component, located in the middle of the interface, is the graphical perspective of the interconnection between the operators. Ports are critical to this

component because it enables users to generate inputs or outputs of the operators. Users can interconnect operators by simply dragging a line from port to port.

The 'problems' component, located in the bottom left, contains any warnings or error messages pertaining to the process. It has three columns – message, fixes and location. The message column shows a short description of the problem. For example, the data mining operator 'Gaussian process' cannot handle polynomial attributes. The Fixes column provides users with potential solutions to the problem. The Location column shows where the problem exists.

The 'toolbar' component, located in the top, contains many functions that provide general support for the users. One critical component to the Toolbar is the Icons for Perspectives. The icons enable users to choose from one of the three perspectives – design, result and welcome. The design perspective is where users create and manage the analysis process. The result perspective is where users can view the results of the analysis. The welcome perspective is the Welcome screen that the users see when they start the program.

The 'parameters' component, located in the right, contains parameters that are required for operators so that they could function correctly. For example, the 'read CSV' operator used in this analysis needs an indication of a file path so that the operator could read the CSV.

4 A case study for analysing financial news in the web

Analysing financial market via website is interesting to financial researchers. In our study, we attempt to compare financial numeric data (i.e., S&P index) and text data (i.e., financial news). Our implicit propositions are that both data have strong relationship and text data will affect positively financial market.

This study is to retrieve the words listed on five financial websites using a data analytic tool and then compare the results to the daily change in the S&P 500. Today's world is full of data including websites and social media data. Although these data may provide very little value, but the key is to be able to turn this data into information that can help to predict the future and makes decisions that will have a positive impact.

Enter the world of 'big data.' The most successful technology companies today have been using big data to generate value for their organisations. The new mega-rich of Silicon Valley, first at Google and now Facebook, are masters at harnessing the data of the web-online searches, posts and messages with internet advertising (Lohr, 2011). The idea of using data analytics to make decisions will create new jobs and change the way the world works.

4.1 Research propositions

As an empirical test, we attempt to analyse three components;

1 website similarity

2 topic association

3 key words predictability, as shown Figure 5.

Figure 5 A roadmap of empirical test (see online version for colours)

First, we are interested in measuring the similarity of web contents among financial websites although each site is independently maintained. Especially major news in financial market should be consistent among them. If this proposition is proven, then we may not need to search several sources to get information for financial decisions. On the contrary, we may continue to investigate which sites are better predicting financial market.

Proposition 1 The financial news contents are inherently similar among financial websites.

Second, in general, financial news site includes many financial terms. Then, we can use these sites for finding useful information. The results will support the use of financial websites for investigating the relationship between financial numeric data and text data. We will attempt to distinguish between financial words and non-financial words.

Proposition 2 The finance related words are the most frequently seen words on each financial news website.

Third, the major goal of this paper is to show the relationship between financial numeric data and text data. If this turns out to be true, there will be the potential to predict how the market closes. For example, if more 'buys' I would expect the market to end up for the day and if more 'sells' I predict the market would be down for the day. If this turns out to be true, the further testing would be needed to find a relationship before the market closes to get the appropriate trades in.

Proposition 3 The key words can predict the overall market performance for the day.

4.2 Modelling and procedure

The model created for this study utilised the web mining extension that is available with RapidMiner. The model utilised the following procedure; *read data, get pages, data to documents,* and *process documents operators* in order to extract the 'words' from a pre-populated list of websites recorded in Excel. The simulation read the excel list of websites and pushed it through the process flow to be able to pull the words on each website, which were then analysed for any predictive measures against the days financial markets change. Below is a step-by-step description of the model.

- *Step 1: read data*

  We use MS Excel application to crawl web data. The import of data using Excel is the important step to change the attribute to a file path, since each site is a URL as an external website. So, in Excel, we type the URLs to collect data.

- *Step 2: get pages*

  We need to configure the scope; for example, what was read in Excel and retrieves the data from the URLs.

- *Step 3: data to documents*

  This step generates documents from the URLs.

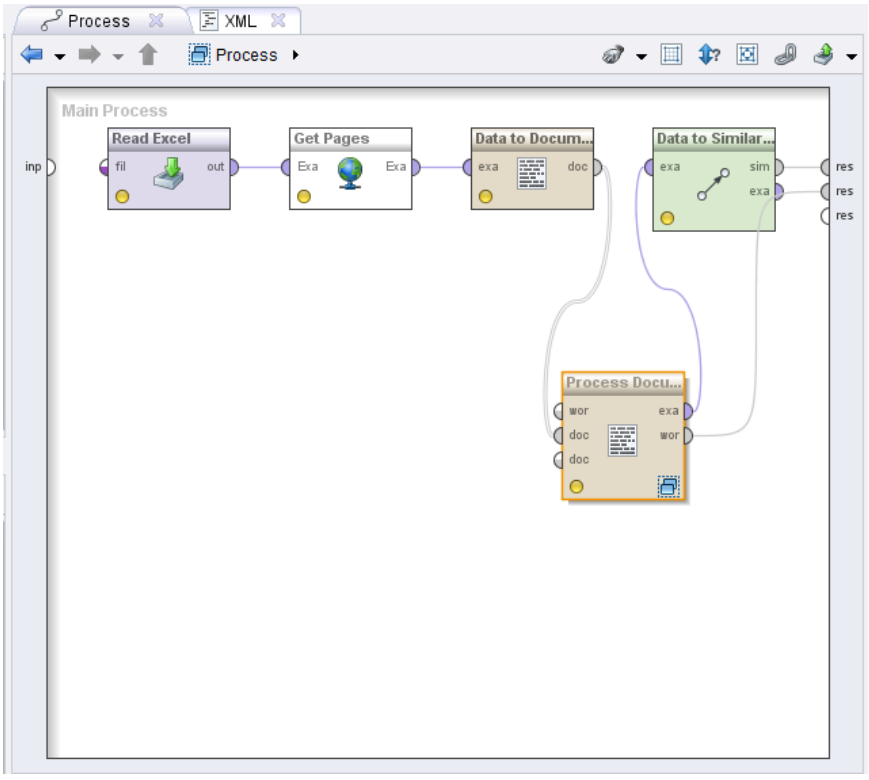
- *Step 4: process documents*

  The next step is to process document using nested operators. A operator tokenise documents which is to split a document into sequence of tokens (i.e., words). After that, it transforms all characters to lowercase for consistency throughout the model.

- *Step 5: Filter stopwords*

  The last step is to filter stopwords from a document. Stopwords means dictionary words. So, we can only retrieve meaningful words.

Figure 6 Simulation model (see online version for colours)



4.3 Data collection

The data collection is an integral part of this study. We choose to collect data every day from five financial market websites over a period of time. The data is then exported to a spreadsheet and analysed to determine if there is any correlation between key words listed on each site and the performance of the market in whole. Figure 7 shows a sample of data collection.

Figure 7 A sample of data collection (see online version for colours)

	A
1	**Links**
2	http://online.wsj.com/public/page/news-financial-markets-stock.html
3	**Links**
4	http://finance.yahoo.com/
5	**Links**
6	http://money.cnn.com/?hpt=sitenav
7	**Links**
8	http://www.businessinsider.com/moneygame
9	**Links**
10	http://www.cnbc.com/

As shown in Figure 8, the Excel file is put through the model and the words from each website listed are extracted.

Figure 8 Word list (see online version for colours)

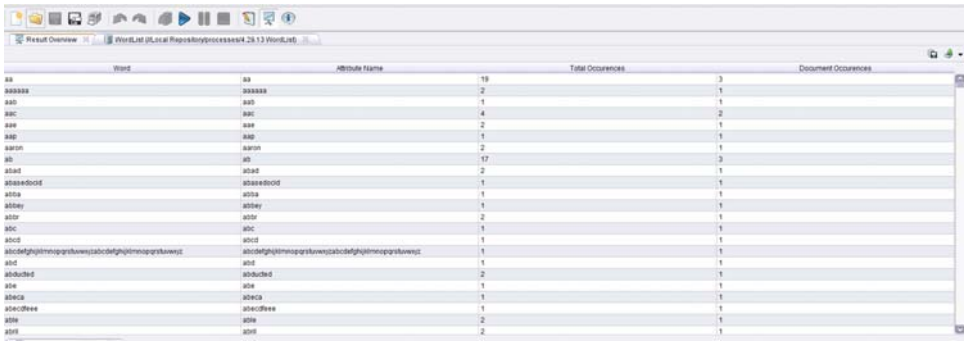


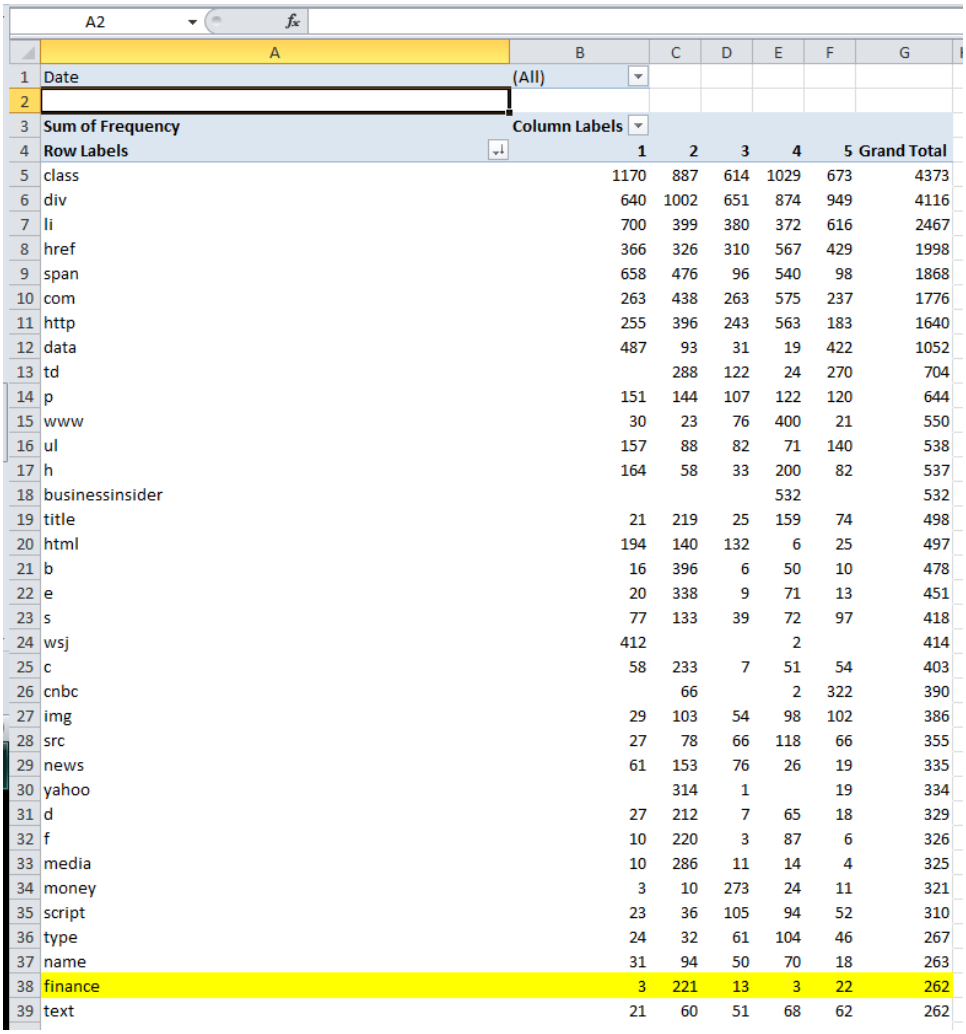
Figure 9 Word frequency (see online version for colours)

	A	B	C	D	E
1	**Word**	**Attribute name**	**Frequency**	**Website**	**Date**
2	aap	aap	1	1	4/29/2013
3	ab	ab	2	1	4/29/2013
4	abasedocid	abasedocid	1	1	4/29/2013
5	abcdefghijklmnopqrstuvwxyzabcdefghijklmnopqrstuvwxyz	abcdefghijklmnopqrstuvwxyzabcdefghijklmnopqrstuvwxyz	1	1	4/29/2013
6	abd	abd	1	1	4/29/2013
7	abril	abril	2	1	4/29/2013
8	access	access	2	1	4/29/2013
9	account	account	4	1	4/29/2013
10	acctid	acctid	1	1	4/29/2013
11	acfooter	acfooter	1	1	4/29/2013
12	acheadline	acheadline	1	1	4/29/2013
13	action	action	5	1	4/29/2013
14	activist	activist	2	1	4/29/2013
15	actual	actual	1	1	4/29/2013
16	ad	ad	36	1	4/29/2013
17	adbf	adbf	2	1	4/29/2013
18	adclass	adclass	7	1	4/29/2013
19	addclass	addclass	1	1	4/29/2013
20	added	added	2	1	4/29/2013
21	admanager	admanager	14	1	4/29/2013
22	adow	adow	4	1	4/29/2013
23	adp	adp	1	1	4/29/2013
24	ads	ads	2	1	4/29/2013
25	adsummary	adsummary	13	1	4/29/2013

Once extracted, the data is copied into Microsoft Excel and the date is added as a field (Figure 9).

Using the pivot Table functionality of Microsoft Excel, the data is summarised and allowed for each word to be analysed in each website (Figure 10).

Figure 10 Word frequency (see online version for colours)

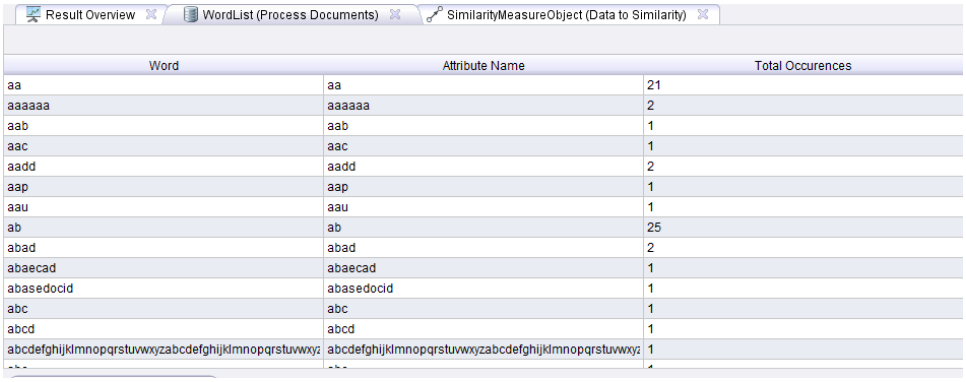


	A	B	C	D	E	F	G
1	Date	(All)					
2							
3	Sum of Frequency	Column Labels					
4	Row Labels	1	2	3	4	5	Grand Total
5	class	1170	887	614	1029	673	4373
6	div	640	1002	651	874	949	4116
7	li	700	399	380	372	616	2467
8	href	366	326	310	567	429	1998
9	span	658	476	96	540	98	1868
10	com	263	438	263	575	237	1776
11	http	255	396	243	563	183	1640
12	data	487	93	31	19	422	1052
13	td		288	122	24	270	704
14	p	151	144	107	122	120	644
15	www	30	23	76	400	21	550
16	ul	157	88	82	71	140	538
17	h	164	58	33	200	82	537
18	businessinsider				532		532
19	title	21	219	25	159	74	498
20	html	194	140	132	6	25	497
21	b	16	396	6	50	10	478
22	e	20	338	9	71	13	451
23	s	77	133	39	72	97	418
24	wsj	412			2		414
25	c	58	233	7	51	54	403
26	cnbc		66		2	322	390
27	img	29	103	54	98	102	386
28	src	27	78	66	118	66	355
29	news	61	153	76	26	19	335
30	yahoo		314	1		19	334
31	d	27	212	7	65	18	329
32	f	10	220	3	87	6	326
33	media	10	286	11	14	4	325
34	money	3	10	273	24	11	321
35	script	23	36	105	94	52	310
36	type	24	32	61	104	46	267
37	name	31	94	50	70	18	263
38	finance	3	221	13	3	22	262
39	text	21	60	51	68	62	262

4.4 Results analysis

The simulation provided two key results, the word list and similarity measure. Figure 11 shows the list of word.

Figure 11 The list of word (see online version for colours)


Result Overview | WordList (Process Documents) | SimilarityMeasureObject (Data to Similarity)

Word	Attribute Name	Total Occurences
aa	aa	21
aaaaaa	aaaaaa	2
aab	aab	1
aac	aac	1
aadd	aadd	2
aap	aap	1
aau	aau	1
ab	ab	25
abad	abad	2
abaecad	abaecad	1
abasedocid	abasedocid	1
abc	abc	1
abcd	abcd	1
abcdefghijklmnopqrstuvwxyzabcdefghijklmnopqrstuvwxyz	abcdefghijklmnopqrstuvwxyzabcdefghijklmnopqrstuvwxyz	1

4.4.1 Website similarity (Proposition 1)

Figure 12 shows to measure similarity of sample set (five websites) with each of the others. Since our study has all financial market websites, the assumption was each should be quite similar to each other.

As show Figure 12, the websites were not as similar as we predicted. So, there is no proof of similarity between websites. The two most similar according to the model are http://www.finance.yahoo.com (2) and http://www.cnbc.com (5), however the similarity is low. Looking at the data below, (2) and (5) were consistently the most similar, but nothing above a .07 similarity. One possibility is the main page of each website offers different articles, as a result of opinion or even political reasons, as certain websites may lean towards a certain party. Overall, the model developed illustrates there is a poor similarity between the chosen websites.

Figure 12 Similarity result (see online version for colours)

First	Second	Similarity
1.0	2.0	0.012
1.0	3.0	0.011
1.0	4.0	0.004
1.0	5.0	0.007
1.0	6.0	0.009
2.0	3.0	0.019
2.0	4.0	0.024
2.0	5.0	0.088
2.0	6.0	0.011
3.0	4.0	0.005
3.0	5.0	0.023
3.0	6.0	0.004
4.0	5.0	0.006
4.0	6.0	0.004

4.4.2 Topic association (Proposition 2)

When examining the results of all the sites combined, the most frequent words do not relate to finance. The results are less than ideal as the most frequent words related to the html coding of each website. In future experiments, the html should be eliminated in order to develop a true view of the sites.

- *Buy:* overall, the market was mostly positive over the five day period. Analysing the keyword 'buy' illustrates that there is no correlation between the market performance and the frequency. One possibility is the main page of each site does not get into as much detail as necessary to capture the word 'buy'.

- Sell: there is no correlation between the word 'sell' and the performance of the market over the period examined. Here, the same issue may have occurred as with 'buy', not enough detail on the main homepage of each website to garner enough data to make any predictions.

When looking at the best performing and worst performing days of the S&P 500 (Figure 13) and comparing that to the frequency of the words 'buy' and 'sell' over the same two days, there is no correlation between 'buy' and highest performing and 'low' and worst performing. Again, there are numerous other factors that could have impacted this. For example, there may have been more important stories published by each website over the period of time analysed that did not include advice on whether to 'buy' or 'sell' (Figure 14). Furthermore, the similarity in frequency could be attributed to different investment philosophies of each writer on the websites. One may see this as a time to 'buy' with the market going up and others may say 'sell' since the market is at a high point.

Figure 13 S&P index change (see online version for colours)

S&P 500

Date	Open	High	Low	Close	Change
3-May-13	1,597.60	1,618.46	1,597.60	1,614.42	1.05%
2-May-13	1,582.77	1,598.60	1,582.77	1,597.59	0.94%
1-May-13	1,597.55	1,597.55	1,581.28	1,582.70	-0.93%
30-Apr-13	1,593.58	1,597.57	1,586.50	1,597.57	0.25%
29-Apr-13	1,582.34	1,596.65	1,582.34	1,593.61	0.71%

Figure 14 Buy and sell (see online version for colours)

Sum of Frequency	Column Labels					
Row Labels	1	2	3	4	5	Grand Total
buy	11	29	6	5	13	64
buybacks	2					2
buyer	2	1				3
buyers					1	1
buying	3	9	2	12	12	38
buyout	1					1
buys		15		8	8	31
homebuyer			1			1
homebuyers			6			6
Grand Total	**19**	**54**	**15**	**25**	**34**	**147**

Sum of Frequency	Column Labels					
Row Labels	1	2	3	4	5	Grand Total
sell	6	29	6	12	14	67
selling	4		11	21	4	40
sells	2		3	5		10
seller		3			1	4
selloff			1			1
Grand Total	**12**	**32**	**21**	**38**	**19**	**122**

5 Conclusions

5.1 Summary

Until now, we present new IT environment including mobile cloud computing, social media, and big data. The internet is the backbone of our society, while mobile cloud computing is a central source of social change. Social media has created big data which is beyond the ability of typical database software tools to capture, store, manage, analyse, and visualise.

The model constructed provided a poor measure for predicting the close of the market. The similarity was low, the most frequent words ended up being html code and the key words 'buy' and 'sell' did not illustrate any predictive measures for the close of the market. There are numerous factors that could have impacted the results such as each websites home page not focusing solely on markets and investing, the impact of the html code from each website causing the similarity to be low and the bias of the contributors of each site. Again, there are no predictive measures derived from the experiment that could be used to accurately predict the ups and downs of the market.

5.2 Managerial implications and limitations

Today businesses firms are challenged by big data because it grows so large that they become awkward to work with using on-hand database management tools. They are also becoming challenge for application developments. Big data are produced on a daily bases from diverse social media and even they are not well structured. However, big data has big potential that it can generate significant value across sectors, such as healthcare, retail, manufacturing, public sector, etc. In our paper, we attempt to find meaning from text data as a type of big data.

Although our preliminary result shows that there is no strong relationship between financial numeric data and text data. We don't have enough data to investigate this relationship. One of our major limitations is the length of period to collect data. We need to extend it to month or year to get more meaningful result.

Another limitation is that our data (i.e., financial news data) is static, not dynamic. We need to use social media data (i.e., twitter) for better result because financial market is real-time change. However, there is difficult to access twitter data because it may occur privacy issue.

Finally, learning RapidMiner is valuable experience as an introduction to data analytics. It is a powerful tool and I was only able to examine a small portion of the software. While the results are not as expected, they prove that the individual words on a website do not provide any value in predicting the stock market performance.

5.3 Future study

In the future, to gain better and potentially useful results, a different avenue, besides websites, should be utilised. For example, analysing the tweets about finance throughout the day and comparing that to the market performance may result in a better correlation since each tweet will be much more specific compared to an entire website because of Twitter's character limit.

Overall, this experiment was only a start to the world of data mining and analytics. Ideally, a full semester would be spent building models and developing a better process rather than learning the software and then having limited time to analyse a model.

References

Bhatnagar, V. (2013) 'Data mining-based big data analytics: parameters and layered framework', *International Journal of Computational Systems Engineering*, Vol. 1, No. 4, pp.265–276.

Blei, D.M. (2012) 'Probabilistic topic models', *Communications of the ACM*, Vol. 55, No. 4, pp.77–84.

Blitz (2012) *Get Your Facebook Dashboard*, Blitz Local [online] http://www.blitzlocal.com/ (accessed 21 October 2012).

Bulearca, M. and Bulearca, S. (2010) 'Twitter: a viable marketing tool for SMEs?', *Global Business & Management Research*, Vol. 2, No. 4, pp.296–309.

Chaudhuri, S., Dayal, U. and Narasayya, V. (2011) 'An overview of business intelligence technology', *Communications of the ACM*, Vol. 54, No. 8, pp.88–98.

Chen, H., Roger, H., Chiang, L. and Storey, V. (2012) 'Business intelligence and analytics: from big data to big impact', *MIS Quarterly*, Vol. 36, No. 4, pp.1165–1188.

Chiang, R., Goes, P. and Stohr, E. (2012) 'Business intelligence and analytics education and program development: a unique opportunity for the information systems discipline', *ACM Transactions on Management Information Systems*, Vol. 3, No. 3, pp.12–25.

Cothey, V. (2004) 'Web-crawling reliability', *Journal of the American Society for Information Science and Technology*, Vol. 55, No. 14, pp.1228–1238.

Cusumano, M.A. (2011) 'Technology strategy and management: platform wars come to social media', *Communications of the ACM*, Vol. 54, No. 4, pp.31–33.

Davenport, T.H. (2006) 'Competing on analytics', *Harvard Business Review*, Vol. 84, No. 1, pp.98–107.

Fernando, N., Loke, S.W. and Rahayu, W. (2013) 'Mobile cloud computing: a survey', *Future Generation Computer Systems*, Vol. 29, No. 1, pp.84–106.

Foster, M.K., Francescucci, A. and West, B.C. (2010) 'Why users participate in online social networks', *International Journal of e-Business Management*, Vol. 4, No. 1, pp.3–19.

Gartner (2011) *The Top 10 Strategic Technologies for 2012*, Gartner [online] http://www.gartner.com/newsroom/id/1826214 (accessed 10 April 2013).

Hathi, S. (2009) 'How social networking increases collaboration at IBM', *Strategic Communication Management*, Vol. 14, No. 1, pp.32–35.

Hayes, B. (2008) 'Cloud computing', *Communications of the ACM*, Vol. 51, No. 7, pp.9–11.

Hinchcliffe, D. (2011) *The 'Big Five' IT Trends of the Next Half Decade: Mobile, Social, Cloud, Consumerization, and Big Data*, ZDNet News.

IBM (2013) *What Can You Do with Big Data?*, IBM [online] http://www.ibm.com/big-data/us/en/ (accessed 23 March 2013).

IDC (2012) IDC Press [online] http://www.idc.com/getdoc.jsp?containerId=prUS23355112#.UM-Swm9TzHQ (accessed 12 February 2013).

Karpf, D. (2009) 'Blogsphere research: a mixed-methods approach to rapidly changing systems', *IEEE Intelligent Systems*, Vol. 24, No. 5, pp.67–70.

LaValle, S., Lesser, E., Shockley, R., Hopkins, M. and Kruschwitz, N. (2011) 'Big data, analytics and the path from insights to value', *MIT Sloan Management Review*, Vol. 52, No. 2, pp.21–31.

Lohr, S. (2011) 'The age of big data', *The New York Times*, 11 February [online] http://wolfweb.unr.edu/homepage/ania/NYTFeb12.pdf (accessed 5 March 2013).

Manyika, J., Chui, M., Brown, B., Bughin, J., Doobs, R., Roxburgh, C. and Byers, A.H. (2011) *Big Data: The Next Frontier for Innovation, Competition, and Productivity*, McKinsey Global Institute report.

Mierswa, I. and Klinkenberg, R. (2010) *Interview by G Piatetsky-Shapiro []. Interview with RapidMiner's Ingo Mierswa, Ralf Klinkenberg*, February [online] http://www.kdnuggets.com/2010/02/f-interview-rapid-i-founders.html (accessed 13 February 2013).

Milojicic, D. (2008) 'Cloud computing: interview with Russ Daniels and Franco Travostino', *IEEE Internet Computing*, No. 5, pp.7–9.

Olston, C. and Najork (2010) 'Web crawling', *Journal of Foundations and Trends in Information Retrieval*, Vol. 4, No. 3, pp.175–246.

Pabreja, K. and Datta, R. (2012) 'A data warehousing and data mining approach for analysis and forecast of cloudburst events using OLAP-based data hypercube', *International Journal of Data Analysis Techniques and Strategies*, Vol. 4, No. 1, pp.57–82.

Pavalam, S., Raja, S., Akorli, F. and Jawahar, M. (2011) 'A survey of web crawler algorithms', *International Journal of Computer Science*, Vol. 8, No. 6, pp.309–313.

Rapid-I (2012) *Rapid Miner User Manual* [online] http://docs.rapid-i.com/files/rapidminer/rapidminer-5.0-manual-english_v1.0.pdf (accessed 20 December 2012).

Sotomayor, B., Montero, R., Llorente, I. and Foster, I. (2009) 'Virtual infrastructure management in private and hybrid clouds', *IEEE Internet Computing*, Vol. 13, No. 5, pp.14–22.

Violino, B. (2011) 'Social media trends', *Communications of the ACM*, Vol. 54, No. 2, pp.17–17.

Werbach, K. (1997) *Digital Tornado: The Internet and Telecommunications Policy*, No. 29, Working Paper, FCC Office of Plans and Policy.