World Academy of Science, Engineering and Technology
International Journal of Computer and Information Engineering
Vol:10, No:4, 2016

Unsupervised Text Mining Approach to Early Warning System

Ichihan Tai, Bill Olson, Paul Blessner

***Abstract*—**Traditional early warning systems that alarm against crisis are generally based on structured or numerical data; therefore, a system that can make predictions based on unstructured textual data, an uncorrelated data source, is a great complement to the traditional early warning systems. The Chicago Board Options Exchange (CBOE) Volatility Index (VIX), commonly referred to as the fear index, measures the cost of insurance against market crash, and spikes in the event of crisis. In this study, news data is consumed for prediction of whether there will be a market-wide crisis by predicting the movement of the fear index, and the historical references to similar events are presented in an unsupervised manner. Topic modeling-based prediction and representation are made based on daily news data between 1990 and 2015 from The Wall Street Journal against VIX index data from CBOE.

***Keywords*—**Early Warning System, Knowledge Management, Topic Modeling, Market Prediction.

I. Introduction

WRITTEN form of news has existed for centuries to broadcast events that impact people's lives and influence their decisions, and news is usually stored in unstructured forms such as newspapers, where people need to read line-by-line to acquire the knowledge embedded in them. People generally agree that more knowledge of the past and some experience of similar events can help them to better deal with upcoming events. However, historical knowledge often is forgotten over time, or never even acquired to start with, for younger people who did not have a chance to read the newspapers or experience the events. Various lesson-learned systems are developed by commercial, government, and military organization to capture and provide lessons that benefit employees who encounter situations that closely resemble a previous experience in a similar situation [1].

A system that can automatically read news to capture and organize knowledge from the past and provide lessons learned will be able to help avoid repeating the mistakes of the past, and effectively navigate through the time of crisis. An early warning system (EWS) is proposed to early detect crisis and present lessons learned from the gigantic archive of historical news data, in a timely manner. As a case study, news is consumed for prediction of whether there will be a market-wide crisis, and the historical references to similar events are presented in the event of crisis, in an automated manner. Such an automated text-based EWS is particularly useful as the majority of existing EWS rely on numerical data, and textual data is an uncorrelated data source.

Ichihan Tai, Bill Olson, and Paul Blessner are with The George Washington University, Washington, DC 20052 USA (e-mail: tai@gwu.edu, bolson@gwu.edu, pbless@gwu.edu).

II. Related Work

A. Surveying Problems Solved

Earlier early warning systems have been successfully developed using machine learning techniques such as support vector machines [2] or neural networks [3] on numerical data, or using financial market volatilities and machine learning algorithms [4]. However, existing systems have leveraged only numerical data; a text mining (e.g., leveraging news) approach to construction of EWS is missing from the body of knowledge. In previous EWS design, a sudden increase of volatilities in the stock market is used as the definition of crisis [3], which aligns with CBOE Volatility Index's definition of crisis that is going to be used in this paper [5]. The crisis measure used in this paper is observed in the actual financial market and the data is provided by CBOE. On the other hand, text mining systems are used to predict various financial metrics such as stock indexes and exchange rate [6] or gold price volatility [7] and have been proven successful. However, text mining has not yet been applied to predict the VIX index, which is also an important financial index closely watched by financial market participants.

In terms of evaluation of market prediction systems, almost all of the previous work compared their results with the odds of chance of 50%, and practitioners will agree that a system that is accurate for more than half of the time will have value to them on a day-to-day basis [8].

B. Surveying Methodologies Used

A text mining system is generally divided into three components: data, pre-processing module, and machine learning module [6]. Existing literature related to each methodology is discussed individually in this section.

1) Data Methodologies

Text mining systems utilize at least two sources: news data and market data [8]. This behavior emulates human decision makers when they read news and think about what events are described in the news, and what are the implication of those events to the market which can be observed in the market data. The majority of systems use financial news, since it is considered less noisy than the general news; the most popular sources are major news websites like The Wall Street Journal, Financial Times, Reuters, Financial Times, Reuters, Dow Jones, Bloomberg, Forbes, or Yahoo! Finance [8]. Naturally, those news sources are highly tailored to financial market participants, and it is reasonable to assume that financial

market participants have also read the same news and shaped their reactions in the financial markets. Within each news source, news headlines are argued to be more straight-to-the-point and less noisy [9], and use of breaking news can also avoid noise [8]. In this paper, The Wall Street Journal will be used since its historical data is easily accessible by research communities, and it is a popular newspaper among both financial market participants and researchers who developed market predictive text-mining systems.

2) Pre-processing Methodologies

The main purpose of pre-processing of data is to transform unstructured data into structured format that can be processed using machine learning algorithms through feature selection, dimension-reduction, and feature representation [6].

As each unique appearance of words can be considered as a dimension, the high dimensionality of textual data makes it difficult to analyze. To deal with words that mean the same thing but occupy multiple dimensions or synonyms, thesaurus models are built using information extracted from public thesaurus websites [10]. Beyond comparing textual data just at the word level, a methodology named Latent Semantic Analysis is proposed to model the implicit topic or meaning of documents utilizing semantic structure [11]. Latent Semantic Analysis is shown to be useful for lesson-learned systems [12] and stock prediction [13].

A generalized probabilistic version of topic modeling methodology, Latent Dirichlet Allocation, is also developed [14]. Latent Dirichlet Allocation is shown to be useful for predicting both the stock market [15] and the forex market [16]. In this paper, Latent Dirichlet Allocation is used in Early Warning System or prediction of stock market volatility for the first time. Latent Dirichlet Allocation has the ability to visualize each topic using words that can help users of the system to intuitively understand the logic behind the recommendations from the system.

3) Machine Learning Methodologies

Machine learning algorithms that are widely used in market prediction are Support Vector Machine, Regression Algorithms, Naive Bayes, Decision Trees, and their combinations [6]. For easy visualization purpose, decision tree-based classification algorithm C4.5 [17] are used in this paper for determining whether or not there will be a crisis.

In designing an early warning system, it is important to note that errors of false negative (a crisis happened but was not alarmed), and false positive (crisis was alarmed but did not actually happen), should not be treated equally, since a false negative is a lot costlier to users than a false positive. In dealing with a similar problem such as bankruptcy prediction, cost-sensitive machine learning algorithms are used to address the asymmetric natures of those errors [18].

III. System Construction

In construction of the system, the same general framework widely used by many researchers to construct a system that predicts the directions of financial markets using textual data is used [6]. This framework can be applied to the system since the system consumes textual news data just like other market prediction systems, to make a prediction about a financial instrument that can be observed in the financial market. The main difference is that the system predicts the crisis indicator that is derived from the movement of a volatility index, whereas other market prediction systems predict movements of stock or financial indices directly. The three main modules of the system are data, pre-processing, and machine learning.

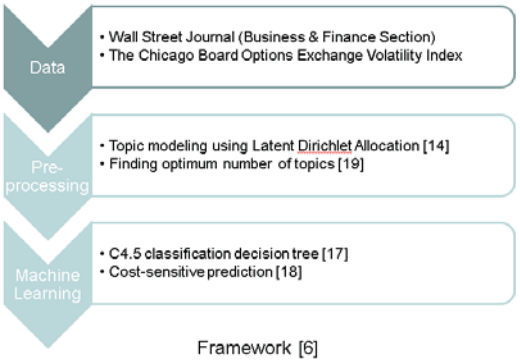

Fig. 1 High level design of the early warning system

A. Data

There are two components of data: Textual and numerical. The textual news data is consumed by the system to construct a model, and the numerical data is used to formulate the crisis indicator which indicates whether or not crisis happened on the date. The textual news data is obtained from the Business and Finance section of The Wall Street Journal from 1990 to 2015 on a daily basis, and only the abstract section is used for simplicity. The numerical data, daily VIX Index data from 1990 to 2015, is obtained from CBOE. The crisis indicator of whether crisis happened or not is defined by whether VIX has moved more than 10% based on the end of date (EOD) value compared with the previous EOD value. For example, if EOD VIX value is 19 on day 0, 20 on day 1, and 23 on day 2, the crisis indicators will be "No" on day 1 since it only has increased by 5.3%, and a "Yes" on day 2 since it has increased by 15% which is greater than 10%. Only a rapid increase is considered a crisis, but not a rapid decrease.

Each data point is composed of date, news, and crisis indicator (Yes/No). Data from 1990 to 2010 is used to train the system and data from 2011 to 2015 is used to evaluate the system. In practical sense, the system analyzes a newly published Wall Street Journal in the morning, against historical data, and warns the user whether he or she is going to have a rough day today, therefore, requiring extra attention.

B. Pre-Processing

One of the main challenges of using news data is its high dimensionality since the appearance of each word can be considered as a dimension. In this paper, high-dimensional words are reduced to and represented in the form of topics. Some classic topic modeling algorithms such as Latent Semantic Analysis are able to show whether two documents belong to the same topic or not, based on document similarity.

However, it is difficult to intuitively interpret newly-reduced dimension and feed the results into another model with confidence. On the other hand, the Latent Dirichlet Allocation (LDA) model allows easy manual interpretations of the resulting topics based on words contained in each topic, if needed. Therefore, in this paper, the dataset is put into LDA to detect topics in the dataset for dimension reduction and feature representation in the pre-preprocessing step. In order to use LDA, the model requires the number of desired topics as an input. This paper uses the natural number of topics [19] as the input for the model. LDA generates a probabilistic distribution of each data point to the topics in the form of: day X is 70% topic C, 20% topic Z, and 10% topic L.

C. Machine Learning

In the machine learning step, patterns are recognized by a machine learning algorithm and rules are generated. However, the majority of classic machine learning algorithms, such as Support Vector Machine or Neutral Network, despite their powerfulness give little information beyond the outputs and the decision rules are black-box. Without understanding how the crisis call is made, the system has little value to the actual users. Therefore, in this paper, decision tree algorithm C4.5 is used to decide whether or not the system thinks there will be a crisis today. In this step, the dataset with its probabilistic topic distribution data is fed into C4.5 using library Weka [20] to generate decision tree, such as: if a data point is greater than 50% topic M and greater than 90% topic N, then there is going to be a crisis. A cost matrix that contains a higher penalty for false negative than for false positive is given to C4.5 to make the classification algorithm cost sensitive, since it's better to have the user better prepared for a crisis that actually never happened, than vice versa.

D. Output and Evaluation

After the system is constructed, test data from 2011 to 2015 is used to test the trained model to evaluate the model. Accuracy, recall, and precision are the main criteria that are used for the evaluation, and their definitions are:

$$Accuracy = \frac{Number\ of\ correct\ classifications}{Number\ of\ classifications} \quad (1)$$

$$Recall = \frac{Number\ of\ crises\ detected}{Number\ of\ actual\ crises} \quad (2)$$

$$Precision = \frac{Number\ of\ crises\ detected}{Number\ of\ crisis\ warning\ calls\ made} \quad (3)$$

Beyond the crisis warning calls the system can deliver, another important output of the system is the visualization of market structure or system decision process using decision trees. Additionally, historical references to similar events are also displayed, to guide the users in studying the implications of the crisis.

IV. Result

Due to computation limitation, the natural number of topics are first searched in a broader step (e.g., every 500 topics initially) and more detailed searches with smaller and smaller step sizes are conducted to more promising ranges. As shown in Fig. 2, the natural number of topics for the dataset is near 100.

$$Penalty\ ratio = \frac{Penalty\ on\ false\ negative}{Penalty\ on\ false\ positive} \quad (4)$$

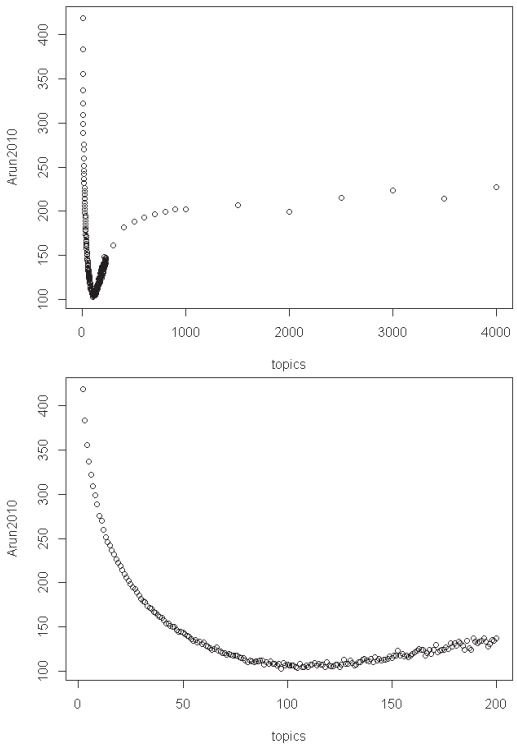

Fig. 2 Natural number of topics search results

Penalty ratio in (4) attempts to capture the asymmetric utility function of early warning system users by penalizing more if system makes less desirable prediction error. In this case, inability to warn against crisis when crisis actually happened (false negative) is considered significantly costlier to users than warning against a crisis when a crisis did not actually happen (false positive). When the penalty ratio is adjusted to 59 and the LDA model is constructed using 100 topics, the system generated 54% accuracy, 55% recall, and 9% precision; the system managed to warn against more than half of the crises (recall) while maintaining better than average accuracy. Fig. 3 illustrates the trade-off relationship between accuracy and recall when adjusting penalty ratio.

Beyond the accurate crisis warning calls the system can generate, the system can also generate historical reference to similar events. For example, on October 10, 2014, the news input indicated that "Global oil prices have fallen about 8% in the past four weeks, potentially crimping the U.S. energy boom." The system predicted that there will be a crisis, and the news belongs to Topic 90 with highest probability. Historically, it is similar to June 2, 2008 event when "Americans are resorting to often-risky ways of shoring up cash as banks tighten lending standards, jobs dwindle, home prices fall, and food and energy prices continue to climb," since the crisis happened on the day and it also belonged to

Topic 90 with highest probability. As shown in Table I, Topic 90 is regarding fear driven by energy price. The system suggests that the user recall what happened to the financial market when energy price changed drastically on October 10, 2014, and pay special attention to the time around June 2, 2008. In reality, the VIX index increased to 21.24 from 18.76, the previous day close, showing a 13% increase on October 10, 2015.

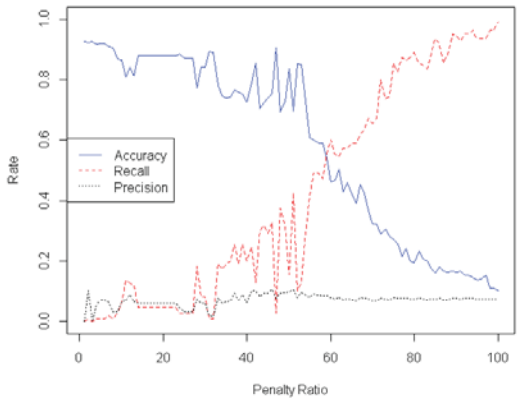

Fig. 3 Change in rates as penalty ratio increases

V. Conclusion

In this paper, it is shown that an early warning system can be successfully constructed to consume news data to alert against crisis. The system helps users understand whether crisis is going to happen for the day based on textual news data, an uncorrelated source from the traditional numerical data. Additionally, the system also advises the user on what are the potential causes of the crisis, using words that ordinary people can easily comprehend, and the system points to the historical time when similar events occurred to help users to further study the implication of the upcoming crisis.

For suggested future work, a sentiment analysis can be incorporated into this early warning system to see if sentiment can improve the performance of the system.

TABLE I
Content of Topic 90

price
fear
energi
cent
Send

Appendix

TABLE II
Content of All of the Topics

Topic 1	**Topic 2**	**Topic 3**	**Topic 4**	**Topic 5**	**Topic 6**	**Topic 7**	**Topic 8**	**Topic 9**	**Topic 10**	**Topic 11**	**Topic 12**	**Topic 13**	**Topic 14**	**Topic 15**
effort	buy	also	help	said	end	bank	oper	board	union	first	major	insur	expect	close
world	amp	wont	averag	giant	stake	sign	atampt	largest	forc	charg	much	latest	call	rais
chang	intern	judg	slid	recent	hold	three	purchas	reject	job	parent	system	fail	spend	still
ever	receiv	won	base	allow	selloff	york	home	well	work	damag	target	collaps	estim	term
aim	transact	right	accept	without	worldwid	one	british	director	contract	juri	project	told	reserv	doesnt
Topic 16	**Topic 17**	**Topic 18**	**Topic 19**	**Topic 20**	**Topic 21**	**Topic 22**	**Topic 23**	**Topic 24**	**Topic 25**	**Topic 26**	**Topic 27**	**Topic 28**	**Topic 29**	**Topic 30**
near	comput	sale	deal	loan	approv	share	make	lead	acquir	financi	nation	agre	move	oil
leav	chip	retail	valu	mortgag	reflect	stock	possibl	corpor	takeov	cost	sharehold	acquisit	two	product
becom	intel	march	peopl	pacif	gms	honeywel	hope	way	launch	biggest	corp	equiti	america	meet
isnt	appl	earlier	familiar	south	tri	ges	action	half	hostil	competit	disclos	main	partner	produc
bring	challeng	good	matter	famili	win	dynegi	clear	five	swap	california	drive	gave	north	output
Topic 31	**Topic 32**	**Topic 33**	**Topic 34**	**Topic 35**	**Topic 36**	**Topic 37**	**Topic 38**	**Topic 39**	**Topic 40**	**Topic 41**	**Topic 42**	**Topic 43**	**Topic 44**	**Topic 45**
group	news	econom	set	commod	declin	offici	talk	merger	court	profit	meanwhil	will	regul	unit
made	technolog	separ	fall	gold	current	health	cabl	creat	file	earn	higher	sever	tobacco	sell
four	data	remain	even	barrel	annual	far	bell	combin	case	quarter	plung	fight	settlement	general
focus	yesterday	indic	though	per	expand	healthcar	discuss	power	protect	strong	sharpli	finish	lawsuit	food
buyer	strength	appear	china	ounc	countri	want	telecom	consolid	bankruptci	net	worri	took	sent	motor
Topic 46	**Topic 47**	**Topic 48**	**Topic 49**	**Topic 50**	**Topic 51**	**Topic 52**	**Topic 53**	**Topic 54**	**Topic 55**	**Topic 56**	**Topic 57**	**Topic 58**	**Topic 59**	**Topic 60**
year	million	firm	american	chairman	trade	back	microsoft	rose	europ	cut	issu	invest	tax	stock
month	pay	european	airlin	execut	part	rival	seek	record	cite	boost	rule	secur	hous	ralli
next	anoth	wake	air	chief	japan	order	give	surg	effect	bush	suit	financ	clinton	less
just	potenti	past	carrier	presid	japanes	decis	softwar	jump	germani	reduc	claim	public	vote	citigroup
straight	alreadi	hire	express	name	open	boe	antitrust	februari	threaten	deficit	employe	program	bill	spark
Topic 61	**Topic 62**	**Topic 63**	**Topic 64**	**Topic 65**	**Topic 66**	**Topic 67**	**Topic 68**	**Topic 69**	**Topic 70**	**Topic 71**	**Topic 72**	**Topic 73**	**Topic 74**	**Topic 75**
propos	ibm	futur	offer	point	includ	time	fed	plan	maker	busi	pact	investor	compani	growth
control	whether	jonesaig	bid	soar	among	demand	inflat	new	big	due	accord	fund	drug	economi
asset	develop	dow	viacom	friday	total	internet	economi	govern	like	least	agreement	manag	start	street
author	question	barrel	join	lost	crisi	warner	suggest	key	car	aid	reach	money	worldcom	wall
initi	posit	spot	commun	sank	surpris	fuel	greenspan	unveil	get	huge	negoti	mutual	doubl	second
Topic 76	**Topic 77**	**Topic 78**	**Topic 79**	**Topic 80**	**Topic 81**	**Topic 82**	**Topic 83**	**Topic 84**	**Topic 85**	**Topic 86**	**Topic 87**	**Topic 88**	**Topic 89**	**Topic 90**
increas	nasdaq	fell	industri	report	billion	use	gain	take	come	may	auto	concern	bond	price
consum	vol	drop	dow	loss	debt	line	climb	servic	global	feder	ford	announc	treasuri	fear
mani	index	tumbl	jone	post	cash	larg	push	step	state	keep	chrysler	today	sinc	energi
foreign	sampp	credit	composit	result	merg	found	advanc	network	requir	critic	vehicl	earli	yield	cent
april	trea	mix	tech	warn	assum	need	rebound	provid	payment	limit	truck	eas	longterm	send

Topic 91	Topic 92	Topic 93	Topic 94	Topic 95	Topic 96	Topic 97	Topic 98	Topic 99	Topic 100
market	former	continu	ceo	dollar	week	rate	amid	face	day
say	account	rise	top	mark	last	interest	analyst	consid	hit
settl	alleg	follow	morgan	despit	capit	lower	high	pressur	war
toward	sec	agenc	turn	signal	hurt	low	show	put	recess
quick	probe	grow	caus	currenc	restructur	edg	weak	broad	suppli

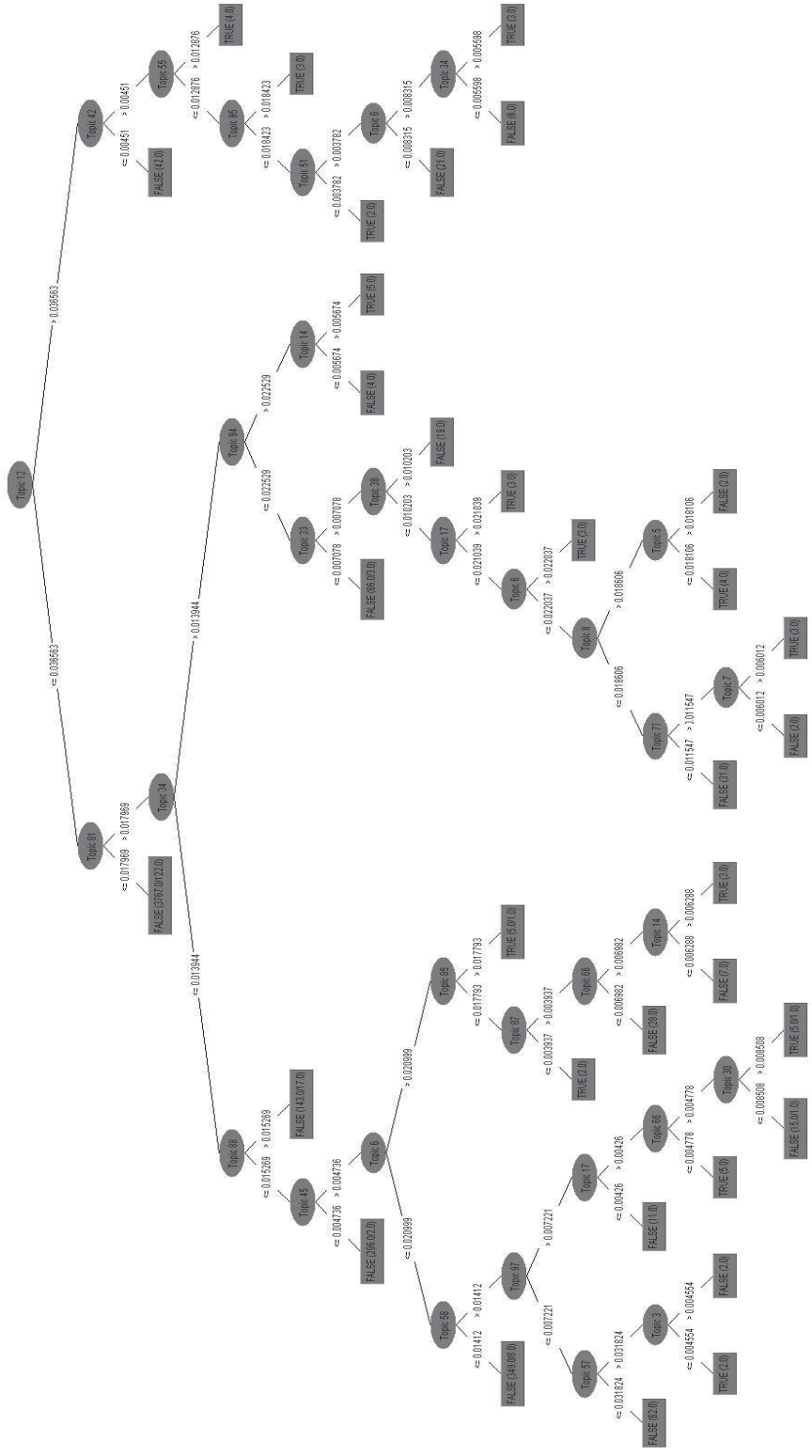

Fig. 4 Sample decision tree generated

ACKNOWLEDGMENT

This research is performed for The George Washington University in partial fulfillment of the requirements for the Doctor of Philosophy degree.

REFERENCES

[1] Weber, R., Aha, D. W., & Becerra-Fernandez, I. (2001). Intelligent lessons learned systems. Expert systems with applications, 20(1), 17-34.
[2] Ahn, J. J., Oh, K. J., Kim, T. Y., & Kim, D. H. (2011). Usefulness of support vector machine to develop an early warning system for financial crisis. Expert Systems with Applications, 38(4), 2966-2973.
[3] Kim, T. Y., Oh, K. J., Sohn, I., & Hwang, C. (2004). Usefulness of artificial neural networks for early warning system of economic crisis. Expert Systems with Applications, 26(4), 583-590.
[4] Oh, K. J., Kim, T. Y., & Kim, C. (2006). An early warning system for detection of financial crisis using financial market volatility. Expert Systems, 23(2), 83-98.
[5] Exchange, C. B. O. (2009). The CBOE volatility index-VIX. White Paper, 1-23.
[6] Nassirtoussi, A. K., Aghabozorgi, S., Wah, T. Y., & Ngo, D. C. L. (2014). Text mining for market prediction: A systematic review. Expert Systems with Applications, 41(16), 7653-7670.
[7] Onsumran, C., Thammaboosadee, S., & Kiattisin, S. (2015). Gold Price Volatility Prediction by Text Mining in Economic Indicators News. Journal of Advances in Information Technology Vol, 6(4).
[8] Nassirtoussi, A. K., Aghabozorgi, S., Wah, T. Y., & Ngo, D. C. L. (2015). Text mining of news-headlines for FOREX market prediction: A Multi-Layer Dimension Reduction Algorithm with semantics and sentiment. Expert Systems with Applications, 42(1), 306-324.
[9] Huang, C. J., Liao, J. J., Yang, D. X., Chang, T. Y., & Luo, Y. C. (2010). Realization of a news dissemination agent based on weighted association rules and text mining techniques. Expert Systems with Applications, 37(9), 6409-6413.
[10] Cheung, C. F., Lee, W. B., Wang, W. M., Wang, Y., & Yeung, W. M. (2011). A multi-faceted and automatic knowledge elicitation system (MAKES) for managing unstructured information. Expert Systems with Applications, 38(5), 5245-5258.
[11] Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. Journal of the American society for information science, 41(6), 391.
[12] Strait, M. J., Haynes, J. A., & Foltz, P. W. (2000, July). Applications of latent semantic analysis to lessons learned systems. In Intelligent Lessons Learned Systems: Papers from the AAAI Workshop (pp. 51-53). AAAI, Menlo Park, CA.
[13] Hollum, A. T. G., Mosch, B. P., & Szlávik, Z. (2013). Economic sentiment: Text-based prediction of stock price movements with machine learning and wordnet. In Recent Trends in Applied Artificial Intelligence (pp. 322-331). Springer Berlin Heidelberg.
[14] Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. the Journal of machine Learning research, 3, 993-1022.
[15] Mahajan, A., Dey, L., & Haque, S. M. (2008, December). Mining financial news for major events and their impacts on the market. In Web Intelligence and Intelligent Agent Technology, 2008. WI-IAT'08. IEEE/WIC/ACM International Conference on (Vol. 1, pp. 423-426). IEEE.
[16] Jin, F., Self, N., Saraf, P., Butler, P., Wang, W., & Ramakrishnan, N. (2013, August). Forex-foreteller: Currency trend modeling using news articles. InProceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 1470-1473). ACM.
[17] Quinlan, J. (1993). R. (1993) C4. 5: Programs for machine learning.
[18] Chen, N., Ribeiro, B., Vieira, A. S., Duarte, J., & Neves, J. C. (2011). A genetic algorithm-based approach to cost-sensitive bankruptcy prediction. Expert Systems with Applications, 38(10), 12939-12945.
[19] Arun, R., Suresh, V., Madhavan, C. V., & Murthy, M. N. (2010). On finding the natural number of topics with latent dirichlet allocation: Some observations. In Advances in Knowledge Discovery and Data Mining (pp. 391-402). Springer Berlin Heidelberg.
[20] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA data mining software: an update. ACM SIGKDD explorations newsletter, 11(1), 10-18.