Knowledge-Based Systems 114 (2016) 128–147

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

A survey of the applications of text mining in financial domain

B. Shravan Kumar[a,b], Vadlamani Ravi[a,*]

[a] Centre of Excellence in Analytics, Institute for Development and Research in Banking Technology, Castle Hills Road No. 1, Masab Tank, Hyderabad, 500057, India

[b] School of Computer & Information Sciences, University of Hyderabad, Hyderabad, 500046, India

ARTICLE INFO

Article history:
Received 24 April 2016
Revised 29 September 2016
Accepted 4 October 2016
Available online 5 October 2016

Keywords:
Text mining
Financial applications
FOREX rate prediction
Stock market prediction
Customer relationship management
Cyber security

ABSTRACT

Text mining has found a variety of applications in diverse domains. Of late, prolific work is reported in using text mining techniques to solve problems in financial domain. The objective of this paper is to provide a state-of-the-art survey of various applications of Text mining to finance. These applications are categorized broadly into FOREX rate prediction, stock market prediction, customer relationship management (CRM) and cyber security. Since finance is a service industry, these problems are paramount in operational and customer growth aspects. We reviewed 89 research papers that appeared during the period 2000–2016, highlighted some of the issues, gaps, key challenges in this area and proposed some future research directions. Finally, this review can be extremely useful to budding researchers in this area, as many open problems are highlighted.

1. Introduction

Nowadays, in Internet-dependent world, enormous amount of data is generated from several sources. Of this, a huge amount of data is available in an unstructured format. Analyzing the unstructured data using Text mining and data mining techniques can enable better decision-making. Text mining includes a lot of tasks such as document clustering, document classification, text summarization, sentiment analysis, social network analysis, topic detection, web page classification, identification of author, plagiarism detection, phsihing/spam/malware analysis, patent analysis, financial decision making, etc.

Traditionally, finance domain is replete with studies on FOREX rate prediction and stock market prediction. However, the ever-increasing dependence on technology and availability of vast amount of customer related data not only enabled financial industry to solve many customer related problems such as customer acquisition, market basket analysis, churn prediction etc., effectively but also to provide cyber security by efficiently solving assoicated problems such as phishing/spam/malware detection, fraud detection and intrusion detection. Undoubtedly, CRM and cyber security have become quinessential parts of financial industry. Therefore, CRM and Cyber security are included as important dimensions of the survey.

In this paper, a comprehensive review of research works dealing with various financial applications of Text mining is presented. These applications include FOREX Rate Prediction, Stock Market Prediction, Customer Relationship Mangement (CRM) subsuming Churn Prediction, and cyber security subsuming Phishing Detection, Spam Detection, Malware Detection, Fraud Detection and Intrusion Detection.

The fundamental challenge in Text mining is the unstructured format of data. We need to convert it into a structured format before starting the data mining process. The contribution of this study is reviewing the past works comprehensively with respect to dimensions such as (i) The relevant data mining techniques applied in developing predictive models (ii) data sources used for their analysis and (iii) prediction accuracy measures employed.

Financial forecasting domain is an interdisciplinary field consisting of Data Mining (subsuming statistics and machine learning), Text mining, Natural language processing (NLP), and Behavioral Economics. Data mining algorithms such as Support Vector Machines, Linear Regression, Logistic regression, neural networks, Naive Bayes and Decision Trees play a predominant role in this field.

Earlier to the current study, extant surveys on Text mining dealt with individual applications such as phishing/spam/malware detection, financial market etc. However, till date, to the best of our knowledge, there is no single review paper, where all the above financial applications including CRM are reviewed. Therefore, this study makes an attempt to address this gap. This paper also discusses the advantages and disadvantages of various methods used.

* Corresponding author. Fax: +914023535157.
E-mail addresses: bskumar@idrbt.ac.in (B.S. Kumar), rav_padma@yahoo.com, padmarav@gmail.com (V. Ravi).

http://dx.doi.org/10.1016/j.knosys.2016.10.003
0950-7051/

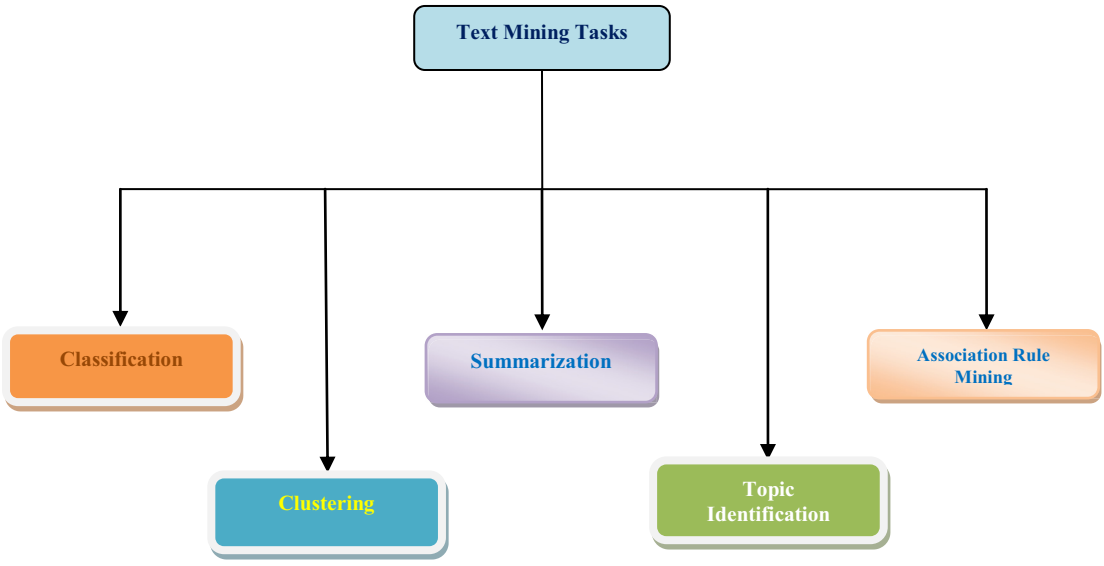

Fig. 1. Primary tasks of Text mining.

The rest of the paper is structured as follows: Section 2 describes the Text mining tasks. Section 3 briefly describes the previous surveys. Review methodology is presented in Section 4. An overview of the most commonly using data mining techniques is presented in Section 5. Later, in Section 6, various applications of Text mining in financial domain are systematically reviewed. Finally, we present conclusions and future research directions in Section 7.

2. Text mining tasks

According to Dorre, Gerstl, and Seiffert [40], text mining applies mostly the same techniques of data mining to the corpus of textual data after converting it to a structured format. Text mining techniques extract the knowledge from text documents. The Text mining process consists of two steps: text preprocessing and knowledge extraction. In short, text preprocessing step converts unstructured data into a document-term matrix, and knowledge extraction step involves data mining. The preprocessing step involves tokenization, stop words removal, stemming, etc. followed by the formulation of a document-term matrix, as proposed by Salton and McGill [111]. In this form, documents are represented in rows and words in columns. There are various ways of forming a document-term matrix based on the term weight. Salton and Buckley [110] presented three different ways of presentations viz., Binary Term (BT), Term Frequency (TF), Term Frequency-Inverse Document Frequency (TF-IDF). Given the high dimensionality of the data, feature selection or dimensionality reduction is performed before invoking a classifier.

We categorized the Text Mining tasks, depicted in Fig. 1, broadly into five categories; Classification, Clustering, Association Rule Mining (ARM), Text summarization and Topic detection/ identification. Most of the text mining applications fall into these categories.

Preprocessing plays a significant role in Text mining. Any task of the Text mining fully depends on the preprocessing. High-quality of preprocessing is always yielded superior results. Uysal and Gunal [126] presented a paper on the text preprocessing and its effect on the classification task. They examined on two datasets (news and email) with two different (English and Turkish) languages. They experimented with various combinations of preprocessing.

Researchers are trying to improve the elementary vector space model that leads to Part of Speech (PoS) tagger [20]. Through this process, for each document, the words are assigned to the PoS identifier (noun, verb, adjective, etc.). Sentence structure is useful for creation of tags.

Another, simple model for the text representation is Bag of words [58]. In this model, text is represented as a set of words. This is the most common approach in text mining and used in various works [72]. Another method is n-gram model [116]. It is a sequence of a set of terms (syllables, words, gesture, etc.) in a text. Based on the value of n, there exist different names i.e. Unigram (n is 1), Bigram (n is 2), Trigram (n is 3), etc. This model is applied not only in text mining but also applied in various other fields like Communication Theory, Computational Biology, and Data Compression techniques, etc.

Though Text mining was started in 1960s, it became popular in 1990s, as it was identified as the primary field of Information systems. The introduction of Machine Learning algorithms for conducting Text mining tasks reduced human intervention drastically as well as considerable amount of time to process the text. The beginning of text mining task i.e. document classification was carried out in 1960s [88]. One of the initial applications of Text mining was developed by Borko and Bernick [18] for automation of document indexing. Documents have one or more keywords for content description. These words are a subset of controlled dictionary which contains the collection of synonyms and concepts.

Word Sense Disambiguation (WSD) is another application of Text mining which helps us to find out whether the text is ambiguous or not. It also helps us to understand the meaning of the word. WSD has a lot of applications including, navigation in hyperlinks, spell checker, summarization, etc. To solve these issues, Escudero et al. [41] invoked the AdaBoost algorithm. They experimented with Linguistic corpus which consists of 192,800 documents and in which they identified that 191 are the frequent ambiguous words.

Ontologies are also another frequently using technique in text mining. It contains the domain knowledge and these are in the form of relationships [62] between the entities i.e. concepts and their level order. WorldNet is the one of example for Ontology which contains 110 K unique concepts and 150 K words. Feldman and Hirsh [44] developed a system called FACT (Finding Associations in Collections of Text). It discovers the co-occurrence and associations of the terms in the text corpus.

Feature selection is a major step in text mining. Removing redundant and unnecessary features and retaining relevant/ important features is called feature selection. Literature abounds with several feature selection techniques. Yang and Pedersen [133] conducted a comparative study of feature selection methods in statistical learning as applied to text categorization. The focus is on dimensionality reduction. Five methods were evaluated, including term selection based on Document Frequency (DF), Information Gain (IG), Mutual Information (MI), Chi-square (Chi) statistic, and Term Strength (TS). They found IG and CHI most compelling. They experimented with *k*-NN classifier on the Reuters corpus.

3. Previous related surveys

In this section, we describe the previous surveys related to Text mining and its applications.

Most recently, Ittoo et al. [68] performed a review on the applications of text analytics in various industries. They majorly classified the applications into categories such as market research, automatic tweet generation, domain-specific search engine, case-based reasoning, aviation safety reports, and health related content. The study reviewed the datasets, tools employed in these individual applications and also compared the performance of the methods.

The primary assumption of the data mining algorithms is that the training and test data have the same distribution. If the distribution of the features of the test data changes from that of the training data, then we get inaccurate predictions on test data. To avoid this difficulty, transfer learning is proposed, where, a model is trained one domain and tested on another domain unlike the regular machine learning algorithms. Lu et al. [85] presented a survey paper on Transfer learning. They categorized the Transfer learning techniques into three parts namely neural network based, Fuzzy based and Bayes based and described its applications in Finance, Image processing, NLP, etc. of transfer learning.

Lim et al. [83] presented the research trends of Business Intelligence and Analytics (BIA) that are mainly focused on three major research areas namely Big data, text analytics and network analytics. They also presented the futuristic models and their applications with respect to the BIA.

Finding the groups of similar objects based on their pattern of text is called text clustering [70]. It has various applications like customer segmentation, outlier detection, recommender system, etc. Aggarwal and Zhai [5] presented a survey paper on text clustering algorithms. They explained various types of clustering algorithms (partitioning, agglomerative, etc.) and different feature selection methods like document frequency, term strength, Non-negative Matrix Factorization (NMF), etc. They also mentioned new explorations of the applications of text clustering such as online chat/ social network analysis and diverse applications (Flickr etc.).

Assigning of text documents to the predefined class is called text categorization (classification). Sebastiani [115] presented a survey of text categorization. In this work, they discussed text categorization concerning single as well as multiple classes. An application of document classification includes organization of documents, language processing, web pages categorization, etc. They described various classifiers such as NB, DT, and NN and evaluated these using different performance metrics viz., Accuracy, Precision, and Recall using some of the benchmark datasets (Reuters, etc.) for text categorization. Finally, they analyzed the previous results of few classifiers.

The World Wide Web (WWW) is the information super highway. Kosala and Blockeel [77] presented a survey paper on web mining. In this article, they described the methodology of Web mining as follows: Documents retrieval, Preprocessing, and Analysis. The Web mining categorized into three types- Content mining, Structure mining, and Usage mining.

Table 1
List of papers reviewed for survey.

Source	No. of papers
ACM	20
Blackwell	2
Elsevier	39
IEEE	11
IOS	1
John and Wiley	2
Kluwer Academic	2
Springer	10
USENIX	3
Total	89

Thus, it is clear that there is not comprehensive review yet that exclusively deals with text mining applications to finance domain. Therefore, the current review fills that gap.

4. Review methodology

To conduct the present literature survey on the applications of Text mining to financial domain, we collected information from various sources available on the web. Primarily, we referred to databases such as ACM digital library, Taylor and Francis, Science Direct, Wiley, Google Scholar, Springer and IEEE-Xplore. Book chapters and edited volumes are excluded from the scope of the review. We refined our search using various words like “Text mining”, “Applications of Text mining”, “Stock market prediction + text”, etc. We extracted 89 papers for this survey that were published during 2000–2016 in various conferences and journals. The scope of the current survey covers applications of all Text mining tasks in the domain of financial services and the topics include FOREX rate prediction, Stock market forecasting, Customer Relationship Management and Cyber security subsuming Phishing detection, Spam detection, Fraud detection, Malware detection, and intrusion detection. This survey does not include Text mining applications in non-financial service industries such as Telecom, Healthcare, Hospitality, Retail, Manufacturing, and Travel.

The various publication resources of this survey and distribution of these are presented in Table 1. The review comprises articles mostly from Elsevier followed by ACM, Springer, and IEEE so on. Similarly, we presented the distribution of articles Journal wise in Table 2. Expert Systems with Applications turns out to be having a higher number of articles related to the paper, followed by Decision Support Systems, Knowledge-Based Systems, Computing Surveys. In the same way, KDD conference occupies top position in the list of conferences presented in Table 3. It is followed by, ECML, AAAI and so on. The number of papers reviewed from 2000 to 2016 is shown in Fig. 2. Highest number of articles (13) appeared in 2010, followed by 2007 and 2008. We also included papers from the current year also.

The Fig. 2 depicts the year-wise distribution of the papers. The list of the different methods applied to Text mining is depicted in Fig. 3, which consists of the regular classification methods i.e. DT [105], SVM [30], NB [39], and *k*-NN [32]. Among various approaches, SVM is found to be employed most often and is depicted in Fig. 5, followed by, NB, Neural Network, Decision Tree, Linear Regression and Logistic Regression etc. Similarly, among the regular classifiers, SVM is the most popular technique followed by NB and DT. List of abbreviations are described in Table 4. The performance metrics for evaluation of the models are presented in Table. 5. Similarly, in Fig. 6 we presented the corpora used in various applications where, Financial News is occupying the first position.

There are various types of data available online. A huge amount of data is generated every day. Consequently, there is a need for more sophisticated techniques to analyze the data. The schematic

Table 2
Distribution of articles in journals.

Journal Name	No. of papers
Communications Surveys & Tutorials	1
Computers & Operations Research	1
Computers in Industry	1
Computing Surveys	2
Data Mining and Knowledge Discovery	1
Decision Support Systems	10
Engineering Applications of Artificial Intelligence	1
European Journal of Operational Research	1
Expert Systems with Applications	15
Foundations and Trends in Information Retrieval	1
Information Processing & Management	1
Journal of Accounting Information Systems	1
Journal of Accounting Research	1
Journal of Artificial Intelligence Review	1
Journal of Computational Science	1
Journal of Computer Security	1
Journal of Emerging Technologies in Accounting	1
Journal of Finance	2
Journal of Financial Economics	1
Journal of Information Assurance and Security	1
Journal of Information Management	1
Journal of Intelligent Systems in Accounting, Finance and Management	1
Journal of Systems and Software	1
Journal of International Money and Finance	1
Knowledge-Based Systems	3
Management Science	1
Transactions on Asian Language Information Proessing	1
Transactions on Information Systems	1
Transactions on Internet Technology	1
Transactions on Systems, Man, and Cybernetics	1
Web Semantics: Science, Services and Agents on World Wide Web	1
Total	58

Table 3
Distribution of articles in conferences.

Conference name	Abbreviation	No. of papers
Association for the Advancement of Artificial Intelligence	AAAI	2
Annual Computer Security Applications Conference	ACSAC	1
Association of Computational Linguistics	ACL	1
Australasian Computer Science Conference	ACSC	1
Canadian Conference on Artificial Intelligence	Canadian AI	1
Conference on Email and Anti-Spam	CEAS	2
Human-Computer Interaction	CHI	1
Computational Linguistics	COLING	1
Conference on Automation Science and Engineering	CASE	1
Detection of Intrusions and Malware, and Vulnerability Assessment	DIMVA	1
European Conference on Machine Learning	ECML	2
Empirical Methods in Natural Language Processing	EMNLP	2
Genetic and Evolutionary Computation Conference	GECCO	1
Advances in Pattern Recognition	ICAPR	1
Neural Information Processing	ICONIP	1
Pattern Recognition	ICPR	1
Joint Conference on Neural Networks	IJCNN	1
Knowledge Discovery and Data Mining	KDD	6
Pacific Asia Conference on Knowledge Discovery and Data Mining	PAKDD	1
Research in Attacks, Intrusions and Defenses	RAID	1
Symposium on Principles of Distributed Computing	PODC	1
Workshop on Recurring Malcode	WORM	1
Total		31

view of Text mining applications in finance that are within the scope of the review is depicted in the Fig. 4, while Fig. 5 depicts the distribution of various data mining techniques. Fig. 6 depicts different types of data analyzed in various works reviewed here.

5. Overview of data mining techniques

In this section, we briefly describe some of the traditional machine learning algorithms that are frequently employed in the financial applications along with Text mining. However, complete details of these techniques is intended to be presented here as these are well-known in the Machine learning as well as Data mining community.

5.1. Support vector machine (SVM)

Support Vector Machine, proposed by Vapnik [127], performs the classification task by constructing the hyperpalne in such a way that linearly separable data will be classified into two categories. For dealing with non-linear data, it uses the kernel (sig-

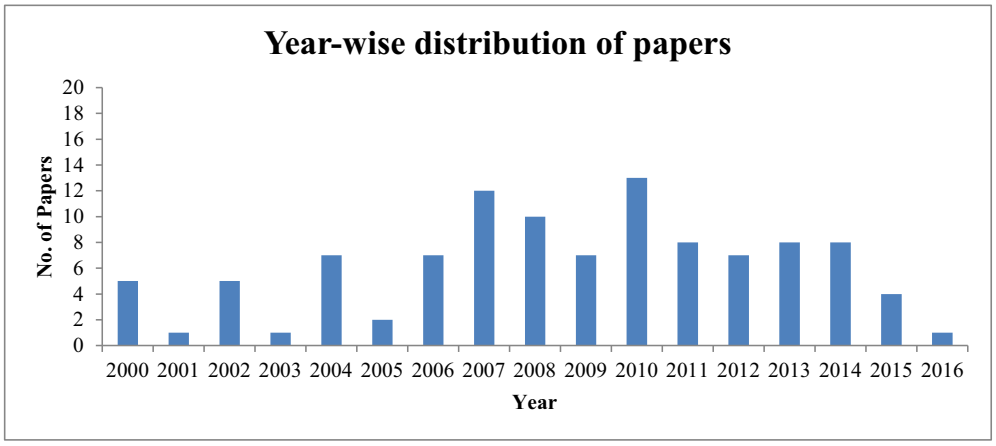

Fig. 2. Papers distribution.

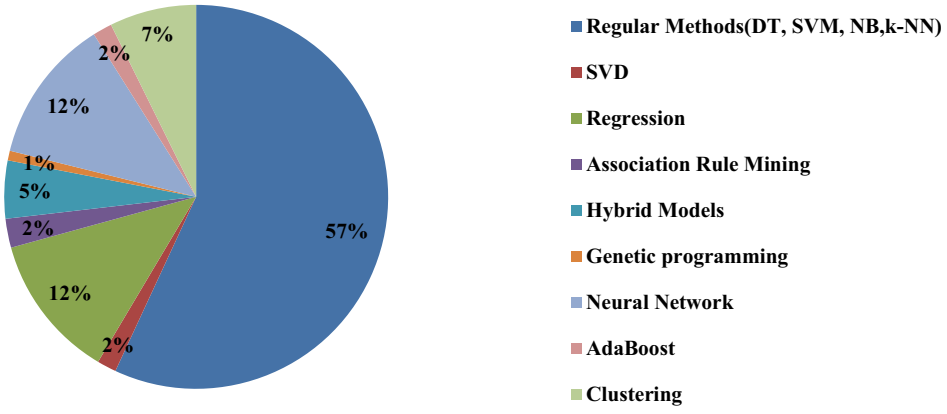

Fig. 3. Approaches in various papers.

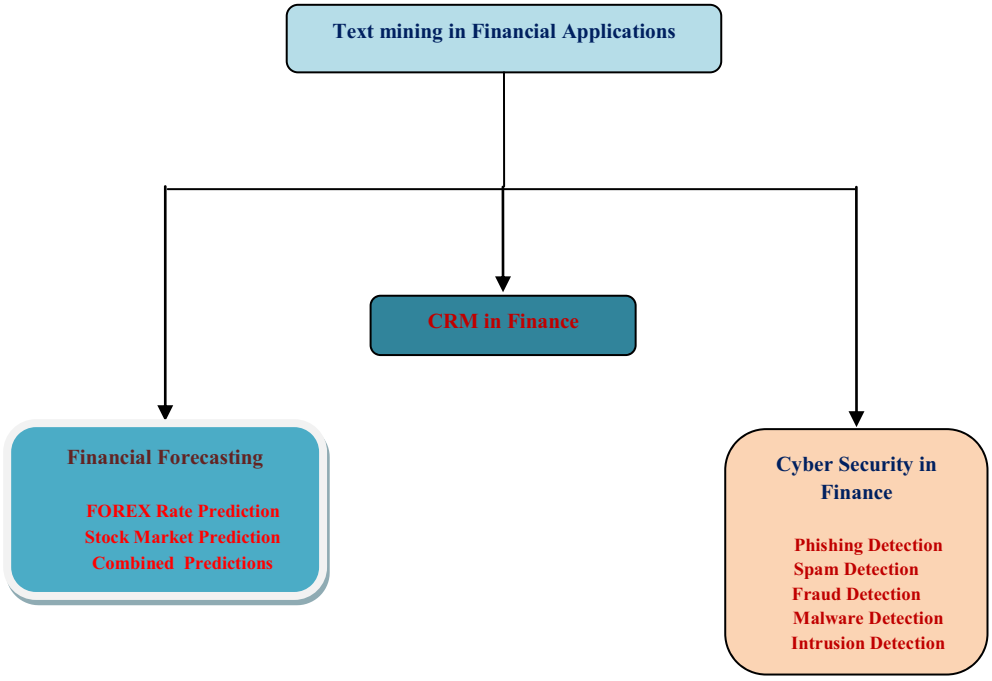

Fig. 4. Financial applications with Text mining.

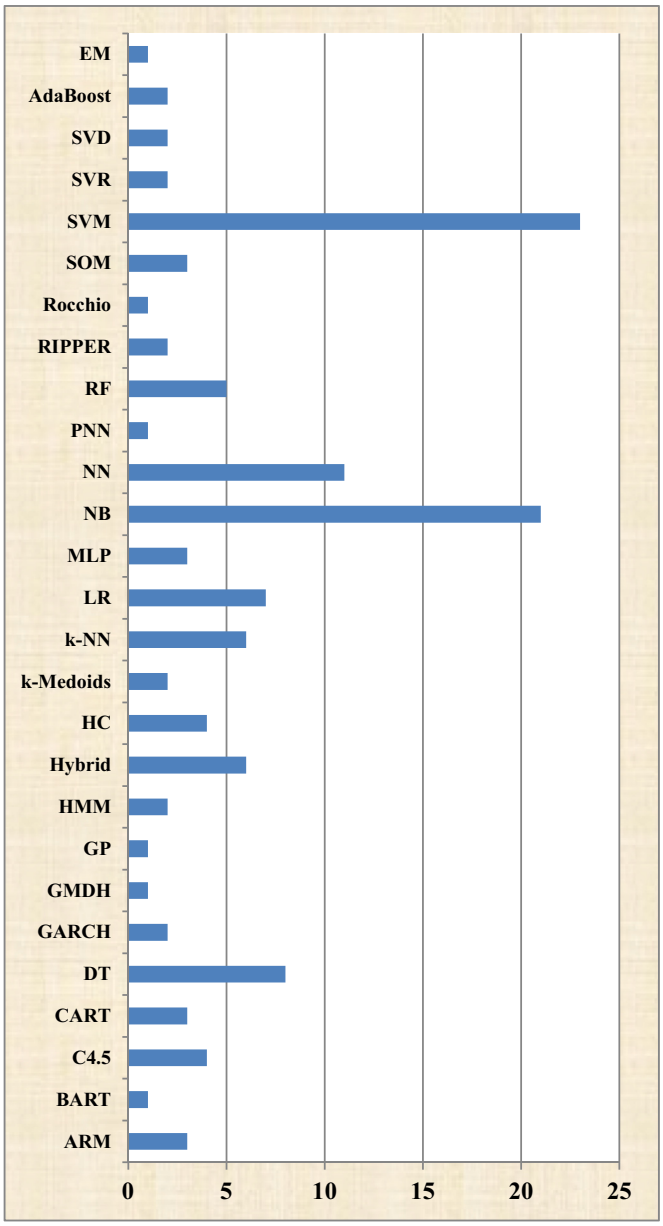

Fig. 5. Methods distribution in various works.

Table 4
Abbreviations (alphabetical order).

Abbreviation	Interpretation
ANN	Artificial neural network
ARM	Association rule mining
BART	Bayesian additive regression trees
CART	Classification and regression tree
DT	Decision tree
GARCH	Generalized autoregressive conditional heteroskedasticity
GP	Genetic programming
GMDH	Group method data handling
HMM	Hidden Markov model
HC	Hierarchical clustering
k-NN	*k*-nearest neighborhood
LR	Linear regression
MLP	Mult-layer perceptron
NB	Naive Bayes
NN	Neural network
OLSR	Ordinary least squares regression
PNN	Probabilistic neural network
RF	Random forest
RIPPER	Repeated incremental pruning to produce error reduction
ROC	Recievers operating characteristics
SVM	Support vector machine
OWL	Web ontology language

moid, radial, polynomial) function for projection of the data into higher dimensions so that data can be linearly separable. It performs the modeling task by separating the samples of one class variables on one side, and other samples are on the other aspect of the plane. The samples near to the plane are called support vectors. SVM identifies the support vectors such that the distances between them were maximum. The performance of the SVM depends on the selection of the kernel and its user-defined parameters.

5.2. Naive Bayes (NB)

Naive Bayes is based on the principle of Bayes theorem. It assumes that the predictors are independent i.e. complete independence among the features set. Due to its predicting capability, it is also employed quite frequently. It is suitable for binary class as well as multi-class problems [39].

5.3. k-nearest neighborhood (k-NN)

The *k*-NN algorithm finds the *k* number of samples in the training which are nearer to the test samples [45]. In this method, three components are playing a key role: data samples, distance metric and number of the neighbors i.e. *k*-value. For any classification task, initially, it computes the distance between the unlabeled data

Table 5
Performance metrics for evaluation of the models.

Metric	Interpretation
$Accuracy = \frac{TP+TN}{TP+FP+TN+FN}$	The number of correctly classified samples on test data.
$Sensitivity = \frac{TP}{TP+FN}$	The number smaples correcttly classified as positive class
$Specificity = \frac{TN}{TN+FP}$	The number of samples correctly classified as negative class
$\Pr ecision = \frac{TP}{TP+FP}$	Positive predicted value i.e. retrieved instances that are relevant
$Recall = \frac{TP}{TP+FN}$	It is same as Sensitivity or we can say relevant instances that are retrieved
$F-Score = \frac{2*\Pr ecision*Recall}{\Pr ecision+Recall}$	It is the harmonic mean of precision and recall
$AUC = (Sensitivity + Specificity)*0.5$	Average value of sensitivity and specificity
$MSE = \frac{\sum_{i=1}^{n}(Actual_i - \Pr edicted_i)^2}{n}$	It tells how the fitted line is close with the data points
$MAPE = \frac{1}{n}\sum_{i=1}^{n}\frac{\lvert Actual_i - \Pr edicted_i\rvert}{Actual_i} * 100$	Mean absolute relative error in terms of percentage

TP = True Positive, TN = True Negative, FP = False Positive, FN = False Negative

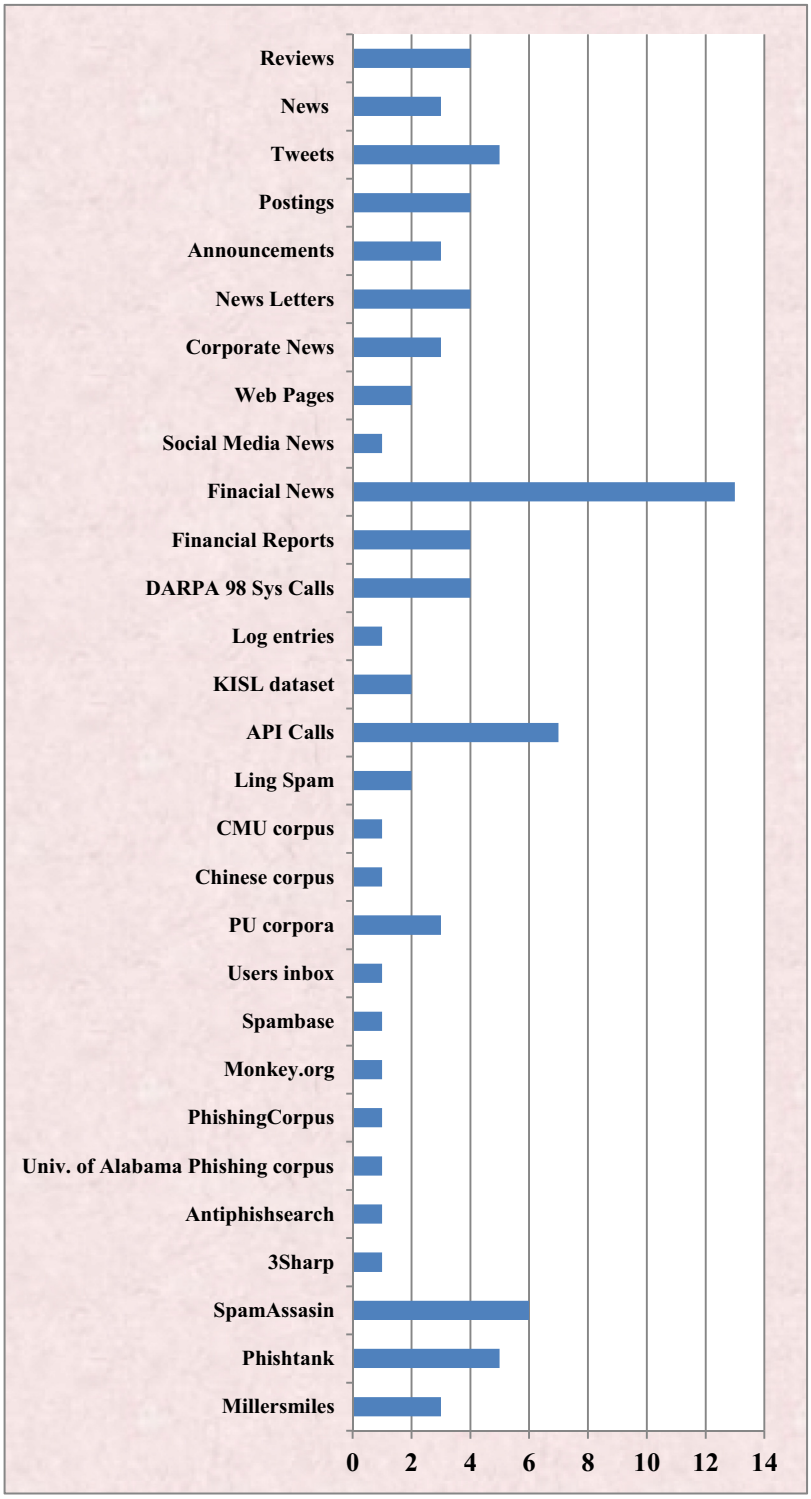

Fig. 6. Datasets used in different works.

samples and other labeled samples. Based on the computed distance, labeled data sample is assigned to the nearest labeled sample.

5.4. Decision tree (DT)

Decision Tree, introduced by Quinlan [104], is one of supervised learning methods and widely used in many of the data mining applications. It consists of a simple structure where each non-terminal node represents the test condition, and terminal or leaf node represents the class output. It produces "if-then" rules that are very much helpful in decision-making process. Usually, the root node can be picked up using Gini Index, Information Gain, Gain Ratio, or Entropy.

5.5. Multilayer perceptron (MLP)

MLP is one of the frequently used feed forward neural networks [108] for solving classification and prediction problems. It consists of three layers- input, output and hidden. Input layer consists of the number of input variables of a dataset. One or more hidden layers accept the content from input layer. Later, the output layer receives the output of hidden layer and produces the class label or prediction. Hidden layer serves the following purposes: updating the weights and activating the (nonlinear) functions that help us to capture the nonlinearity of the data. It is based on backpropagation algorithm. It is too popular to be described here.

5.6. Linear regression (LR)

It is one of the methods for finding the relationship between independent variables and discrete dependent variable. Linear Regression [48] deals with one independent variable and one continuous dependent variable. Similarly, the regression that deals with two or more independent variables and one dependent variable is known as multiple linear regression.

5.7. Classification and regression trees (CART)

Most of the Decision Tree algorithms are useful for solving the classification problems. CART is having the capability to handle both classification and regression problems [19]. It also generates the "if-then" rules. It is powerful than other Decision Tree algorithms since it deals with real values, enumerated type data as well as missing data also. It also follows the same structure of Decision Tree for splitting of a node and all other things except the output. In CART the output may contain discrete as well as continuous values.

5.8. Group method of data handling (GMDH)

GMDH is one of the powerful neural network architectures for modeling the complex problems [69]. It is the first deep learning neural network architecture. It works based on the construction of polynomial terms, and the numbers of layers are determined by the genetic component. The output depends on the combination (polynomial) of the inputs.

Quadratic form of GMDH is as follows:

$$Q = c_0 + c_1x_1 + c_2x_2 + c_3x_1^2 + c_4x_2^2 + c_5x_1x_2$$

Where x_1 and x_2 are the inputs, c_i is the weight vector, and Q is the output of the node. The weights are found by solving the regression equations with Q=Y, the response vector.

It trains the model based on polynomial function such that error in the output variable is minimum. It builds polynomials repeatedly, and the algorithm selects the best one among them. The process is complete only when the algorithm meets the selection criteria.

6. Review of financial applications of Text Mining

The current survey highlights various applications of Text mining in finance. The applications are broadly categorized into FOREX Rate Prediction, Stock Market Prediction, CRM applications and Cyber Security in finance. We describe all the works related to each application separately here in this section. At the end of the each and individual application, key observations drawn from them are added.

6.1. Financial forecasting applications

In this section, we described the works where either FOREX Rate or Stock Market prediction was studied. It also reviews the papers, where both studies are made.

6.1.1. FOREX rate prediction

The foreign exchange (FOREX) market changed drastically over the past years. The basic idea of this study is to automate the human thinking and other aspects that anticipate the direction of financial market movements before making an investment decision. An investor should carefully study the historical trends in financial markets and assess the current situation in order to predict the future [54]. Accurate predictions of FOREX rate indeed reduces the market risk emanating from the fluctuations in the FOREX rates.

Numeric time series data and textual data not only contain the effect, but also the possible causes of the event. Better predictions can be made by combining numeric data with text. A recent study on exchange rate forecasting proposed a new model based on the current status of world financial markets. Stock market fluctuation based on non-quantifiable information, reveals the impact of news reports on the time series based on efficient market hypothesis. Fung et al. [46] proposed a new statistical based partition algorithm using Hierarchical clustering and *t*-test piecewise segmentation algorithm to identify the trends in time series. Based on this algorithm, the partitioned trends were clustered into two groups i.e. rise and drop based on the slope of trends and the coefficient of determination. In order to filter the news article with the help of clusters obtained from trends a guided clustering algorithm was proposed which was based on incremental k-Means. They also proposed new differentiated weighting scheme that assigns higher weights to the features occurred in the rise.

Evans and Lyons [42] also experimented with macro news for studying the currency flow. In this work, they observed that the arrival of macro news can account for more than 30% of daily price variance. By considering the macro news, it was observed that the market participants and the prices of the market were affected by the arrival of macro news. They concluded that macro news impacted two thirds of the exchange rate.

Vu et al. [129] proposed a model for price movements of the stocks based on Twitter messages. They labeled the sentiment into two categories – positive and negative. Based on this, they predicted stock price of four companies viz., Amazon, Apple, Microsoft, and Google with past 41 days' data. They employed DT for the classification task under 10 Fold Cross Validation (10-FCV).

Jin et al. [71] proposed a model called for Forex-foreteller which mines news articles and forecasts the movement of foreign currency markets. A combination of language models, topic clustering, and sentiment analysis was used to identify the news articles. These were combined with historical stock index and currency exchange values for prediction. They employed linear regression model for currency forecasting.

Yu et al. [136] studied the impact of social media on the stock market performance of a company using NB classifier. A total of 824 firms belonging to six industries' (Pharmacy, Software, Health sector, Hotel, and Savings institutions) were analyzed using the postings collected from various sources (forums, blogs, and Twitter). They used stock return value and risk as the performance metrics for evaluation. For each document, sentiment score was calculated based on these values.

Chatrath et al. [26] carried out a study on currency up and downs based on macro news. They analyzed the rates of four currencies for 2005–2010 with intra-day and frequency sample of 5-min duration. Currency jumps are a good stand-in for news arrival, and they found that 9–15% of currency changes were directly affected by U.S. announcements. News effect explains 22–56% of the

Table 6
Summary of different studies on stock market prediction and FOREX rate forecasting studies.

Study	Dataset	Model	Performance measure
FOREX rate forecasting/ prediction			
Chatrath et al. [26]	Financial news	MV regression	Mean, SD
Jin et al. [71]	Financial news	LR	Precision/recall
Yu et al. [136]	Social media news	NB	Positive/negative, F-measure
Vu et al. [129]	Tweets	DT	Up/down, accuracy
Evans and Lyons [42]	Macro news	Heteroskedasticity	Up/ down
Fung et al. [46]	Corporate news	SVM	Up/down
Summary of different studies of stock market prediction			
Nassirtoussi et al. [96]	News headlines	Multilayer model	Accuracy
Moniz and Jong [94]	Dow Jones newswires	RF	Precision, recall and F-score
Nizer and Nievola [98]	Stock news	GARCH	Sensitivity/ specificity/ accuracy
Chan and Franklin [25]	Bloomberg.com, Quamnet.com	HMM, DT	Sensitivity
Wang et al. [131]	Financial news, Shangai stock exchange	NB	Precision
Gilbert and Karahalios [52]	LiveJournal site, S & P 500 Index	Monte Carlo simulation	p-value
Mellouli et al. [90]	Financial news	Ontology + Bayesian networks	–
Fasanghari and Montazer [43]	Tehran stock exchange	Fuzzy modeling	Positive/ negative
Dey et al. [37]	Financial news	LDA	Positive/ negative
Wang et al. [132]	Finacial news	Ontology (WOL)	–
Koppel and Shtrimberg [76]	S & P news	SVM	Accuracy
Kloptchenko et al. [75]	Financial reports	SOM	Return on equity
Fung et al. [47]	Hang Seng index, reuters	SVM	Positive/ negative
Back et al. [11]	Annual reports	SOM	Error rate
Lavrenko et al. [78]	Yahoo news (biz.yahoo.com), Yahoo stock	Linear regression	Gain/ loss
Thomas and Sycara [124]	Textual data of the web	GA	Return value
FOREX rate and Stock Market prediction combined studies			
Hagenau et.al. [57]	Corporate news	SVM, SVR	Up/ down
Schumaker et al. [114]	US financial news	SVR	Error rate
Groth and Muntermann [55]	German ad hoc announcements	NB, k-NN, SVM, NN	Up/ down
Bollen et al. [17]	Tweets	SOFNN	Error rate
Huang et al. [65]	Financial news	Weighted AR	Up/ down
Li [79]	The US corporate filings	NB	Positive/ negative
Schumaker and Chen [113]	US financial news	SVM	MSE
Butler and Keselj [22]	S & P 500 index	n-gram + SVM	Good/ bad
Tetlock et al. [123]	The US financial news	Negative words ratio	Up/ down
Mahajan et al. [87]	Financial news	DT+SVM (hybrid)	Up/ down
Das and Chen [33]	The US message postings	Hybrid classifiers	Accuracy
Antweiler and Frank [9]	The US message postings	Bayes + SVM	Buy/ sell
Mittermayer [92]	US financial news	SVM	Good/ bad

5-min jump returns as a negative impact on currency change. Co-jump statistics are strictly dependent on macro news among European currencies, particularly in between Euro and Swiss franc. Some of the research works are carried out on FOREX rate prediction as well as Stock Market prediction also. They are discussed in Next Section.

Various works carried out on the FOREX Rate problems are summarized in terms of Model proposed, Performance measure used, and dataset analyzed in Table 6.

Some of the key observations from the section are as follows:

- Quick impacts of news articles (i.e. breaking news etc.) are better for short-term prediction. It helps us to find out the directional movement of the currency exchange rate on the intraday.
- It is observed that most of the works were explored with financial news and rarely with macro news. So, the combination of news articles with the macro news yields better predictions.
- A Good number of methods (SVM, DT, k-NN and NB) are being applying frequently. Not only these methods but also the other classification methods are also needed to be explored.
- Most of the works are related with restricted zone only.
- The majority of the works are considering the time when news articles released and if we consider the whole news articles it may lead to being the best prediction.

6.1.2. Stock market prediction

Stock market prediction is an interesting problem which involves data mining and statistics. Today's biggest economies of the world have higher stock market value. Every market is bounded by supply and demand equilibriums. Human intervention is limited in stock market prediction. Researchers have been working to predict the stock exchange prices. No methodology has been conceived to accurately predict the price movement. It is therefore difficult to predict the stock market price dynamically. The stock market behavior depends on news which is in the unstructured format. This extracted knowledge from the unstructured data is used for effective decision making.

Lavrenko et al. [78] presented a model to identify the news stories which affects the financial markets trend. They identified the patterns in time series with the help of piecewise linear fit followed by labels assignment with automated binning process. Thomas and Sycara [124] worked on the behavior of financial markets. Textual information available on the websites is impacting their business. They proposed two models based on maximum entropy and Genetic algorithm to predict financial markets. They concluded that the combination of these two models yielded better predictions than stand-alone models.

Back et al. [11] proposed a model based on Clustering approach and they employed neural network clustering algorithm called Self Organizing Maps (SOMs). They experimented with numerical data as well as combination of the numerical data with text. They concluded that this combination yielded best results.

Fung et al. [47] proposed a model for stock prediction using textual data and time series data mining. All of the existing approaches were based on mining only a single time series. They primarily experimented with multiple time series mining with text data in the proposed framework. Through this method, they identified inter-relationships between the time series. For evaluation

purpose, they conducted experiments on stock data and news articles of Reuters Market 3000 by employing SVM classifier.

By combining the data and Text mining techniques, financial reports were analyzed by Kloptchenko et al. [75]. They proposed a model based on SOMs for financial reports analysis. They investigated the three financial reports of major telecom companies. They collected the data from Motorola, Nokia and Ericsson companies which consist of both quantitative (financial ratios) and qualitative (textual data).

Koppel and Shtrimberg [76] proposed a model based on the news articles for stock prediction. They extracted the features from the Multex Significant Development corpus and experimented with the Standard & Poor 500 (S&P 500) stock index. During the process of modeling, they labeled the news as positive or negative. Later, they employed the SVM and reported an accuracy of 70%.

Wang et al. [132] proposed an Ontology-based model for stock market prediction. In the first phase, they presented a framework based on financial news, market investors, and financial liabilities and in the second phase, they found out the causality among the news articles and liabilities. Through this proposed approach, they evaluated the stock trading activity of China Petroleum Corporation (New York Stock Exchange) and 9/11 attack articles.

Dey et al. [37] proposed a model based on LDA for stock market analysis using financial news. The model identifies the events that would affect the stock market and its impact. They extracted the topics from the news articles and then clustered them with k-means algorithm. They analyzed groups within SENSEX raw data.

The prevalent conditions that exist in the economy as a whole, rather than in a particular region varies. In general, the macro environment includes trends in GDP, inflation, employment, monetary and fiscal policy. In our present review, we also discussed the works of researchers for financial market prediction based on global news. The key factor of stock market decision making is the selection of the right stock at the right time. Choice of the superior stocks for investment, a finite number of alternatives have been ranked considering several criteria. Multiple Criteria Decision Making (MCDM) has to solve these types of problems. Fasanghari and Montazer [43] proposed a fuzzy-based model for selecting better stocks to model the uncertainty.

Mellouli et al. [90] presented a model for financial headlines representation using ontology. The proposed model had dealt with 201 attributes that belong to 31 concepts. They concluded that the headline articles are almost of 99% correctly categorized with this approach. They evaluated the model with 227 financial headlines related to the different companies concerning the Toronto stocks.

Gilbert and Karahalios [52] performed research on how real world emotions affect the real world things. They demonstrated the estimation of emotions from weblogs information about the future stock market prices. Anxiety, worry, and fear are estimated from the 20 million posts on LiveJournal site. Based on these emotions, they predicted the movements of S&P 500 index. They analyzed the results of Monte Carlo simulation.

Wang et al. [131] proposed a framework for finding the correlation between news articles and financial liabilities based on Ontology. They preprocessed the news articles and financial liabilities using Ontology. By invoking the NB algorithm, they figured out the trading function. They classified the news events based on their type using Ontology. They concluded that the consideration of news articles and their polarity (positive, negative and neutral) is producing best results compared to considering polarity alone after experimenting with China Petroleum & Chemicals Corporation, SP Power Development, and Shangai Composite Index stocks.

Chan and Franklin [25] proposed a novel text-based decision support system which extracts event sequences from text patterns and predicts the likelihood of the occurrence of events using a classifier-based inference engine. They investigated more than 2000 financial reports with 28,000 sentences. Experiments showed that the prediction accuracy of the model outperformed similar statistical models by 7% for the seen data while significantly improving the prediction accuracy of the unseen data. Further comparisons substantiate the experimental findings. Nizer and Nievola [98] applied the Text mining techniques with GARCH model for predicting the volatility in the stock market behavior. They built the model on Portuguese news content about companies and their stocks and analyzed its effect on the Brazilian stock market.

Moniz and Jong [94] proposed a method based on the counting of negative terms from the dictionary and word counting approach to capture contextual information. LDA model was used to infer the negative effect. They identified the top words associated with the topic clusters. Random Forest algorithm was used for classification and F-Score used for the performance of the model. To test the efficacy of the proposed model, they used Corpus of Dow Jones Newswires.

Recently, Nassirtoussi et al. [95] conducted a survey on various methods for market prediction using Text mining. They examined various articles based on the data sources used, feature selection methods, applied techniques, and comparison of these techniques.

Nassirtoussi et al. [96] predicted FOREX market with the news headlines. They proposed a multi-layer dimension reduction technique with the help of semantics and sentiment. In their study, they found out that the previous works on FOREX market prediction did not consider high dimensionality and ignored opinion and semantics of textual language. Therefore, they proposed a model with multi-layer architecture. Description of the layers is as follows: first layer consisted of semantic abstraction layer used for the co-occurrences of the terms. It created the way to identify words with same root to be treated as single entity. Similarly, in the second layer they used Sum-Score to integrate the weights of sentiments. It assigned the weights to the terms appeared in sentiments. It was helpful for reducing the dimensions of terms in view. Finally, the third layer consisted of dynamic model creation. They updated the models with recent information and which was necessary for prediction. Through this approach, they reported an accuracy of 83.33% on real-time data.

Various works carried out on the Stock Market Prediction problems are summarized in terms of Model proposed, Performance measure used, and dataset analyzed in Table 6.

We have drawn few of the observations from this section:

- It is needed to construct stock prediction models that accept the news articles rather than news headlines as input for ontology.
- As more text data (e.g., news articles) is being generated daily, we need more sophisticated techniques for distilling knowledge.
- Many of the works based on the specific news sources (financial news/ corporate news etc.). However, combining multiple news articles can be very much helpful for better stock prediction.

6.1.3. Combined works of stock market and FOREX prediction

A model for stock price prediction based on the news called NewsCATS (News Categorization and Trading System), classified the new press releases based on existing classes associated with stock prices [92]. The SVM algorithm used here for classification examined the stocks of NYSE, NASDAQ, AMEX and five other regional stocks with press releases of Business wire.

Antweiler and Frank [9] presented a study on stock message boards. Through Computational Intelligence (CI) techniques they forecasted the stock returns of the next day using positive messages. Through these message postings, they predicted the volatility of the daily frequencies.

Das and Chen [33] developed a model for extraction of the sentiment from stock message boards. They combined different algorithms through the voting scheme. They reported that the performance of the proposed model was better, keeping in view the lower false positive rate as well as accuracy. Aggregation of the cross-sectional message and time series would improve the quality of the sentiment index. They analyzed investor opinions based on news, regulatory changes and announcements made by the company's management.

Mahajan et al. [87] analyzed impact of news on the stock market. They identified the events by using Latent Dirichlet Allocation (LDA) based on topic extraction method. They analyzed the actual market data to understand the impact on the market. They generated the model by combining the DT and SVM. Through this approach they reported an accuracy of 60%.

A quantitative measure of the language was used to predict the organizations earnings and stock returns by Tetlock et al. [123]. They found out the negative words in the firm-specific news and forecasts of the stock returns. They investigated the S&P 500 companies stock return values based on The Wall Street Journal (WSJ) and Dow Jones News Service (DJNS).

Schumaker and Chen [113] estimated the stock price after releasing the financial news articles over a time of 20 minutes with SVM. The effectiveness of the proposed method was evaluated on 9211 financial news articles and 10,259,042 stock quotes covering the S&P 500 stocks during a five-week period. They reported MSE value as 0.04261 for the actual future stock price, direction of price movement i.e. directional accuracy is 57.1% and 2.06% of the highest return. Further, they found that Proper Noun scheme performs better than others.

Butler and Keselj [22] assessed the performance of stock price based on the textual financial reports. Initially, they used Character n-gram (CNG) based method and readability scores method. SVM was employed for classification. They experimented on S&P500 index lists.

According to the behavioral economics, emotions of the humans are very useful in a decision-making process. Bollen et al. [17] extracted feedbacks from the Twitter about the Dow Jones Industrial average over a period. They analyzed the content of Twitter feeds with tracking tools (Opinion Finder and Google-Profile of Mood States). The Opinion Finder measures the positive vs. negative mood and Google-Profile of Mood measures the mood concerning the following dimensions: calm, alert, sure, vital, kind, and happy. To test the hypothesis of this mood prediction they used the self-organizing fuzzy neural network model. They concluded that the Dow Jones predictions were significantly improved by the inclusion of specific mood dimensions. They reported that 6% reduction in Mean Average Percentage Error (MAPE) and the overall prediction accuracy of 87.6%.

To assist the investors in deciding to buy and sell stocks based on financial news headline Huang et al. [65] proposed a model. Text mining and arbitrary association rules are used to find out the significance of newly arrived news articles. They demonstrated this approach on Taiwan Stock Exchange Financial Price Index.

Li [79] presented the work of analyzing 0-Q and 10-K filings using Text mining. For this experimentation, they considered the filings of SEC Edgar (1994–2007). Later, they extracted the Management Discussion and Analysis section (MD&A) from these filings. After preprocessing, they labeled each sentence into four categories viz., positive, negative, neutral, and uncertain and NB classifier was chosen for classification task.

Groth and Muntermann [55] explained the implications of the news on the stock price. They identified the risk factors which have existed in the text data. They employed different models viz., NB, *k*-NN, NN, and SVM for finding the patterns in the textual data. They reported that *k*-NN is performing well compared to other models.

Schumaker et al. [114] proposed a model called Arizona Financial Text system for prediction of the price movement using financial news. They developed a system based on financial news articles and combined it with sentiment analysis tool. They invoked the Support Vector Regression (SVR) model for this research work. In their study, they found that negative words are more useful for the prediction as compared to the positive emotions. One of the key observations in their research was – good news would impact stock sell, negative articles leads to buying of the stock.

Hagenau et al. [57] by enhancing more expressive features to represent text from the market feedbacks, proved that robust feature selection helps to improve the accuracy as compared to complex features. Selection of semantically relevant features reduces the problem of over-fitting.

The various combined works are summarized in Table 6.

6.2. CRM applications

Business intelligence and analytics deal with systems and technologies, practices, and applications to analyze data and mine the new knowledge of the markets. These new insights can be useful not only for improving the services but also profits. The subject of Customer Relationship Management (CRM) has become cornerstone for financial services industry and it encompasses many business problems such as customer acquisition, market basket analysis and customer churn prediction to mention a few and fortunately, all these problems can be formulated as data mining problems. With the advent of call centers, various communication channels and social media, humungous unstructured data is generated which is a treasure trove of useful business insights. In order to extract this business knowledge, one needs text mining.

In the current scenario, as products are sold online, the customers' review/feedback is also taken online. This leads to generation of vast amount of data, which is useful for companies to both analyze and summarize it and therefore identify the customer needs to serve them better. Besides, social media tools like Facebook, Twitter are used by most of the companies to interact with their clients. Opinion mining is helpful to estimate the sentiment associated with the product and its feature. Sentiment analysis is a classification problem where statements were classified into two categories – positive or negative.

Opinion mining impacts the economic growth of the organization also. Numeric ratings alone are not sufficient to analyze the behavior of the customer. Customer feedback (text) analysis plays an important role to explain the behavior of customers.

Recommendation systems improve the e-commerce sales in many ways like, viewers to buyers, cross-sell, etc. Recommendation systems are used in electronic commerce by various researchers. Schafer et al. [112] presented work on how the e-commerce websites are benefited based on the recommender systems. They examined how it affected on sales improvement. They analyzed six popular sites namely Amazon.com, CDNOW, Drugstore.com, eBay, MovieFinder.com and Reel.com. They created taxonomy for the recommender system (i.e. customers' requirements, techniques for recommendation systems and personalization. They identified the five commonly used recommender systems applications in e-commerce (raw retrieval, statistical summaries, attribute-based, item-to-item correlation, user-to-user correlation).

Pang et al. [100]'s study on sentiment classification using unigram, bigram and n-gram, applied some of the well-known machine learning methods such as NB, Maximum Entropy and SVM on the movie reviews. They concluded that results using these techniques were better compared to human generated values.

Hu and Liu [64] conducted a research study on the text summarization based on customer reviews. Initially, they identified the product features followed by the opinion of the client i.e. positive or negative. Later, they summarized these results.

Customer opinions/reviews are the gold mine for the market competitors. A huge amount of information is available on the web. Popesu and Etzioni [102] proposed a model called OPINE (Tool) for mining the reviews of the customers. They carried out the experiments on Amazon reviews and they reported best values of precision and recall values.

Ghani et al. [50] proposed a model for extraction of attributes and its associated values from the textual description of the products. They extracted data from the retail stores, URLS, prices, etc. from the web. They used wrappers for extraction of information from the sites. As per the domain expert's suggestion, they primarily used eight attributes of each product. NB and Expectation Maximization were used for text classification with almost of 600 products to train the model. They used 5-FCV to evaluate future demand for the products, recommendations about the product and similarities between various providers.

Devitt and Ahmad [35] proposed a model based on lexical cohesion for finding the sentiment polarity in the financial news. They examined the relationship between financial news and the stock market. They investigated on the polarity direction (positive/ negative) and how much the strength of this on the stock value.

Coussement and Van den Poel [31] proposed a new method for complaint handling strategies through email classification that distinguishes complaints from non-complaints by combining the linguistic information with classification model. Linguistic style features were extracted from 9176 emails out of which 3299 were complaint-based emails and rest of them were general. They proposed three different models using Adaboost (ADA) classifier viz., Adaboost-Singular Valued Decomposition (ADA-SVD), Adaboost-Linguistic Style feature (ADA-LS) and Adaboost-Singular Valued Decomposition with Linguistic Style feature (ADA-SVD-LS). Percentage Correctly Classified (PCC) and AUC metrics were used for evaluation of the model performance. They collected and evaluated the call center mails of the Belgian newspaper. They reported that ADA-SVD-LS is performing well compare to other two approaches.

Pang and Lee [99] presented a survey paper that describes the methods of sentiment analysis and its applications. They compared various traditional (fact-based) methods. They also mentioned different types of publicly available datasets and competitions regarding opinion mining analysis.

Bifet and Frank [15] proposed a model for analysis of twitter data streams and extracted the features from twitter data. They used datasets from twittersentiment.appspot.com and Edinburgh corpus. They obtained 10,000 unigrams using WEKA. They used term presence for vector space model creation. They employed various models like Multinomial Naive Bayes, Stochastic Gradient Descent (SGD) and the Hoeffding Tree. They reported that SGD outperformed other models in terms of Kappa statistic value of 62.6% and accuracy of 82.8%. Dey et al. [36] proposed a framework using NLP and Ontology to extract the knowledge from the customer opinions.

The success factors of the companies also depend on the information available on their sites. Based on this, Thorleuchter and Van den Poel [125] developed a new model. They extracted the web content of top 500 companies they employed the Latent Semantic Indexing (LSI) to identify the semantic patterns in the text. They classified the Top 100 (positive class) and the remaining 500 (negative class) e-commerce companies by employing Logistic Regression (LR). The performance of the regression model was evaluated with the measures of lift, ROC, precision and recall.

Liu and Zhang [81] presented a survey of opinion mining and sentiment analysis. They defined the objective of opinion mining in terms of sub tasks (extraction, classification, etc.). They also discussed various types of opinion mining like aspect-based, sentiment classification, dictionary-based approach, etc.

Twitter messages are used to determine the sentiment of users. Most of the companies use sentiments to find out the brand/product sentiment. Ghiassi et al. [51] worked on these type of problems, and proposed a new model based on supervised feature reduction using n-grams. They developed a new model called Dynamic Architecture for Artificial Neural Networks (DAN2) for sentiment classification. They compared the proposed model with SVM classifier and reported that DAN2 performed better than SVM classifier in terms of accuracy and recall. They carried out the experiments on the randomly selected tweets with manual labels.

Many people express their opinions in the social media like Twitter, Facebook, etc. and these vary from one demographic field to other areas. Ikeda et al. [67] proposed a Hybrid model based on Text mining and community analysis for demographic user's analysis. They experimented on the tweets of 100,000 user profiles. They evaluated the proposed model with the following measures: Recall, Precision and F-measure.

He et al. [60] describes the study of the pizza suppliers. They collected data of Pizza Hut, Domino's Pizza and Papa John's pizza from Facebook and Twitter sites. They identified the behavioral patterns occurred in Facebook and Twitter. They carried out the research using SPSS Clementine and Nvivo 9 tools. SPSS Clementine used to extract the key concepts, indexing and grouping the text. They employed NVivo 9 for query search that retrieves pattern and connections. Most of the tweets were about the ordering and delivering, pizza quality, purchase decision and marketing. The recommendations provided from this study helped to improve the business.

Ballings and Van den Poel [12] assessed the feasibility of predicting in Facebook usage frequency with six classification algorithms namely Random Forest, Kernel Factory, Logistic Regression, Neural Networks, Support Vector Machines and Stochastic Adaptive Boosting. They studied the deviation from the regular patterns which would help in customizing the services like advertisements and recommendations. They reported the highest accuracy of 74% and AUC of 0.66 with Stochastic Adaptive Boosting algorithm.

Recently, a survey on the works published during 2002–2014 on sentiment analysis and opinion mining was conducted by Ravi and Ravi [106]. They reviewed the tasks and applications of the opinion mining. They presented data sources and methods for building models.

Various works carried out on the CRM applications are summarized in terms of Model proposed, Performance measure used, and dataset analyzed in Table 7.

The understanding and analysis of CRM applications are still limited because of the Natural Language Processing concerning the text analysis is a typical task. Due to the great revolution in IT sector, many startup companies initiated in the analysis field. So, it is still a live problem in our area.

6.3. Cyber security in finance

Cyber security is of paramount importance in financial service industry, as it is increasingly depending on the information technology for delivering products and services. While on hand, technology provides convenience and comfort to customers, it also opens up doors to numerous forms of cyber crimes. Cybercrime is closely related to the economic impact of a company/ Country. According to the MacAfee [97] report the economic or financial losses of the firms due to cybercrime are varied from $375 to $575 Billion which are greater than some of the countries' annual income. This figure shows that importance of the problem called Cyber Security. In 2015 more than 54 Million people in Turkey, 40 Million

Table 7
Summary of various studies on Text mining in CRM applications.

Study	Dataset	Model	Performance measure
Ballings and Van den Poel [12]	Facebook	LR, RF,NN, SVM, adaptive boosting	AUC, accuracy
He et al. [60]	Tweets/Facebook	SPSS/ Nivea tools	Satisfy/not (accuracy)
Ghiassi et al. [51]	Tweets	N-gram+ANN	Accuracy
Ikeda et al. [67]	Tweets	Hybrid (SVM+clustering)	Recall, precision and F-measure
Thorleuchter and Van den Poel [125]	Top 100 e-commerce Companies	SVD+LR	Lift, ROC, precision and recall
Dey et al. [36]	Amazon reviews	NLP, ontology	Precision, recall, F-score
Bifet and Frank [15]	Tweets	NB/stochastic gradient Descent/hoeffding tree	Accuracy/kappa
Coussement and Van den Poel [31]	e-mails	SVD and LSI	AUC
Devitt and Ahmad [35]	Financial news	Lexical cohension	Precision, recall, F-score
Ghani et al. [50]	URL's	NB/EM	Accuracy
Popescu and Etzioni [102]	Amazon reviews	OPINE tool	Precision, recall
Hu and Liu [64]	Amazon reviews	Feature based summarization (FBS)	Precision, recall
Pang et al. [100]	Movie reviews	NB, maximum entropy, SVM	Accuracy
Schafer et al. [112]	Theoritical study	Theoritical study	Theoritical study

people in the USA, 20 Million people in China and Korea, and 16 million citizens in Germany are affected by cybercrime activities. According to Symantec Internet Security Threat Report [122], malicious activities are more from China followed by the United States and India. They analyzed various malicious activities and their effect on various countries. There is a lot of gap between actual cost and recover cost due to cybercrime. It shows an excellent scope for us to explore the cybercrime activities (prevention, detection, and recovery).

In this paper, we categorized the cyber security applications concerning the financial services industry into five types viz., Phishing detection, Spam detection, Malware detection, Intrusion Detection and Fraud detection. We discussed each of them with respect to their presence in various subsections.

6.3.1. Phishing detection

Phishing is a widespread problem that is affecting both businesses and consumers. Of late, phishing attacks have increased drastically. The goal of phishing is to first steal the identities and credentials of a genuine user and then siphon off funds from his/her account remotely without his/her knowledge. Attackers adopt numerous strategies to attract or take control of users through counterfeit websites. The common factor among all these phishing websites is that they make the users believe that they are actual websites and mislead them in the process. Although phishing can be detected at both email and website levels, analyzing phishing detection is a pressing problem, on which researchers have been working across the world.

Pan and Ding [101] proposed an SVM based approach, which was independent of any specific phishing implementation. They examined the anomalies in web pages, discrepancies between website identity, its structural features, and HTTP transactions, which doesn't require the user knowledge of the site. Reports as per the data collected from the most frequently attacked websites revealed that majority of the victims were from financial institutions and e-commerce companies. The values obtained through this approach achieved low miss rate and low false-positive rate.

Dhamija et al. [38] provided the first experimental proof that explains about how malicious strategies work. They mentioned various strategies for phishing such as the lack of awareness among users, deceitfulness and lack of attention after analyzing 20 websites. Abu-Nimeh et al. [3] compared the performance of CART, LR, Bayesian Additive Regression Trees (BART), SVM, Random Forest (RF), and Neural Networks for phishing emails detection. They experimented on dataset consisting of 2889 emails with 43 features.

Ludl et al. [84] presented a paper on phishing detection. They analyzed 1000 phishing sites using two basic approaches namely blacklist and page analysis. In the first approach, they compared the URL with the existing blacklist URLs to find out whether the particular URL is phishing or not. In the second method, they analyzed HTML source code which contained 18 features. They employed C4.5 for classification purpose. Garera et al. [49] identified some measures to find phishing URLs. In this study, they employed Logistic Regression (LR) leading to higher accuracy and found some key features such as page-based, domain-based, type-based and word-based features.

Miyamoto et al. [93] constructed a model with nine methods viz., SVM, NN, NB, LR, Random Forest, CART, AdaBoost, Bagging, and BART for detection of phishing sites. They analyzed dataset comprising 1500 phishing sites and 1500 legitimate sites. For performance measure, they used F1 measure, error rate and Area Under the ROC Curve (AUC). They reported that Adaboost yielded the highest an F1 value of 0.8581, error rate of 14.15%, and AUC value of 0.9342.

Based on structural properties of an email, Basnet et al. [13] proposed both supervised as well as unsupervised viz., SVM, NN, SOM and K-Means with 16 features. In this work, they used SpamAssassin and Phishing Corpus, which consist of 4000 emails in which 3027 are legitimate and 973 are phishing. They reported that among all, SVM produced the highest accuracy.

Sheng et al. [118] presented a paper on the performance of phishing detection based on blacklist approach. They experimented with eight anti-phishing toolbars viz., Internet Explorer 7 and 8, Firefox 2 and 3, Google Chrome, Netcraft Toolbar, McAfee Site Advisor, and Symantec Norton. Alabama Phishing Team experimented on their legitimate URL's and Phishing URL's.

Bergholz et al. [14] proposed a model for phishing detection based on the categorization of different features like structural (body parts), link, element (HTML, Javascript), spam filter, wordlist (account, update, confirm, verify, secure, notify, log). They experimented with corpus consisting of 16,364 non-phishing and 3636 phishing emails. Aburrous et al. [2] presented e-banking phishing detection model developed using fuzzy data mining approach that combines the data mining and fuzzy techniques. They also generated different models with C4.5, RIPPER and CBA.

He et al. [59] extracted 12 features from the web pages and trained the model with SVM classifier. They reported high accuracy rate with relatively low-false positive and low-false negative rates. Chen et al. [29] analyzed 1030 phishing sites phishing sites with the combination of Text mining and financial attributes. An accuracy of 89% was achieved through their proposed model.

Khonji and Iraqi [73] presented a Literature Survey on Phishing Detection that included four methods viz., visual similarity, machine learning based, blacklists, and rule-based. Detection rate and low false positive values were used for evaluation purpose. They

Table 8
Various studies of phishing detection.

Study	Dataset	Model	Performance measure
Abdelhamid et al. [1]	Millersmiles/ PHISHTANk	Associative rule	Accuracy
Chen et al. [29]	Millersmiles	DT/ NN/ SVM	Accuracy
He et al. [59]	Millersmiles\ PhishTank\ 3Sharp	SVM	True positive/ False positive
Aburrous et al. [2]	PhishTank	C 4.5/ RIPPER	Accuracy/ error rate
Bergholz et al. [14]	AntiPhish search	Markov model	Precision/ recall/ F-score
Sheng et al. [118]	University of Alabama phishing corpus	Anti-phishing toolbars	Accuracy/ false positive
Basnet et al. [13]	SpamAssassin/phishing corpus	SVM/ biased SVM/ NN/ SOM	Accuracy
Miyamoto et al. [93]	PhishTank	SVM/ CART/ LR/ RF/NN/NB/BART/ AdaBoost/ Bagging	Accuracy/ F-score
Abu-Nimeh et al. [3]	Monkey.org/ Spambase	LR/ CART/ SVM/ NN/ BART/ RF	Error rate/ false positive
Garera et al. [49]	Google safe browsing toolbar	LR	Accuracy
Ludl et al. [84]	PhishTank	C 4.5	Accuracy
Pan and Ding [101]	Page pool (paged pool)	SVM	False positive

concluded that machine learning techniques outperformed than other approaches.

Abdelhamid et al. [1] presented a review paper on "Phishing Detection based on Association Rule Mining Technique". They developed an algorithm based on association rule named as Multi-label Classifier based Associative Classification (MCAC) for phishing detection. They described various approaches followed by the researchers for phishing detection namely Blacklist based fuzzy rule-based, machine learning techniques, CANTINA, image-based methods. They presented different sets of features which are discriminative for identifying the phishing or not. The feature subset selection has been using by chi-square method.

Various works carried out on the Phishing Detection are summarized in terms of Model proposed, Performance measure used, and dataset analyzed in Table 8.

Some of the key observations from the above works are:

- Most of the researchers addressed the phishing detection problem as a static problem only.
- PhishTank is the corpus used in many of the works.
- Identification of feature subset is varying from one researcher to other.
- Standard feature set needed to be identified so that problem will be resolved in a better manner.
- More feature selection methods are needed to be explored.

6.3.2. Spam detection

Spam is defined as follows: It is a junk mail sent by unsought person. It consists of Viruses, Trojans etc. Symantec Report (2016) consists of spam activity based on Demographic fields and organizational wise. Email spam is one of the major concerns in the internet world as it has potential in creating financial losses. Spam detection can be carried out mainly in two ways. One is text-based analysis, and another is an image-based method. There exist various feature selection algorithms in literature. Some of them used in spam filtering analysis methods are Document Frequency, Information Gain, Chi-square Statistic, Odds Ratio and Term Frequency.

Androutsopoulos et al. [7] proposed Naive Bayes based anti-spam model. They experimented with Ling-Spam corpus. Earlier to this, Androutsopoulos et al. [8] did a similar type of work in the same year, which proposed a model to detect the spam messages. They tested on the PU1 corpus which consisted of 1099 messages (481 spam and 618 legitimate) by NB model.

Massey et al. [89] reviewed machine learning methods for spam filtering. They applied various feature selection methods and evaluated the model with different classifiers (ID3, NN, etc.) on four corpuses namely Ling-Spam, Spam Assassin, Annexia Spam Archive and their own email collection. Zhang et al. [138] applied SVM, AdaBoost, Maximum Entropy Model, NB, and Memory-Based Learning with various feature subsets for spam filtering. They employed Information Gain, Document Frequency, and Chi-square for Feature Subset selection. They analyzed datasets viz., PU1 (481 spam and 618 legitimate), Ling-Spam (481 spam and 2412 legitimate), SpamAssassin (1897 spam and 4150 legitimate) and ZH1 Chinese Spam Corpus (1205 spam and 428 legitimate).

Klimt and Yang [74] performed spam detection with a new dataset on Enron corpus with SVM classifier. They reported F-Score of 0.7. Metsis et al. [91] worked on the spam filtering with the Naive Bayes (NB) approach. Sending spam emails into the hijacked systems for a short period are called Transient spam-bots. Brodsky and Brodsky [21] presented a framework that solves these types of problems.

Chen et al. [27] analyzed spam filtering using three variants of Bayesian methods namely Aggregating One-Dependence Estimators (AODE), Hidden Naive Bayes (HNB), and locally weighted learning with Naive Bayes (LWNB). They selected relevant features with the following feature selection methods: Gain Ratio, Information Gain, Symmetrical Uncertainty and Relief. They compared the results of three classifiers with the linear classifiers NB, *k*-NN, SVM, and C4.5. They reported that Aggregating One-Dependence Estimators (AODE) performed the best with respect to accuracy.

Similarly, Blanzieri and Bryl [16] conducted a survey on email filtering techniques. They provided an overview of the data mining and machine learning techniques applied to spam filtering. Besides analyzing various methods, they also discussed commercial and non-commercial software solutions.

Guzella and Caminhas [56] presented a comprehensive study of spam filtering with data mining techniques. They covered both textual and image-based approaches, the structure of spam filter and representation (i.e. document frequency, information gain, term frequency variance). They listed out the publicly available datasets and various classifiers including SVM, NB, LR, ANN and hybrid methods employed by researchers.

Anomaly detection is one of the important ways to identify the abnormal patterns/behaviors to quickly determine the suspicious transactions which deviate from the normal behavior. Finding deviations from the regular patterns is called Anomaly Detection [34]. There are various existing methods to characterize these anomalies. In this study, we presented some of the cases based on system call traces. There were few researchers who explained about the intrusion detection through machine learning approaches.

Zhan et al. [137] applied anomaly detection in the email system to find out whether an email is normal or spam. They proposed the weak estimators such as Stochastic Learning-Based Weak Estimator (SLWE) and Maximum Likelihood Estimator (MLE) for estimating the distributions of events which are deviate from the normal pattern. In this study, they used Information Gain to find out the top 200 discriminative features and performed classification with NB classifier. They compared the Precision, Recall and ROC values obtained by the above two approaches.

Table 9
Various studies on the spam detection.

Study	Dataset	Model	Performance measure
Zhan et al. [137]	SpamAssasin	NB	Precision, recall, ROC
Chen et al. [27]	PU corpus	AODE/ HNB/ LWNB/ NB/ SVM/ C 4.5/ *k*-NN	Accuracy
Brodsky and Brodsky [21]	SpamAssassin	Trinity	Precision
Metsis et al. [91]	SpamAssassin	NB	Recall, ROC
Zhang et al. [138]	Chinese corpus	NB/ SVM/ C 4.5	Accuracy
Klimt and Yang [74]	CMU	SVM	F-Score
Massey et al. [89]	SpamAssassin/Ling-spam/ Annexia spam archive	NB/ ID3/ NN	Error rate
Androutsopoulos et al. [7]	Ling-spam	NB	Accuracy
Androutsopoulos et al. [8]	PU corpus	NB	Accuracy

Identification of Spam is an expensive and time-consuming process. Caruna and Li [23] presented a survey on this, which focused on emerging approaches to spam filtering built on recent developments in computing technologies. These include peer-to-peer computing, grid computing, semantic web, and social networks. It also addressed many perspectives related to personalization and privacy in spam filtering. They concluded that despite advanced methodologies being used, attaining high performance and detection rate is still an open problem.

Various works carried out on the Spam Detection are summarized in terms of Model proposed, Performance measure used, and dataset analyzed in Table 9.

We had drawn some of the observations from the existing works.

- Most of the works are based on the SpamAssasin corpus.
- Very few works exist for Spam identification based on dynamic behavior.
- Need to generate/ create vast corpus such that we will analyze the expected behavior of Spam.
- Machine Learning algorithms are needed to be explored in depth.

6.3.3. Malware detection

Another challenging problem for cyber world is malware detection. It is both important as well as relevant, which is evident from a Gartner's Magic Quadrant report [86]. As per the survey, companies spent approximately $2.8 billion in 2011 for malware protection. For malware protection, organizations must have anti-malware, anti-spyware, vulnerability assessment, personal firewalls and finally host-based intrusion prevention. Despite the high-level security mechanisms, system vulnerabilities became a powerful weapon for malware authors. There are two types of malware detection techniques: anomaly-based detection and signature-based method. In this work, we analyzed the malware based on the anomaly-based approach.

Numerous machine learning techniques have been applied from the past few years to analyze malware. Wang and Stolofo [130] developed a model called Payload-based Anomaly Detector (PAYL) and reported an overall 60% of detection rate with 1% of false positive rate. Vasudevan and Yerraballi [128] defined variants of Malware as Viruses, Trojans, and Spywares. Ye et al. [135] developed the Intelligent Malware Detection System (IMDS) using Association Rule Mining, and further employed NB, SVM and J4.8 classifiers which produced accuracies of 83.86%, 90.54% and 91.49% respectively, while their proposed (IMDS) approach yielded 93.07%. Provos et al. [103] also performed research on web malwares. Idika and Mathur [66] carried out research on different malware detection techniques and reported their limitations. In their research work, they compared 45 techniques about malware detection.

Ahmed et al. [6] proposed an amalgam malware detection scheme using run-time Application Program Interface (API) calls of Windows OS. It consists of offline training and online testing phases. Out of thousands of API calls in Windows, 237 calls were chosen for both Benign and Malware programs. Further, these API calls are grouped into seven based on their functionality. They categorized different API calls, resulting in 0.97 detection rate of accuracy.

Malwares are grouped based on their common characteristics by Ye et al. [134]. They proposed Automatic Malware Categorization System (AMCS) by combining k-medoids algorithm, Hierarchical clustering and weighted subspace k-medoids algorithm. They evaluated the efficacy of the proposed model on the malware dataset of Kingsoft Anti-Virus Lab. The performance of the model was measured by using F-measure. Hou et al. [63] analyzed the characteristics of a web page whether it is malicious or not. They collected 1141 URLs from StopBadWare site. Information Gain was used for selection of distinctive features followed by model construction with NB, DT, SVM and Boosted Decision Tree classifiers. They performed 10FCV and reported that Boosted Decision Tree outperformed other models.

Zhuang et al. [139] developed a model for malware and phishing websites categorization using cluster ensemble method involving hierarchical clustering and k-medoid algorithm. In this work, they represented the malware with static feature extraction methods. They collected the malware dataset from Kingsoft Internet Security Laboratory between June 10 and 16, 2012. Feature vector generated with tf and tf-idf weighting.

Suarez-Tangil et al. [120] proposed a model to classify the smart phone malware's using Text mining approach. Recently, Sundarkumar et al. [121] presented a method based on topic modeling and machine learning to detect the malware. They used API calls for detection of malware based on features selected by topic modeling. They employed different algorithms viz., DT, SVM, PNN, GMDH, MLP and RF for classification purpose. They experimented with two datasets, and concluded their work with the statistical significance testing.

Various works carried out on the Malware Detection are summarized in terms of Model proposed, Performance measure used, and dataset analyzed in Table 10.

The following are some of the highlights from the above study:

- Malware detection is not explored fully through Text mining process.
- It is needed to see the sights of API calls for identifying Malware.
- API calls/ System calls need to be dynamically analyzed.
- More advanced intelligent techniques need to be explored.

Most of the works are based on Windows API calls. We also need to extend other Operating Systems call so that there will be an opportunity to explore common properties between Malwares on different Operating Systems.

6.3.4. Intrusion detection

Cyber fraudsters adopt novel and potent methods to hack computer systems in order to cause maximum damage in just one go.

Table 10
Malware detection with Text mining approach.

Study	Dataset	Model	Performance measure
Sundarkumar et al. [121]	Windows API calls	SVM/ DT/ GMDH/ PNN./ MLP/ RF	AUC
Suarez-Tangil et al. [120]	Android OS malware	Hierarchical clustering, NN	Error value
Zhuang et al. [139]	Kingsoft internet security laboratory	k-medoids, hierachical clustering	F_1 measure
Hou et al. [63]	Web pages	DT	Accuracy, ROC Curve
Ye et al. [134]	Kingsoft anti-virus lab	k-medoids, hierachical clustering and weighted subspace k-medoids	F_1 measure
Ahmed et al. [6]	API calls	DT	Accuracy
Vasudevan and Yerraballi [128]	API calls	SpiKE	TP
Provos et al. [103]	We pages	Map reduce framework	TP
Ye et al. [135]	API calls	ARM, NB, SVM and J4.8	Accuracy
Wang and Stolofo [130]	DARPA 99	PAYL	TP

Table 11
Intrusion detection with Text mining approach.

Study	Dataset	Model	Performance Measure
Adeva and Atxa [4]	Log entries	Rocchio/ *k*-NN/ NB	F_1 score
Sharma et al. [117]	System calls sequences of DARPA 98	*k*-NN	False positive
Rawat et al. [107]	System calls of DARPA 98	*k*-NN	False positive
Chen et al. [28]	System calls of DARPA 98	ANN/ SVM	Detection rate/ False positive
Liao and Vemuri [80]	System calls	*k*-NN	False positive
Liu et al. [82]	System calls	ANN	Accuracy
Helmer et al. [61]	System calls	RIPPER	True positive

Pharming or DDos attack are some of the ways of doing it. If successful, it can result in maximum damage.

Liao and Vemuri [80] employed Text mining techniques to predict whether a program behavior is normal or not and evaluated the model with *k*-NN classifier on DARPA BSM dataset. They reported low false positive value.

Helmer et al. [61] described the network based multi-agent distributed system for intrusion detection. For a process, they maintained the feature vector representation, and based on execution of a sequence (at attack) it was labeled as good or bad. They performed the feature selection through Genetic Algorithm (GA). They compared the results with and without feature selection. They employed RIPPER for classification and found if-then rules for intrusion detection. Finally, they concluded that the combination of GA (for feature selection) and RIPPER (for classification) yielded the best results.

Liu et al. [82] presented a model for intrusion detection based on neural networks. They employed three types of neural networks viz., Back propagation, RBF networks, and Self Organizing Map. They experimented with the Unix System calls of lpr (2703 normal and 1001 abnormal traces) and send mail (172 normal and three abnormal traces) programs, which were extracted from University of Mexico.

Chen et al. [28] performed intrusion detection on the DARPA 98 dataset with ANN and SVM classifiers. They generated the tf-idf matrix with the system calls and classified using the above classifiers. . They listed out the most commonly used 50 system calls and reported the FP and the detection rate values. By using server logs of a website, intrusion detection was performed by

Rawat et al. [107] proposed a model for intrusion detection. They introduced the binary weighted cosine metric for similarity measure, and classified the system calls into normal and abnormal. They evaluated the model with *k*-NN classifier on DARPA 98 project. They reported low false positive rate with the proposed approach. Sharma et al. [117] described the intrusion detection using Text mining techniques. They employed *k*-NN classifier for intrusion detection and evaluated the proposed approach on DARPA 98 data set. Adeva and Ataxa [4] carried out research work on intrusion detection in web applications by using the log entries generated by the web server. The proposed model evaluated with Rocchio, *k*-NN and Bayes classifiers with F-measure as metric. They reported the highest F-value with NB classifier.

Various works carried out on the Intrusion Detection are summarized in terms of Model proposed, Performance measure used, and dataset analyzed in Table 11.

- Like Malware analysis, Intrusion detection is also not yet explored in-depth through Text mining.
- Most of the works are analyzed using *k*-NN classifier.
- Few more methods are needed to be applied for better detection.
- The majority of the works are based on System calls only. We need to investigate other ways also.

6.3.5. Fraud detection and other problems

Churn Prediction or Bankruptcy predictions have been most interesting area of research in financial domain [119]. Most of the extant works are based on numerical data such as financial ratios etc. Recently, some of the works are carried out based on quantitative as well as qualitative also. Few of the works are presented in this section.

Appavu et al. [10] proposed a method based on DT called AD Infinitum for identifying the threatening emails. They classified the emails based on the incremental method, and created two corpuses namely TCEThreatening1 and TCEThreatening2. They compared the results with regular classifiers namely DT, SVM, NB. They reported that the performance of proposed method yielded best results in terms of F-measure and Accuracy.

There are various methods including finding out the fraud in financial statements. Cecchini et al. [24] proposed a model to analyze financial events with text-based ontology creation. They created a dictionary called Management Discussion and Analysis section (MD & A) based on companies bankruptcy. They carried out the research on Bankruptcy dataset (78 companies) and fraud dataset (61 fraud and 61 non-fraud companies). Performance of the model was evaluated with accuracy values. They reported the prediction accuracy values as follows: 83.87% for Bankruptcy and 81.97% for Fraud datasets.

Shirata et al. [119] proposed a model to predict bankruptcy based on text existed in financial statements in Japan. They experimented on annual reports of 180 companies for the period of 1999–2005 (90 Bankruptcy and 90 non-Bankruptcy) which are col-

Table 12
Fraud detection with Text mining approach.

Study	Dataset	Model	Performance measure
Saha et al. [109]	Financial statement	LR	Accuracy, sensitivity, specificity
Glancy and Yadav [53]	Financial reports	Hierarchical clustering	Accuracy
Shirata et al. [119]	Financial reports	CART	Accuracy
Cecchini et al. [24]	US securities and Exchane commision releases	Ontology	Accuracy
Appavu et al. [10]	TCEThreating emails, PU1	DT/ SVM/ NB	Accuracy, F-measure, recall, precision

lected from the Tokyo stock exchange. They identified some of the specific words which are helpful for distinguishing non-Bankruptcy and Bankruptcy. They employed CART-based SAF (Simple Analysis of Failure) model for prediction in the OmniFind Analytic Edition (OAE) tool.

Sometimes the data is present in an imbalance mode, and it is very tough for prediction tasks. Glancy and Yadav [53] applied quantitative approaches to text corpus in order to find out the frauds in financial statements. They proposed computational fraud detection model that can detect fraud in financial reports.

Manual auditing is highly inefficient and is not so much accurate because of human miscalculation. To avoid this type of problems, Saha et al. [109] proposed a model for processing Bank loans. Through Text mining approaches, they analyzed the 100 fraudulent cases of small-scale and medium-scale industries. They analyzed these cases with the help of five reputed auditors to reveal risk level, risk impact, and risk detection. A binary value was assigned based on the domain experts' evaluation, and these values are being served as output labels for the regression model. They employed the Logistic Regression to classify the fraudulent applications. Various works carried out on the Fraud Detection are summarized in terms of Model proposed, Performance measure used, and dataset analyzed in Table 12.

Some of the key observations from this section:

- Full evaluation audit score model is required for comparative analysis of more companies.
- Limited for single geographical location only need to explore in multiple geographic locations for more generalization and better predictions.
- There is a need to construct ontologies which are useful in multiple domains.
- Handling of missing and ambiguous data is the challenging problem for fraudulent cases.
- There is a need to build the fault tolerance systems for banking applications

7. Conclusions and future directions

This paper presents a comprehensive review of Text mining applications in the financial domain. The applications include four major categories viz., FOREX prediction, stock market prediction, CRM and Cyber fraud detection. Although there exist several text mining applications, devising more efficient techniques is essential for handling and predicting a significant amount of data. Combining numerical data with textual data yields better predictions. In our study, we found that

- Considerable progress is achieved in the application of text mining to solve problems in financial domain. However, there remains a lot to be achieved further. Identifying suitable feature selection method is still an open problem. Since datasets in this domain have high dimensionality, dimension reduction becomes critical to the success of the data mining techniques.
- Stock market prediction and FOREX rate problems are frequently solved using text articles. Intrusion detection and fraud detection are rarely addressed with Text mining approaches.
- The most frequently used Text mining task is Classification followed by Forecasting.
- SVM, NB, *k*-NN, DT are most often used Data mining techniques in financial applications. Out of them, SVM is the predominant technique applied in various applications because of its high prediction capability.
- Performance evaluation metrics are not unique and vary from application to application.

The study concludes with some of the key future directions:

- Most of the works are related to specific companies about their stock values. We need sophisticated techniques such as NLP etc. for investor behavior analysis.
- Sentence subjectivity classification and semantic structure identification are the most challenging problems.
- Construction of Ontologies is required in each domain (finance, insurance, etc.)
- Churn Prediction problem is not yet fully explored using Text mining.
- Hybrid techniques models are explored but not in many studies and applications. Ensemble techniques have to develop to obtain highly accurate predictions.
- Evolutionary methods are conspicuous by their absence. They haven't been explored at all perhaps due to large number of features in text corpus. However, they are worth exploring given their comparable performance vis-a-vis traditional data mining techniques.
- Most of the works related to stock market prediction are with respect to limited zones only.
- Fuzzy logic based techniques such as fuzzy rule based classification, fuzzy clustering etc. need to be explored. They are conspicuous by their absence.
- Ontologies prove to be helpful in sentiment classification and therefore they need to be employed in future studies in financial domain.
- Social media analytics is also playing an important role in the financial sector. Hence, the paper includes this area in the scope.
- Deep learning is potentially useful in dealing with large feature space dimensions in textual corpus. Hence, it needs to be effectively integrated into the prediction /classification phase.
- Developments in big data analytics can be exploited successfully in Text mining applications to finance too.
- Spiking neural networks significantly improve prediction accuracies in temporal data. They too are conspicuously absent in the previous studies.
- Evolutionary computation holds a lot of promise in solving these problems, as it is very versatile and powerful in that it can solve classification, regression, clustering and association rule mining problems in finance. It can rival many traditional data mining techniques in these tasks in terms of performance.
- News articles should be used instead of headlines to avoid ambiguity and correctness.
- Benchmark datasets are not available in financial market analysis. Many of the researchers are experimenting with their available data. This leads to inappropriate prediction rendering them

useless for comparing the results. A Common Framework is required to integrate the both input and output of the models.
- Standardized performance metrics are preferable for comparison among the various works.

References

[1] N. Abdelhamid, A. Ayesh, F. Thabtah, Phishing detection based associative classification data mining, Expert Syst. Appl. 41 (13) (2014) 5948–5959 Elsevier.
[2] M. Aburrous, M.A. Hossain, K. Dahal, F. Thabtah, Intelligent phishing detection system for e-banking using fuzzy data mining, Expert Syst. Appl. 37 (2010) 7913–7921 Elsevier.
[3] S. Abu-Nimeh, D. Nappa, X. Wang, S. Nair, A Comparison of machine learning techniques for phishing detection, APWG eCrime Researchers Summit, October 4-5, ACM, 2007.
[4] J.J.G. Adeva, J.M.P. Atxa, Intrusion detection in web applications using text mining, Eng. Appl. Artificial Intell. 20 (4) (2007) 555–566 Elsevier.
[5] C.C. Aggarwal, C.X. Zhai, A survey of text clustering algorithms, in: Mining Text Data, Springer, 2012, pp. 77–128.
[6] F. Ahmed, H. Hameed, Z. Shafiq, M. Farooq, Using Spatio temporal Information in API calls with machine learning algorithms for malware detection, 2nd ACM worshop on Security and Artificial Intelligence (AISec), November 9th, 2009.
[7] I. Androutsopoulos, J. Koutsias, K.V. Chandrinos, G. Paliouras, C.D. Spyropolus, An evaluation of Naive Bayesian anti-spam filtering, in: ECML, Barcelona, Spain, Springer, 2000a, pp. 9–17.
[8] I. Androutsopoulos, J. Koutsias, K.V. Chandrinos, C.D. Spyropoulos, An experimental comparison of Naive Bayesian and keyword-based anti-spam filtering with personal e-mail messages, in: SIGIR ' 00, Athens, Greece, ACM, 2000, pp. 160–167.
[9] W. Antweiler, M.Z. Frank, Is all that talk just noise? The information content of internet stock message boards, J. Finance 59 (3) (2004) 1259–1294.
[10] S. Appavu, R. Rajaram, M. Muthupandian, G. Athiappan, K.S. Kashmeera, Data mining based intelligent analysis of threatening e-mail, Knowl.-Based Syst. 22 (2009) 392–393 Elsevier.
[11] B. Back, J. Toivonen, H. Vanharanta, A. Visa, Comparing numerical data and text information from annual reports using self-organizing maps, Int. J. Account. Inform. Syst. 2 (4) (2001) 249–269 Elsevier.
[12] M. Ballings, D. Van den Poel, CRM in social media: predicting increases in Facebook usage frequency, Eur. J. Operat. Res. 244 (1) (2015) 248–260 Elsevier.
[13] R. Basnet, S. Mukkamala, A.H. Sung, Detection of phishing attacks: a machine learning approach, Soft Comput. Appl. Ind., STUDFUZZ 226 (2008) 373–383 2008, Springer.
[14] A. Bergholz, J.D. Beer, S. Glahn, M.-F. Moens, G. Paal, S. Strobel, New filtering approaches for phishing email, J. Comput. Secur. 18 (1) (2010) 7–35 IOS Press.
[15] A. Bifet, E. Frank, Sentiment knowledge discovery in twitter streaming data, in: 13th International Conference Discovery Science (DS), October 6-8, Canberra, Australia, Springer, 2010, pp. 1–15.
[16] E. Blanzieri, A. Bryl, A survey of learning-based techniques of email spam filtering, J. Artificial Intell. Rev. 29 (1) (2008) 63–92 Kluwer Academic Publishers.
[17] J. Bollen, H. Mao, X.-J. Zeng, Twitter mood predicts the stock market, J. Comput. Sci. 2 (2010) 1–8 Elsevier.
[18] H. Borko, M. Bernick, Automatic document classification, J. ACM (JACM) 10 (2) (1963) 151–162 ACM.
[19] L. Breiman, J.H. Friedman, C.J. Stone, R.A. Olshen, Classification and Regression Trees, CRC Press, Taylor & Franics, 1984.
[20] E. Brill, A simple rule-based part of speech tagger, in: Third Conference on Applied Natural Language Processing (ANLC), Trento, Italy, ACM, 1992, pp. 152–155.
[21] A. Brodsky, D. Brodsky, Trinity: distributed defense against transient spam-bots, in: 26th ACM Symposium on Principles of Distributed Computing (PODC), New York, 2007, pp. 378–379.
[22] M. Butler, V. Keselj, Financial forecasting using character n-gram analysis and readability scores of annual reports, in: 22nd Canadian Conference on Artificial Intelligence (Canadian AI), Kelowna, Canada, Springer, 2009, pp. 39–51.
[23] G. Caruna, M. Li, A survey of emerging approaches to spam filtering, ACM Comput. Surv. 44 (2) (2012) ACM.
[24] M. Cecchini, H. Aytug, G.J. Koehler, P. Pathak, Making words work: using financial text as a predictor of financial events, Dec. Support Syst. 50 (1) (2010) 164–175 Elsevier.
[25] S.W.K. Chan, J. Franklin, A text-based decision support system for financial sequence prediction, Dec. Support Syst. 52 (1) (2011) 189–198 Elsevier.
[26] A. Chatrath, H. Miao, S. Ramchander, S. Villupuram, Currency jumps, cojumps and the role of macro news, J. Int. Money Finance 40 (2014) 42–62 Elsevier.
[27] C. Chen, Y. Tian, C. Zhang, Spam filtering with several novel Bayesian classifiers, in: 19th Conference on Pattern Recognition (ICPR), Tampa, FL, IEEE, 2008, pp. 1–4.
[28] W.-H. Chen, S.-H. Hsu, H.-P. Shen, Application of SVM and ANN for intrusion detection, Comput. Operat. Res. 32 (10) (2005) 2617–2634 Elsevier.
[29] X. Chen, I. Bose, A.C.M. Leung, C. Guo, Assessing the severity of phishing attacks: a hybrid data mining approach, Dec. Support Syst. 50 (4) (2011) 662–672 Elsevier.
[30] C. Cortes, V. Vapnik, Support-vector networks, Mach. Learn. 20 (3) (1995) 273–297.
[31] K. Coussement, D. Van den Poel, Improving customer complaint management by automatic email classification using linguistic style features as predictors, Dec. Support Syst. 44 (4) (2008) 870–882 Elsevier.
[32] T.M. Cover, P.E. Hart, Nearest neighbor pattern classification, IEEE Trans. Inform. Theor. 13 (1) (1967) 21–27.
[33] S.R. Das, M.Y. Chen, Yahoo! for amazon: sentiment extraction from small talk on the web, J. Manage. Sci. 53 (9) (2007) 1375–1388 INFORMS.
[34] D.E. Denning, An intrusion-detection model, IEEE Trans. Softw. Eng. 13 (2) (1987) 222–232.
[35] A. Devitt, K. Ahmad, Sentiment polarity identification in financial news: a cohesion-based approach, in: 45th Association of Computational Linguistics (ACL), Prague, Czech Republic, 2007, pp. 984–991.
[36] L. Dey, Sk.M. Haque, Raj Nidhi, Mining customer feedbacks for actionable intelligence, in: WI-IAT, Toronto, Canada, IEEE, 2010, pp. 239–242.
[37] L. Dey, A. Mahajan, Sk.M. Haque, Document clustering for event identification and trend analysis in market news, in: ICAPR 09, Kolkata, IEEE, 2009, pp. 103–106.
[38] R. Dhamija, J.D. Tygar, M. Hearst, Why phishing works, in: CHI , April 22–27, Montréal, Québec, Canada, ACM, 2006, pp. 581–591.
[39] P. Domingos, M. Pazzani, On the optimality of the simple Bayesian classifier under zero-one loss, Mach. Learn. 29 (2) (1997) 103–130 Kluwer Academic Publishers.
[40] J. Dorre, P. Gerstl, R. Seiffert, Text mining: Finding nuggets in mountains of textual data, in: KDD-99, San Diego, CA, ACM, 1999, pp. 398–401.
[41] G. Escudero, L. Marquez, G. Rigau, Boosting applied to word sense disambiguation, in: ECML 00, Barcelona, Spain, Springer, 2000, pp. 148–155.
[42] M.D.D. Evans, R.K. Lyons, How is macro news transmitted to exchange rates, J. Financ. Econ. 88 (1) (2008) 26–50 Elsevier.
[43] M. Fasanghari, G.A. Montazer, Design and implementation of fuzzy expert system for Tehran stock exchange portfolio recommendation, Expert Syst. Appl. 37 (9) (2010) 6138–6147 Elsevier.
[44] R. Feldman, H. Hirsh, Mining associations in text in the presence of background knowledge, in: KDD, Aug. 2-4, Portland, Oregon, USA, AAAI, 1996, pp. 343–346.
[45] E. Fix, J.L. Hodges Jr., Discriminatory Analysis, Nonparametric Discrimination, USAF School of Aviation Medicine, Randolph, Proj, 1951.
[46] G.P.C. Fung, J.X. Yu, W. Lam, News sensitive stock trend prediction, PAKDD, LNAI 2236 (2002) 481–493 Springer.
[47] G.P.C. Fung, J.X. Yu, W. Lam, Stock prediction: integrating text mining approach using real-time news, in: International Conference on Computational Intelligence for Financial Engineering, Hong Kong, IEEE, 2003, pp. 395–402.
[48] F. Galton, Regression towards mediocrity in hereditary stature, J. Anthropol. Inst. Great Britain Ireland 15 (1886) 246–263.
[49] S. Garera, N. Provos, M. Chew, A.D. Rubin, A framework for detection and measurement of phishing attacks, WORM'07, November 2, ACM, 2007.
[50] R. Ghani, K. Probst, Y. Liu, M. Krema, A. Fano, Text mining for product attribute extraction, ACM SIGKDD 8 (1) (2006) 41–48.
[51] M. Ghiassi, J. Skinner, D. Zimbra, Twitter brand sentiment analysis: a hybrid system using n-gram analysis and dynamic artificial neural network, Expert Syst. Appl. 40 (16) (2013) 6266–6282 Elsevier.
[52] E. Gilbert, K. Karahalios, Widespread worry and stock market, 4th International AAAI Confernce on Weblogs and Social Media (ICWSM), 2010 2010.
[53] F.H. Glancy, S.B. Yadav, A computational model for financial reporting fraud detection, Dec. Support Syst. 50 (3) (2011) 595–601 Elsevier.
[54] C. Goodhart, News and the foreign exchange market, J. Int. Securities Markets 4 (1989) 333–348.
[55] S.S. Groth, J. Muntermann, An intraday market risk management approach based on textual analysis, Dec. Support Syst. 50 (4) (2011) 680–691 Elsevier.
[56] T.S. Guzella, W.M. Caminhas, A review of machine learning approaches to spam filtering, Expert Syst. Appl. 36 (7) (2009) 10206–10222 Elsevier.
[57] M. Hagenau, M. Liebmann, D. Neumann, Automated news reading: Stock price prediction based on financial news using context-capturing features, Dec. Support Syst. 55 (3) (2013) 685–697 Elsevier.
[58] Z. Harris, Distributional structure, Word 10 (2/3) (1954) 146–162 Taylor & Francis.
[59] M. He, S.-J. Horng, P. Fan, M.K. Khan, R.S. Run, J.-L. Lai, R.-J. Chen, A. Sutanto, An efficient phishing webpage detector, Expert Syst. Appl. 38 (10) (2011) 12018–12027 Elsevier.
[60] W. He, S. Zha, L. Li, Social media competitive analysis and text mining: a case study in the pizza industry, Int. J. Inform. Manage. 33 (3) (2013) 464–472 Elsevier.
[61] G. Helmer, J.S.K. Wong, V. Honavar, L. Miller, Automated discovery of concise predictive rules for intrusion detection, J. Syst. Softw. 60 (3) (2002) 165–175 Elsevier.
[62] A. Hotho, S. Staab, G. Stumme, Ontologies improve text document clustering, in: ICDM, Nov. 19-22, Florida, USA, IEEE, 2003, pp. 541–544.
[63] Y.-T. Hou, Y. Chang, T. Chen, C.-S. Laih, C.-M. Chen, Malcious web content detection by machine learning, Expert Syst. Appl.. 37 (1) (2010) 55–60 Elsevier.
[64] M. Hu, B. Liu, Mining and summarizing customer reviews, in: KDD 04, Seattle, Washington, USA, ACM, 2004, pp. 168–177.
[65] C.-J. Huang, J.-J. Liao, D.-X. Yang, T.-Y. Chang, Y.-C. Luo, Realization of a news dissemination agent based on weighted association rules and text mining techniques, Expert Syst. Appl. 37 (9) (2010) 6409–6413 Elsevier.

[66] Idika, N., and Mathur, A.P., 2007. A Survey of Malware Detection Techniques. Department of Computer Science, Purdue University, Tech. Rep.
[67] K. Ikeda, G. Hattori, C. Ono, H. Asoh, T. Higashino, Twitter user profiling based on text and community mining for market analysis, Knowl.-Based Syst. 51 (2013) 35–47 Elsevier.
[68] A. Ittoo, L.M. Nguyen, A. Van den Bosch, Text analytics in industry: challenges, desiderata and trends, Comput. Ind. 78 (2016) 96–107 Elsevier.
[69] A.G. Ivakhnenk, The group method of data handling – a rival of the method of stochastic approximation, Soviet Automatic Control 13 (3) (1968) 43–55.
[70] A.K. Jain, M.N. Murty, P.J. Flynn, Data clustering: a review, ACM Comput. Surv. 31 (3) (1999) 264–323 ACM.
[71] F. Jin, N. Self, P. Saraf, P. Butler, W. Wang, N. Ramakrishnan, Forex-foreteller: currency trend modeling using news articles, in: KDD 13, Chicago, Illinois, USA, 2013, pp. 1470–1473.
[72] T. Joachims, probabilistic analysis of the Rocchio algorithm with tfidf for text categorization, in: 14th ICML, Nashville, Tennessee, USA, Morgan Kaufmann Publishers Inc, 1997, pp. 143–151.
[73] M. Khonji, Y. Iraqi, Phishing detection: a literature survey, IEEE Commun. Surv. Tut. 15 (4) (2013) IEEE.
[74] B. Klimt, Y. Yang, The enron corpus: a new dataset for email classification research, in: Machine Learning: ECML, LNCS, Volume 3201, Springer, 2004, pp. 217–226.
[75] A. Kloptchenko, T. Eklund, B. Back, J. Karlsson, H. Vanharanta, A. Visa, Combining data and text mining techniques for analyzing financial reports, J. Intell. Syst. Account. Finance Manage. 12 (1) (2004) 29–41 John Wiley and Sons Ltd.
[76] M. Koppel, I. Shtrimberg, Good news or bad news? Let the market decide, in: AAAI Symposium on Exploring Attitude and Affect in Text, Palo Alto, 2006, pp. 86–88.
[77] R. Kosala, H. Blockeel, Web mining research: a survey, in: KDD, 2, ACM, 2000, pp. 1–15.
[78] V. Lavrenko, M. Schmill, D. Lawrie, P. Ogilvie, D. Jensen, J. Allan, Mining concurrent text and time series, KDD, 2000 ACM.
[79] F. Li, The Information content of forward-looking statements in corporate filings - a Naive Bayesian machine learning approach, J. Account. Res. 48 (5) (2010) 1049–1102 Wiley.
[80] Y. Liao, V.R. Vemuri, Using text categorization techniques for intrusion detection, in: 11th USENIX Security Symposium, 2002, pp. 51–59.
[81] B. Liu, L. Zhang, A survey of opinion mining and sentiment analysis, Mining Text Data (2012) 415–463 Springer.
[82] Z. Liu, G. Florez, S.M. Bridges, A Comparison of input representations in neural networks: a case study in Intrusion detection, in: International Joint Conference on Neural Networks (IJCNN), Honolulu, Hawaii, IEEE, 2002, pp. 1708–1713.
[83] E.-P. Lim, H. Chen, G. Chen, Business intelligence and analytics: research directions, ACM Trans. Manag. Inform. Syst. 3 (4) (2013) Article 17, ACM.
[84] C. Ludl, S. McAllister, C. Kruegel, On the effectiveness of techniques to detecting phishing sites, in: Detection of Intrusions and Malware, and Vulnerability Assessment (DIMVA), 4579, Lucerne, Switzerland, Springer, 2007, pp. 20–39.
[85] J. Lu, V. Behbood, P. Hao, H. Zuo, S. Xue, G. Zhang, Transfer learning using computational intelligence: a survey, Knowl.-Based Syst. 30 (2015) 14–23 Elsevier.
[86] Magic Quadrant for end point protection platforms. Visited Dec 2014. http://www.computerlinks.co.uk/FMS/22855.magic_quadrant_for_endpoint_protection_platforms.pdf.
[87] A. Mahajan, L. Dey, Sk.M. Haque, Mining financial news for major events and their impacts on the market, in: WI-IAT '08, Sydney, NSW, IEEE, 2008, pp. 423–426.
[88] M. Maron, Automatic indexing: an experimental inquiry, J. ACM (JACM) 8 (3) (1961) 404–417.
[89] B. Massey, M. Thomure, R. Budrevich, S. Long, Learning spam: simple techniques for freely-available software, FREENIX Track, USENIX Annual Technical Conference, 2003.
[90] S. Mellouli, F. Bouslama, A. Akande, An Ontology for representing finacial headline news, Web Semantics 8 (2-3) (2010) 203–208 Elsevier.
[91] V. Metsis, I. Androutsopoulos, G. Paliouras, Spam filtering with Naive Bayes - which Naive Bayes? 3rd Conference on Email and Anti-Spam (CEAS), ACM, 2006 July 27-28, USA.
[92] M.A. Mittermayer, Forecasting intraday stock price trends with text mining techniques, 37th Annual Hawaii International Conference on System Sciences (HICSS), IEEE, 2004.
[93] D. Miyamoto, H. Hazeyama, Y. Kadobayashi, An Evaluation of machine learning-based methods for detection of phishing sites, in: ICONIP-2008, Auckland, New Zealand, Springer, 2008, pp. 539–546.
[94] A. Moniz, F.D. Jong, Classifying the influence of negative affect expressed by the financial media on investor behavior, in: 5th Information Interaction in Context Symposium (IIiX), Regensburg, Germany, ACM, 2014, pp. 275–278.
[95] A.K. Nassirtoussi, S. Aghabozorgi, T.Y. Wah, D.C. Ling Ngo, Text mining for market prediction: a systematic review, Expert Syst. Appl. 41 (16) (2014) 7653–7670 Elsevier.
[96] A.K. Nassirtoussi, S. Aghabozorgi, T.Y. Wah, D.C. Ling Ngo, Text mining for news-headlines for FOREX market prediction: a multi-layer dimension reduction algorithm with semantics and sentiment, Expert Syst. Appl. 42 (1) (2015) 306–324 Elsevier.
[97] Net Losses: Estimating the Global cost of Cybercrime, 2014. www.mcafee.com/in/resources/reports/rp-economic-impact-cybercrime2.pdf.
[98] P.S.M. Nizer, J.C. Nievola, Predicting published news effect in the Brazilian stock market, Expert Syst. Appl. 39 (12) (2012) 10674–10680 Elsevier.
[99] Bo. Pang, L. Lee, Opinion mining and sentiment analysis, Foundat. Trends Inform. Retrieval 2 (1) (2008) 1–135 NOW Publishers.
[100] Bo. Pang, L. Lee, S. Vaithyanathan, Thums up? Sentiment classification using machine learning techniques, in: Conference on Empirical Methods in Natural Language Processing (EMNLP), July 6-7, Philadelphia, PA, USA, ACM, 2002, pp. 79–86.
[101] Y. Pan, X. Ding, Anomaly based web phishing page detection, in: 22nd Annual Computer Security Applications Conference (ACSAC' 06), December 11-15, Miami Beach, FL, IEEE, 2006, pp. 381–392.
[102] A.-M. Popescu, O. Etzioni, Extracting product features and opinions from reviews, in: Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP), October 6-8, Vancouver, Canada, ACM, 2005, pp. 339–346.
[103] N. Provos, D. McNamee, P. Mavrommatis, K. Wang, N. Modadugu, The ghost in the browser analysis of web-based malware, in: First Workshop on Hot Topics in Understanding Botnets (HotBots), April 10, Berkely, CA, USENIX, 2007, p. 4.
[104] J.R. Quinlan, Induction of decision trees, Mach. Learn. 1 (1986) 81–106 Kluwer Academic Pubkishers.
[105] J.R. Quinlan, Simplifying decision trees, Int. J. Man-Mach. Studies 27 (3) (1987) 221–234 Academic Press Ltd.
[106] K. Ravi, V. Ravi, A survey on opinion mining and sentiment analysis: tasks, approaches and applications, Knowl.-Based Syst. 89 (2015) 14–46 Elsevier.
[107] S. Rawat, V.P. Gulati, A.K. Pujari, V.R. Vemuri, Intrusion detection using text processing techniques with a binary-weighted cosine metric, J. Inform. Assurance Security 1 (1) (2006) 43–50 MIR Labs.
[108] D.E. Rumelhart, G.E. Hinton, R.J. Williams, Learning representations by back--propagating errors, Nature 323 (1986) 533–536.
[109] P. Saha, I. Bose, A. Mahanti, A Knowledge base scheme for risk assesment in loan processing by banks, Dec. Support Syst. 84 (2016) 78–88 Elsevier.
[110] G. Salton, C. Buckley, Term-weighting approaches in automatic text retrieval, Inform. Process. Manage. 24 (5) (1988) 513–523 Elsevier.
[111] G. Salton, M.J. McGill, Introduction to Modern Information Retrieval, McGraw-Hill, Inc., New York, NY, USA, 1986.
[112] J. Schafer, J. Konstan, J. Reidl, Electronic commerce recommender applications, Data Mining Knowl. Discovery 5 (1) (2000) 115–153 Kluwer Academic Publishers.
[113] R.P. Schumaker, H. Chen, Textual analysis of stock market prediction using breaking financial news: the AZFin text system, ACM Trans. Inform. Syst. 27 (2) (2009) Article 12, ACM.
[114] R.P. Schumaker, Y. Zhang, C.-N. Huang, H. Chen, Evaluating sentiment in financial news articles, Dec. Support Syst. 53 (3) (2012) 458–464 Elsevier.
[115] F. Sebastiani, Machine learning in automated text categorization, ACM Comput. Surv. 34 (1) (2002) 1–47 ACM.
[116] C.E. Shannon, A mathematical theory of communication, Bell Syst. Tech. J. 27 (1948) 379–423 and 623–656.
[117] A. Sharma, A. Pujari, K.K. Paliwal, Intrusion detection using text processing techniques with a kernel based similarity measure, Comput. Security 26 (7-8) (2007) 488–495 Elsevier.
[118] S. Sheng, B. Wardman, G. Warner, L.F. Cranor, J. Hong, C. Zhang, An empirical analysis of phishing blacklists, Sixth Conference on Email and AntiSpam (CEAS), July 16-17, ACM, 2009.
[119] C.Y. Shirata, H. Takeuchi, S. Ogino, H. Watanabe, Extracting key phrases as predictors of coroporate bankrtuptcy: empirical analysis of annual reports by text mining, J. Emerging Technol. Account. 8 (2011) 31–44.
[120] G. Suarez-Tangil, J.E. Tapiador, P. Peris-Lopez, J. Blasco, DENDROID: a text mining approach to analyzing and classifying code structures in android malware families, Expert Syst. Appl. 41 (4) (2014) 1104–1117 Elsevier.
[121] G.G. Sundarkumar, V. Ravi, I. Nwogu, V. Govindaraju, Malware detection via API calls, topic models and machine learning, in: 11th International Conference on Automation Science and Engineering (CASE), Aug. 24-28, Gothenburg, Sweden, IEEE, 2015, pp. 1212–1217.
[122] Symantec Internet Security Threat Report (ISTA), 2016. https://www.symantec.com/security-center/threat-report.
[123] P.C. Tetlock, M. Saar-Tsechansky, S. Macskassy, More than words: quantifying language to measure firms' fundamentals, J. Finance 63 (3) (2008) 1437–1467 Blackwell publishing Inc..
[124] J.D. Thomas, K. Sycara, Integrating genetic algorithms and text learning for financial prediction, in: GECCO, July 8-12, Las Vegas, USA, ACM, 2000, pp. 72–75.
[125] D. Thorleuchter, D. Van den Poel, Predicting e-commerce company success by mining the text of its publicly-accessible website, Expert Syst. Appl. 39 (17) (2012) 13026–13034 Elsevier.
[126] A.K. Uysal, S. Gunal, The impact of preprocessing on text classification, Inform. Process. Manage. 50 (1) (2014) 104–112 Elsevier.
[127] V.N. Vapnik, Statistical Learning Theory, JohnWiley & Sons, NewYork, 1998.
[128] A. Vasudevan, R. Yerraballi, Spike: Engineering malware analysis tools using unobtrusive binary-instrumentation, in: 29th Australasian Computer Science Conference (ACSC), January 16-19, Hobart, Tasmania, ACM, 2006, pp. 311–320.

[129] T.T. Vu, S. Chang, Q.T. Ha, N. Collier, An experiment in integrating sentiment features for tech stock prediction in twitter, in: Workshop on Information extraction and entity analytics on social media data, COLING, Mumbai, India, 2012, pp. 23–38.
[130] Wang, J.S. Stolfo, Anomalous payload-based network intrusion detection, in: Symposium on Research in Attacks, Intrusions and Defenses (RAID), September 15-17, Sophia Antipolis, France, Springer, 2004, pp. 203–222.
[131] S. Wang, K. Xu, L. Liu, B. Fang, S. Liao, H. Wang, An ontology based framework for mining dependence relationships between news and financial instruments, Expert Syst. Appl. 38 (10) (2011) 12044–12050 Elsevier.
[132] S. Wang, Z. Zhe, Ye. Kang, H. Wang, X. Chen, An ontology for casual relationships between news and financial intruments, Expert Syst. Appl. 35 (3) (2008) 569–580 Elsevier.
[133] Y. Yang, J.O. Pedersen, A comparative study on feature selection in text categorization, in: 14th International Conference on Machine Learning (ICML), Nashville, TN, USA, Morgan Kaufmann Publishers, 1997, pp. 412–420.
[134] Y. Ye, T. Li, Y. Chen, Q. Jiang, Automatic malware categorization using cluster ensemble, in: KDD 10, July 25-28, Washington, DC, USA, ACM, 2010, pp. 95–104.
[135] Y. Ye, D. Wang, T. Li, D. Ye, IMDS: intelligent malware detection system, in: 13th KDD, August 12-15, San Jose, California, USA, ACM, 2007, pp. 1043–1047.
[136] Y. Yu, W. Duan, Q. Cao, The impact of social and conventional media on firm equity value: a sentiment analysis approach, Dec. Support Syst. 55 (4) (2013) 919–926 Elsevier.
[137] J. Zhan, B.J. Oommen, J. Crisostomo, Anamolay detection in dynamic systems using weak estimators, ACM Trans. Internet Technol. (TOIT) 11 (1) (2011) Article 3, ACM.
[138] L. Zhang, J. Zhu, T. Yao, An evaluation of statistical spam filtering techniques, ACM Trans. Asian Language Inform. Process. (TALIP) 3 (4) (2004) 243–269 ACM.
[139] W. Zhuang, Y. Ye, Y. Chen, T. Li, Ensemble clustering for internet security applications, IEEE Trans. Systems Man Cybern Part C 42 (6) (2012) 1784–1796 IEEE.