ONLINE DICTIONARY LEARNING TECHNIQUES FOR FINANCIAL NEWS ANALYSIS

Elmira Navidbakhsh

B.Eng. Ryerson University , Canada, 2012

PROJECT

presented to Ryerson University

in partial fulfillment of the
requirements for the degree of
Master of Engineering
in
Electrical and Computer Engineering

Toronto, Ontario,Canada

DECLARATION

I hereby declare that I am the sole author of this PROJECT. This is a true copy of the PROJECT, including any required final revisions.

I authorize Ryerson University to lend this PROJECT to other institutions or individuals for the purpose of scholarly research.

I further authorize Ryerson University to reproduce this PROJECT by photocopying or by other means, in whole or in part, at the request of other institutions or individuals for the purpose of scholarly research.

I understand that my PROJECT may be made available to the public in electronic form.

ONLINE DICTIONARY LEARNING TECHNIQUES FOR FINANCIAL NEWS ANALYSIS

by

Elmira Navidbakhsh

Master of Engineering in Electrical and Computer Engineering

Department of Electrical & Computer Engineering

Ryerson University

Toronto, Ontario, Canada, 2014

ABSTRACT

The investor's sentiments can be defined an investor's attitude and opinion towards investing in the stocks market. Investor sentiment has been traditionally regarded as a myth by classical financial theories and has received little attention by researchers prior to 1990. The standard argument was that in the highly competitive financial market, suboptimal trading behaviors, such as paying attention to sentiment signals, is unrelated to fundamental value. It has been proposed by the efficient market hypothesis (EMH) that markets are efficient in that opportunities for profit are discovered so quickly that they cease to be opportunities. The EMH effectively states that no system can continually beat the market because if that system were to become public, everyone would use it, thus negating any potential gain. From the literature, it is evident that the application of investor sentiment for evaluating market behavior is achieving broad acceptance. This paper studies the application of Soft Computing to Investor Sentiment, focusing on the dictionary learning approach. Soft computing methods and various sentiment indicators are employed to obtain sample predictions of future trends in stock market returns. This paper's contribution is to expose the key areas where research is being undertaken, and to attempt to quantify the degree of success associated with the different research approaches.

ACKNOWLEDGEMENTS

I offer my gratitude to Ryerson University faculty, staff and classmates who have inspired me to continue my work in Electrical and Computer Engineering. It would not have been possible without the kind support and help of many individuals and organizations. I would like to thank Dr. Kaamram Raahemifar for his patience and invaluable guidance as the ideal thesis supervisor. I also appreciate Ryerson University for its friendly researching environment and helpful resources. Finally, I would like to thank my family for the constant love and support.

I would like to thank S. P. Kasiviswanathan for kindly sharing his MATLAB code through e-mail communication in order to reproduce the simulation, including the ordered data.

DEDICATION

This thesis is dedicated to my family, which taught me that the best kind of knowledge to have is that which is learned for its own sake and that the largest task can be accomplished if done one step at a time. Special thanks to my family for being there with constant love and support.

Table of Contents

List of Figures

Chapter 1

Introduction

The high volume and velocity of social media have propelled such things as blogs and Twitter to the forefront as sources of breaking news. It would appear that news affects the markets in profound ways by impacting volumes of trade, stock returns, volatility of prices and even future firm earnings. In financial domain of news impact analysis, the focus has expanded in recent years from informational to affective content of text in an effort to explain the relationship between text and the markets. All text, be it news, blogs, accounting reports or poetry, has a non-factual dimension that conveys opinion, invokes emotion, and provides a nuanced perspective of the factual content of the text. At present, the most popular approach to automated sentiment analysis at the level of the text involves using machine learning technology to build automated classifiers from human annotated documents. This method has shown much initial promise, particularly because it allows researchers to abstract away from the messy linguistic details by providing an impressive baseline performance in text polarity identification, even with the simplest of features [1]. Accelerating analysis of information content in online social media streams is the need of the hour since it allows businesses and government to understand public opinion about products and policies. In most of these settings, data points appear as a stream of high dimensional feature vectors. Recently, the dictionary learning approach, as one of machine learning methods, has emerged as a powerful data representation framework in sentiment analysis. Dictionary learning is the problem of estimating a collection of basis vectors over which a given data collection can be accurately reconstructed, often with sparse encodings. It may be formulated in terms of uncovering low-rank structure in the data using matrix factorizations possibly with sparsity inducing priors.

The contribution of this research is a consistent presentation and classification of Dictionary learning techniques, specifically Online l_1-dictionary learning, applied to financial markets news. This may be used for further analysis and evaluation, as well as comparative studies. An obvious benefit of this study is that if one applies online l_1-dictionary learning to the financial news, valuable results will be obtained, which, when analyzed, may offer additional information to market behaviour.

1.1 Organization

Although the end goal of this study is to forecast the stock market using financial news, first, there is much fascinating groundwork to be laid, interesting in and of it.

This study is organized into three substantive chapters. In the rest of this chapter, I present the motivation for the work, as well as a description of how the study is organized.

Chapter 2 sets up a framework for online l_1-dictionary learning algorithm for novel document detection that will be able to detect breaking news and trending relevant to the financial market.

Chapter 3 explores news analytics and its application to finance, including textual data, internet message boards.

Chapter 4 outlines the details of the evaluation of this research.

Chapter 5 presents the results and concludes with a look at future challenges for this research.

1.2 Motivation

There are several motivations for forecasting stock market prices. The most basic of these is financial gain. Any system that can consistently pick winners and losers in the active marketplace would make the owner of the system very wealthy. Moreover it has been proposed in the efficient market hypothesis (EMH) that markets are efficient in that opportunities for profit are discovered so fast that they cease to be opportunities. Though there are many ways to approach the efficient market hypothesis, the intuition behind it is that markets efficiently process all related information into a single price. In principle, past data cannot be used to predict future prices in capital markets. One of the key questions in the definition of efficient markets is what type of information is relevant. Forecasting stock movements based only on past prices is very different than forecasting stock movements based on insider information about a pending merger. In addition, the presence of systematic mispricing in the market remains debatable because of the complexity of empirically examining the issue. After the empirical evidence was documented, the EMH was proposed based on the overpowering logic that if returns were forecast- table, many investors would use them to make unlimited profits. The investor sentiment induced returns that obey the EMH; otherwise, the stock market would become a "money machine" producing infinite wealth. A steady economy cannot allow that. The absence of precise valuation models for stocks makes it difficult to measure deviations from theoretical prices. Similar problems arise due to the difficulty in measuring investor sentiment.

1.3 Literature Review

The work presented in this study does not easily fit into a single field of research; it addresses concerns in both finance and computer science. Even within those broad disciplines, the work draws from diverse research threads, such as studies of trading rules and event studies in finance, ensemble methods, dictionary learning, and sparse encoding.

Ankan Saha and Vikas Sindhwani [2] proposed an online nonnegative matrix factorization framework to capture the evolution and emergence of themes in unstructured text under a novel temporal regularization framework. They develop scalable optimization algorithms for their framework, proposed a new set of evaluation metrics, and reported promising empirical results with traditional TDT tasks as well as streaming Twitter data.

News can contain information which may provide an indication of the future direction of a share or stock market index. A recent paper by Brett Drury et.al [3] clearly demonstrates that a news story's headline provides the greatest assistance for classification. His paper presented a strategy that combined the: rule a classifier, alignment an strategy and self-training to induce a robust model for classifying news stories. The models induced from headlines gained the highest estimated F- Measure and trading returns for each strategy with the exception of the alignment method, which consistently performed poorly.

As the social media, such as Twitter, have become leading sources of breaking news, a key task in the automated identification of such news is to detection novel documents from a voluminous stream of text documents in a scalable manner. Motivated by this challenge, the squared loss, the l_1-penalty, is used for measuring the reconstruction error that has been introduced in a paper by

Shiva.P. Kasiviswanthan et.al [4]. Online convex optimization is an area of active research; for a detailed survey on the literature I refer the reader to [5].

Determining the sentimental polarities of stock market news is another approach in investor sentiment analysis. The study conducted by Keisuke Mizumoto et.al [6] proposes an automatic dictionary construction approach and sentiment analysis of stock market news using a polarity dictionary. A semi-supervised learning approach has been used in the construction of the dictionary.

Previous work involves many other techniques in the wide area of statistics or machine learning. Some important methods are as follows:

Dictionary Learning: The problem of estimating a collection of basis vectors over which a given data collection can be accurately reconstructed, often with sparse encodings.

Linear Regression: The most basic technique to describe the relationship between response variables and explanatory variables. Using β and $\varepsilon(t)$to denote the regression coefficients and the error terms, respectively, the linear regression model often takes the form $Y(t) = X(t)\beta + \varepsilon(t)$.

Approximate Nearest Neighbor: One of the simplest learning algorithms. Objects are classified based on the closest training examples in the feature space.

Bayes Learning: A learning algorithm based on Bayes' theorem regarding conditional and marginal probabilities. Bayes learning is one of the most successful algorithms for classifying text documents and has shown great success in many previous applications. Bayes learning method is simple yet robust.

Artificial Neural Networks: ANNs have also been extensively used as nonlinear mapping for statistical modeling. They have been applied in many areas, especially in applications for classification, clustering or prediction.

Support Vector Machines: Support Vector Machines (SVMs) are a set of related supervised learning methods used for classification and regression. Usually SVMs perform better than multiple linear regression models because the real world cannot always be described with a linear model.

Time Series Analysis: forecasting future data points using historical data sets. Research reviewed in this area generally attempts to predict the future values of some time series. Possible time series include Base time series data (e.g. Closing Prices), or time series derived from base data (e.g. Indicators - frequently used in Technical Analysis).

Chapter 2

2.1 Dictionary learning

Dictionary learning is a useful procedure by which dependencies among input features can be represented in terms of suitable bases. It has found applications in many machine learning and inference tasks, including image de-noising, dimensionality-reduction, bi-clustering, feature-extraction and classification, and novel document detection [4]. Dictionary learning usually alternates between the following two steps: an inference (sparse coding) step and a dictionary update step. The first step finds a sparse representation for the input data using the existing dictionary by solving, for example, an l_1 -regularized regression problem. In contrast, the second step usually employs gradient descent approximation to update the dictionary entries. With the increasing complexity of various learning tasks; it is natural that the size of the learning dictionaries is demanding more and more memory and computation. Therefore it is important to study situations where the dictionary need not be available in a single central location but could be spread out over multiple locations. This is mainly true in big data scenarios where multiple large dictionary models may already be available at separate locations and it unfeasible to aggregate all dictionaries into one location due to communication and privacy considerations. This observation motivates us to examine how to learn a dictionary model that is stored over a network of agents, where each agent is in charge of only a portion of the dictionary elements.

2.1.1 Online Dictionary Learning

Dictionary learning is the problem of estimating a collection of basis vectors over which a given data collection can be accurately reconstructed, often with sparse encodings. It falls into a

general category of techniques known as matrix factorization. Classical dictionary learning techniques (e.g., [7, 8, 9]) address the problem of learning a reconstructive dictionary D in $R^{m\times k}$ well adapted to a training set T which is not jointly convex in (D, α), but convex with respect to each unknown when the other one is fixed. In dictionary learning each data point y is represented as a sparse linear combination Dx of dictionary atoms, where D is the dictionary and x is a sparse vector [1, 2], as shown in Figure 1. Classic dictionary learning techniques for sparse representation ([1, 3, and 2] and references therein) consider a finite training set of signals $S = [s_1 \dots s_i] \in R^{m\times n}$ and optimize the empirical cost function which is defined as:

$$f(D) = \sum_{i=1}^{n} l(s_i, D) \quad (1)$$

Where; D in $R^{m\times k}$ is the dictionary, each column representing a basis vector; l_1 is a loss function with an l_1 -regularization term; $l(s, D) \triangleq \min_{a\in R^k} \frac{1}{2}\|s - D\alpha\|_2^2 + \lambda_1\|\alpha\|_1$ such that $l(s, D)$should be small if D is "accurate" at representing the signal s in a sparse fashion. The optimization problem of dictionary learning as that of minimizing the empirical cost (D) is the following:

$$\min_A f(D) = f(D, X) \stackrel{\text{def}}{=} \min_{D,X} \sum_{i=1}^{n} l(s_i, D) = \min_{D,X} \|S - DX\|_1 + \lambda_1\|X\|_1$$

Additionally, for maintaining interpretability of the results we want dictionary D and X's to contain non-negative entries. Therefore, the problem of dictionary learning becomes

$$\min_{D\in D', X\geq 0} \|S - DX\|_1 + \lambda_1 \|X\|_1 \quad (2)$$

Where D´ is the convex set of matrices; to prevent D from being arbitrarily large (which would lead to arbitrarily small values of X), we add a scaling constant on D as follows: $D' = \left\{D \in R^{m\times k} : D \geq 0_{m\times k} \forall j = 1, \dots k \left\|D_j\right\|_1 \leq 1\right\}$,with D_j being the jth column in D. The

optimization problem (2) is in general non-convex. But if one of the variables, either D or X is known, the objective function with respect to the other variable becomes a convex function (in fact, a linear function) and the global solution to (2) can be found. This iterative alternative minimization is the core idea behind most algorithms for dictionary learning [7, 8, and 9].

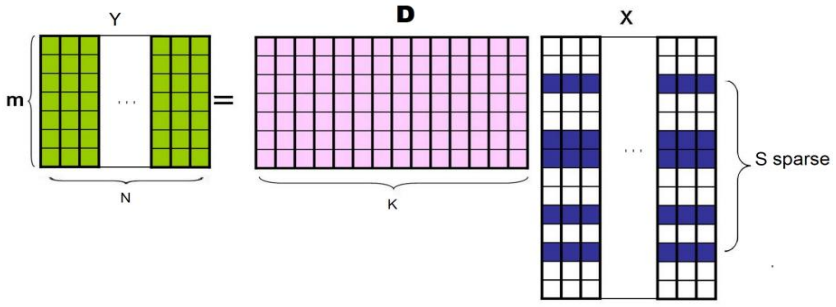

Figure 1.Sparse linear combination Dx of dictionary atoms

2.1.2 Sparse Coding

In classical sparse coding tasks, we consider the signal s in $R^{m\times n}$ and the fixed dictionary D in $R^{m\times k}$. Note that in this setting over-complete dictionaries with $k > m$ are allowed. The number of samples *n* is usually large, whereas the signal dimension *m* is relatively small. In general, we also have $k \ll n$, but each signal uses only a few elements of D in its representation, for instance 10. In this setting, sparse coding with l_1- regularization amounts to computing:

$$l(s, D) \triangleq \min_{\alpha\in R^k} \frac{1}{2}\|s - D\alpha\|_2^2 + \lambda_1\|\alpha\|_1 \qquad (2)$$

Where $l(s, D)$ is defined as the optimal value of the l_1- *sparse coding* problem and λ_1 is a regularization parameter. The l_p regularization term of a vector x for p ≥ 0 is defined as $\|x\|_p^p = \sum_{i=1}^{n}|x[i]|^p$. $\|.\|_p$ is a norm when $p \geq 1$. When $p = 0$, it counts the number of non-zero elements in the vector.

2.1.3 Novel Document Detection Task

Let $D_{t-1} \in R^{m \times k}$, represent the dictionary matrix after time t−1; where dictionary D_{t-1} is a compact summary representation of all the documents in S ≤ t−1. Here, S_t represents the term-document matrix observed at time t. Each column of D_{t-1} is called a basis vector or atom. The exact construction of the dictionary is described later, but ideally we want the dictionary to have a set of representative atoms for each of the old topics. With such a representative dictionary, documents from old topics can be represented as a linear combination of the atoms corresponding to that topic. Let N_t be the number of documents arriving at time ≤ t, then $S[t] \in R^{m \times N_t}$. Under this setup, the goal of novel document detection is to identify documents in S_t that are "dissimilar" to the documents in $S_{[t-1]}$. Given a new document vector y with timestamp t, we see if y could be represented as a sparse linear combination of the columns of D_{t-1}. The sparsest representation is the solution of:

$$\min_x \|x\|_0 \ s.t\ y = D_t - 1\ x,\ x \geq 0 \qquad (3)$$

Where‖. ‖is the l_0 −norm, counting the non-zero entries of a vector. However in general, solving (3) is NP-hard and also difficult to approximate [10]. Recently, a series of papers (see [11, 12] and references therein) have shown that under certain favorable conditions one could obtain the solution to (3) by solving the following:

$$\min_x \|x\|_1 \ s.t\ y = D_t - 1\ x,\ x \geq 0 \qquad (4)$$

In essence, (4) can be viewed as a convex relaxation of (3). In our experiments, we use the alternating directions method of multipliers (ADMM) [8] to solve (3). ADMM has recently gathered significant attention in the machine learning community due to its wide applicability to

a range of learning problems with complex objective functions [13]. For each document **y** arriving at time t, first we solve (3) to check whether **y** could be well approximated as a sparse linear combination of the atoms of D_t . If the objective value $D_{t-1}\, x$ is "big" then we mark the document as novel. Since all the documents should be normalized in S_t to unit l_1- length the objective values should also be in the same scale. In the presence of isotopic Gaussian noise the l_2-penalty on **e** = **y** − $D_{t-1}\, x$ gives the best approximation of x [14, 15, 16]. However, for text documents (and in most other real scenarios), the noise vector e rarely satisfies the Gaussian assumption, and some of its coefficients contain large, impulsive values.

2.1.4 Dictionary Size

Changing the size of the dictionary (k) dynamically with t would lead to a more efficient and effective sparse coding. However, here we make the simplifying assumption that k is a constant independent of t. The problem of designing an adaptive dictionary whose size automatically increases or decrease over time is a remarkable open problem. While the current work uses fixed sized dictionaries, using adaptive dictionaries whose size changes based on the set of active or emerging topics may be more desirable in certain applications.

2.1.5 Online l_1-Dictionary Learning

The standard goal of online learning is to design algorithms whose regret is sublinear in time T, since this implies that "on the average" the algorithm performs as well as the best fixed strategy in hindsight [17]. An efficient online algorithm with sublinear regret would imply an efficient algorithm for solving (1) in the offline case. Therefore making assumptions on either D or X would make it possible to design an efficient online algorithm with sublinear regret. Therefore, the focus of this work is on obtaining regret bounds for the dictionary update, assuming that at

each time-step the sparse codes given to the online algorithms are "close". This motivates the following problem:

Definition 2.4.1 (Online l_1-Dictionary Learning Problem). *At time t, the online algorithm picks* $\widehat{D}_{t+1} \in D'$ *then, the nature (adversary) reveals* $\left(S_{t+1}, \hat{X}_{t+1}\right)$ *with* $S_{t+1} \in R^{m\times n}$ *and* $\widehat{X}_{t+1} \in R^{k\times n}$. *The problem is to pick the* D_{t+1} *sequence such that the following regret function is minimized*

$$R(T) = \sum_{t=1}^{T} \left\| S_t - \widehat{D}\hat{X}_t \right\|_1 - \min_{D\in D'} \sum_{t=1}^{T} \|S_t - DX_t\|_1$$

Where $\widehat{X}_t = X_t + E_t$ *and* E_t *is an error matrix dependent on t.*

2.1.6 Online l_1-Dictionary Algorithm

This section is the detailed design of and algorithm for the online ℓ_1-dictionary learning problem, which is called Online Inexact ADMM (OIADMM), and bound its regret because the algorithm is based on the alternating directions method of multipliers. This algorithm first performs a simple variable substitution by introducing an equality constraint. The update for each of the resulting variables has a closed-form solution without the need of explicitly estimating the sub-gradients.

Algorithm1: Online Inexact ADMM (OIADMM)

Input: $\boldsymbol{S_t \in R^{m\times n}, \widehat{D}_t \in R^{m\times k}, \Delta_t \in R^{m\times n}, \widehat{X}_t \in R^{k\times n}, \beta_t \geq 0, \tau_t \geq 0}$

$$\widetilde{\Gamma}_t \leftarrow S_t - \widehat{D}_t \widehat{X}_t$$

$$\Gamma_t = argmin_{\Gamma} \|\Gamma\|_1 + \langle \Delta_t, \widetilde{\Gamma}_t - \Gamma \rangle + (\beta_t/2) \left\|\widetilde{\Gamma}_t - \Gamma\right\|_F^2$$

$$\left(\Rightarrow \Gamma_{t+1} = soft\left(\widetilde{\Gamma}_t + \frac{\Delta_t}{\beta_t}, \frac{1}{\beta_t}\right)\right)$$

$$G_{t+1} \leftarrow -\left(\frac{\Delta_t}{\beta_t} + \widetilde{\Gamma}_t - \Gamma_{t+1}\right)\widehat{X}_t^T$$

$$\widehat{D}_{t+1} = argmin_{D\in D} \beta_t \left(\langle G_{t+1}, D - \widehat{D}_t \rangle + (1/2\tau_t)\left\|D - \widehat{D}_t\right\|_F^2\right)$$

$$\left(\Rightarrow \widehat{D}_{t+1} = \prod_A \left(max\{0, \widehat{D}_t - \tau_t G_{t+1}\}\right)\right)$$

$$\Delta_{t+1} = \Delta_t + \beta_t\left(S_t - \widehat{D}_{t+1}\widehat{X}_t - \Gamma_{t+1}\right)$$

Return $\widehat{D}_{t+1}$ **and** Δ_{t+1}

OIADMM is summarized in Algorithm 1. Consider the following minimization problem at time t

$$min_{A\in A'} \left\|S_t - D\hat{X}_t\right\|_1$$

We can rewrite this above minimization problem as:

$$\min_{D\in D', \Gamma} \|\Gamma\|_1 such\ that\ S_t - D\hat{X}_t = \Gamma \quad (4)$$

The augmented Lagrangian of (4) is:

$$\mathcal{L}(D, \Gamma, \Delta) = \min_{D\in D', \Gamma} \|\Gamma\|_1 + \langle \Delta, S - D\hat{X}_t - \Gamma \rangle + (\beta_t/2)\left\|S_t - D\hat{X}_t - \Gamma\right\|_F^2 \quad (5)$$

Where $D_t \in R^{m\times n}$ is a multiplier and $\beta_t \geq 0$ is a penalty parameter.

Chapter 3

3.1 News Analytics and Its Application to Finance

News analytics measures news attributes such as sentiment, relevance and novelty by expressing news stories as numbers to permit the management of everyday information in a mathematical and statistical manner. News analytics are used in finance, particularly in quantitative applications and plotting and characterizing company behaviours over time; thus, they yield important strategic insights about a rival company. Predicting the market has always been one of the hottest topics in research, so is the challenge due to its complex, non-stationary, noisy, chaotic, nonlinear and dynamic system that does not follow the random walk process [21]. There are many factors that may cause the fluctuation of financial market movement, including economic conditions, political situations, traders' expectations, catastrophes and other unexpected events. Therefore, predictions of stock market price and its direction are quite difficult. Researchers have proof that suggests that financial time series is predictable. Numerous publications have attempted to construct an accurate model for the stock market. Most of these works focus on time series prediction with various models, such as Artificial Neural Networks [22] [23] Hybrid Genetic Algorithm and Particle Swarm Optimization [24], Fuzzy Logic [25] and some hybrid combinations like the use of ten data mining techniques that include Linear discriminant analysis (LDA), Quadratic discriminant analysis (QDA), K-nearest neighbour classification, Naive Bayes based on kernel estimation, Logit model, Tree based classification, neural network, Bayesian classification with Gaussian process, Support vector machine (SVM) and Least squares support vector machine (LS-SVM) [26]. Analytics is a term that refers to the various modes of using information to make decisions. Traditionally, this is called decision support, which is mostly accomplished through the decision support tools, data warehouse and

business intelligence tools, which have become more sophisticated for data access, data analysis, data manipulation, data mining, forecasting, trend analysis and other metric-based presentations. Analysis tools include packaged analytical applications for specific business domains, such as supply chain analysis, sales channel analysis, performance analysis, etc. Data mining refers to extracting or "mining" knowledge from large amounts of data. Data mining is often treated as a synonym for another popularly used term, Knowledge Discovery from Data, or KDD. Alternatively, others view data mining as simply an essential step in the process of knowledge discovery.

The sources and volume of news have increased significantly over the years; as a result the consumers of financial news must work hard to achieve superior investment returns by using the information available to them. News Analytics offers automatic analyses of published news. One of the tasks of News Analytics (NA) is determining novelty of the published text, i.e. automatically deciding whether the news story at hand is reporting about some new event, or it is merely introducing new facts about an event that is already known. Another example of a news analytics task is detecting sentiment within the published news item. Analysing the language of a text, and the selection of the words the author used, it is possible to determine whether the analysed text is positive or negative in sentiment, as well as the degree of positivity or negativity. The texts lacking high levels of positivity or negativity are usually denoted as neutral. Such scales we call text sentiment measures.

News analytics has introduced technology in order to automate or semi-automate this approach. By automating the judgement process, the human decision maker can act on a larger, hence more diversified, collection of assets. These decisions are also taken more promptly. Automation or semi-automation of the human judgement process widens the horizons of the investment process

[27]. Automated news analysis can form a key component driving algorithmic trading desks' strategies and execution, and the traders who use this technology can shorten the time it takes them to react to breaking stories.

For investors in all market sectors, the goal of news analytics is to; understand market cycles based on the psychological perceptions in each market and asset class and make better asset allocation decisions then drill down into sectors and choose strategies. Another goal of news analytics is to monitor risk perceptions identify fear in different locations to change the pricing of products in response and hedge effectively in times of high uncertainty, monitor risk perceptions indifferent currencies, interest rates in certain economic sectors, and conflict in select countries or as certain levels of concern about unemployment and layoffs which may affect policy decisions and help new companies focus on the subjects people care about the most and understand why and how they care about them.

3.1.1 Types of News Data

Financial news consists mostly of two types: first, regular synchronous announcements, which are scheduled and expected by investors and second, event-driven asynchronous announcements, which are unscheduled and unexpected. Mainstream news rumours and social media normally arrive asynchronously in an unstructured textual form. A substantial portion of prenews arrives at prescheduled times and in a generally structured form. Scheduled announcements often have a well-defined numerical and textual content and may be classified as structured data. These include macroeconomic announcements and earnings announcements. Macroeconomic news, particularly economic indicators from the major economies, is widely used in automated trading. News has an impact in the largest and most liquid markets, such as foreign exchange,

government debt and futures markets. Firms often execute large and rapid trading strategies. These news events are normally well documented, thus thorough back testing of strategies is feasible. Since indicators are released according to a precise schedule, market participants can be well prepared to deal with them. These strategies often lead to companies fighting to be first to the market; speed and accuracy are the major determinants of success.

The different types of news and information flows that can be applied for updating investor beliefs were discussed in books and literature [27] [28]. The study distinguishes four broad classifications of news. Figure 2 shows the news flow architecture.

News This refers to mainstream media and comprises the news stories produced by reputable sources. These are broadcast via newspapers, radio and television. They are also delivered to traders' desks on newswire services. Online versions of newspapers may also exist.

Pre-News This refers to the source data that reporter's research before they write news articles. It comes from primary information sources, such as Securities and Exchange Commission reports and filings, court documents and government agencies. It also includes scheduled 141 announcements such as macro-economic news, industry statistics, company earnings reports, analyst reports, annual reports and other corporate news.

Rumours These are blogs and websites that broadcast "news" and are less reputable than news and pre-news sources. The quality of these varies significantly. Some may be blogs associated with highly reputable news providers and reporters. At the other end of the scale some blogs may lack any substance and may be fuelled entirely by rumour.

Social media These websites fall at the lowest end of the reputation scale. Barriers to entry are extremely low and the ability to publish "information" is easy. These can be dangerously inaccurate sources of information. However, if carefully applied (with consideration of human

behaviour and agendas) some value may be gleaned from these. At a minimum they may help us identify future volatility.

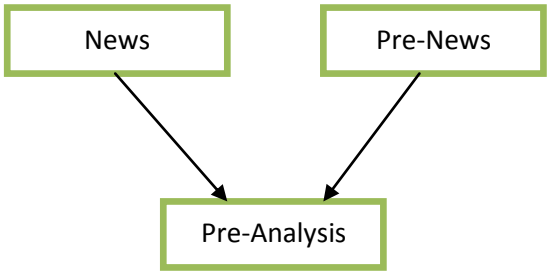

Figure 2.News flow

3.1.2 The Dataset TDT2

The dataset is drawn from the TDT2 corpus (NIST Topic Detection and Tracking corpus) [30] consists of data collected during the first half of 1998 and taken from six sources, including two newswires (APW, NYT), two radio programs (VOA, PRI) and two television programs (CNN, ABC). It consists of 11,201 on-topic documents which are classified into 96 semantic categories. In this subset, those documents appearing in two or more categories were removed, and only the largest 30 categories were kept, thus leaving 9,394 documents in total. To evaluate of this research, a set of 9000 documents was used that represented over 19,528 terms and distributed into the top 30 TDT2 human-labelled topics over a period of 27 weeks. The documents are introduced in groups.

The purpose of the TDT effort is to advance the frontier in Topic Detection and Tracking. TDT processing addresses multiple sources of information, including both newswire (text) and broadcast news (speech). The information flowing from each source is modeled as a sequence of stories. These stories provide information on many topics. The general technical challenge is to identify and follow the topics being discussed in these reports.

The definition of "topic" is a fundamental issue and of the greatest importance. It is also a very difficult problem, one which has not been fully resolved and for which no perfect solution exists. However, for the purposes of the TDT research effort, a topic is defined as "*a seminal event or activity, along with all directly related events and activities"* [29]. Stories will be considered to be "on topic" whenever the story is *directly* connected to the associated event. Topic detection is the task of detecting and tracking topics not previously known to the system. It is characterized by a lack of knowledge of the topic being detected. Therefore the system must embody an understanding of what a topic is, and this understanding must be *independent of topic specifics*. In the topic detection task, the system must detect new topics as the incoming stories are processed and then associate input stories with those topics. Thus this process identifies a set of topics, as defined by their association with the stories that discuss them. Topic detection performance depends on the form of the source and on the maximum permitted delay before topic detection decisions must be outputted. Performance also depends on whether story boundaries are provided to the system.

Chapter 4

4.1 Experimental Setup

All reported results are based on a MATLAB implementation running on a quad-core 2:33 GHz Intel processor with 16GB RAM. The parameters to our l_1-online dictionary learning algorithm are: (i) initial size of dictionary, (ii) regularization parameter, (iii) parameters to OIADMM (β_t and τ_t), and (iv) ADMM parameters for sparse coding. The regularization parameter λ is set to 0.1 which yields reasonable sparsities in the experiments. OIADMM parameters are as follows: $\tau_t$is set to $1/(2\psi_{max}\left(\hat{X}_t\right))$(chosen according to Theorem 4.6) and β_t is fixed to 5 (obtained through tuning). The ADMM parameters for sparse coding and batch dictionary learning are set as suggested in [4]. In the batch algorithms, the dictionary sizes are increased by η=10 in each timestep. The threshold value ζ is treated as a tuneable parameter.

4.2 Experimental Results

This section presents experiments that compare and contrast the performance of l_1-batch andl_1-online dictionary learning algorithms for the task of novel document detection. This section also present results highlighting the advantage of using an l_1-penalty over an l_2-penalty on the reconstruction error for this task.

4.2.1 Implementation of BATCH

This section describes the batch algorithm for detecting novel documents. The Algorithm BATCH alternates between novel document detection and a batch dictionary learning step.

The Batch Dictionary learning step at time *t*, the dictionary learning step is

$$[D_{t+1}, X_{[t]}] = argmin_{\mathrm{D}\in D, X\geq 0} \left\| S_{[t]} - DX \right\|_1 + \lambda \left\| X \right\|_1 \qquad (3)$$

Even though conceptually simple, Algorithm BATCH is computationally inefficient. The bottleneck comes in the dictionary learning step. As t increases, so does the size of

S[t], so solving (3) becomes prohibitive even with efficient optimization techniques. To achieve computational efficiency, in [4], the authors solved an approximation of (3), where in the dictionary learning step they update only the D's and not the X's. This leads to faster running times, but because of the approximation, the quality of the dictionary degrades over time and the performance of the algorithm decreases. But solving (3) with online learning algorithm will show that this online algorithm is both computationally efficient and generates good quality dictionaries under reasonable assumptions. Most algorithms for dictionary learning are iterative batch procedures, which operate on the entire training set; however this method is not feasible for very large and dynamic data.

In this implementation, the dictionary size grows by η in each time step. Growing the dictionary size is essential for the batch algorithm because as *t* increases the number of columns of $S[t]$ also increases. Therefore, a larger dictionary is required to compactly represent all the documents in $S[t]$. For solving (3), we use alternative minimization over the variables. The complete pseudo-code is given Algorithm BATCH-IMPL. The optimization problems arising in the sparse coding and dictionary learning steps are solved using ADMM's.

4.2.2 The Online Algorithm

The online algorithm by S.Kasiviswanathan et.al uses the same novel document detection step as Algorithm BATCH, but dictionary learning is done using OIADMM. In experiments, the number

of documents introduced during each time step is almost of the same order; hence, there is no need to change the size of the dictionary across time steps for the online algorithm.

Algorithm 3 : Online Dictionary Learning

Input: $\boldsymbol{S_t = [s_1, \dots s_{n_t}] \in R^{m \times n_t}, \widehat{D}_t \in R^{m \times k}, \Delta_t \in R^{m \times n_t}, \lambda \geq 0, \zeta \geq 0, \beta \geq 0, \tau \geq 0}$

Novel document detection step:

For j=1 **to** $\boldsymbol{n_t}$ **do**

Solve: $\boldsymbol{x_j = argmin_{x \geq 0} \|S_j - \widehat{D}_t x\|_1 + \lambda \|x\|_1}$

If $\boldsymbol{\|S_j - \widehat{D}_t x_j\|_1 + \lambda \|x_j\|_1 > \zeta}$

Mark $\boldsymbol{S_j}$ **as novel**

Online Dictionary Learning Step:

Set $\boldsymbol{\widehat{X}_t \leftarrow [x_1, \dots x_{n_t}]}$

$\boldsymbol{(\widehat{D}_{t+1}, \Delta_{t+1}) \leftarrow}$ **OIADMM** $\boldsymbol{(S_t, \widehat{D}_t, \Delta_t, \widehat{X}_t \beta, \tau)}$

The sparse coding matrices of the Algorithm BATCH $X_1, \dots, X_t$ could be different from $\hat{X}_1, \dots, \hat{X}_t$. If these sequences of matrices are close to each other, then we have a sublinear regret on the objective function.

4.2.3 Evaluation of Novel Document Detection.

For performance evaluation, it is assumed that documents in the corpus have been manually identified with a set of topics. For simplicity, we assume that each document is tagged with the single most dominant topic with which it is associated, hereafter called the "true topic" of that document. A document *y* arriving at time *t* is called novel if the true topic of *y* has not appeared

before the time *t*. So at time *t*, given a set of documents, the task of novel document detection is to classify each document as either novel (positive) or non-novel (negative). To evaluating this classification task, the standard Area Under the ROC Curve (AUC) [31] is used.

4.2.4 Performance Evaluation for l_1-Dictionary Learning.

A simple reconstruction error measure is used for comparing the dictionaries produced by our l_1-batch and l_1-online algorithms. The dictionary at time *t* needs to be a good basis to represent all the documents in $S_{[t]} \in R^{m \times N_t}$. This would lead to define the *sparse reconstruction error* (SRE) of a dictionary D at time *t* as

$$SRE\ (D) \stackrel{\text{def}}{=} \frac{1}{N_t}\left(\min_{X \geq 0} \left\| S_{[t]} - DX \right\|_1 + \lambda \left\| X \right\|_1\right).$$

A dictionary with a smaller SRE is better, on average, at sparsely representing the documents in $S_{[t]}$.

Experiments performed comparingl_1- penaltyvs.l_2-penalty justify the choice of using an l_1-penalty (on the reconstruction error) for novel document detection. In the l_2-setting, for the sparse coding step fast implementation of the LARS algorithm with positivity constraints [32] was used and the dictionary learning was done by solving a non-negative matrix factorization problem with additional sparsity constraints(also known as the non-negative sparse coding problem [33]).

4.2.5 Experiment on the Data

The experiment setup is as follows. A collection of 1000 documents is presented to the algorithms in order to initialize the dictionary. The algorithm from [4] utilizes the data as a

block. Once the dictionary is initialized, a new collection of 1000 documents is presented to the algorithm. The algorithm then processes the data samples in order to determine if each of the new documents belongs to a topic that has been previously observed, or not. This assessment is done by determining if the value of the cost function is sufficiently large, in order to deem the data sample "novel." The detection result then produces a receiver operating characteristic (ROC) curve, illustrated in Figures 3-7. For this simulation setup, novel documents are introduced only at the first (samples 1001-2000), second (2001-3000), fifth (5001-6000), sixth (6001-7000), and eighth (8001-9000) time-steps. For this reason, we execute only the novel document detection part of the algorithm at those time-steps and then present the ROC curves for those time-steps. Following the production of the ROC curve, the previously new data set becomes the training dataset for the classifier in order to update the dictionary. The dictionary is also expanded at this point by adding nodes to the network. The process then repeats by testing the newly updated dictionary on a new set of document, which later becomes the training set, etc. We will call each generation of an ROC curve a "time-step", and we will designate it with the variable $1 \leq t \leq 8$ (since the TDT2 dataset contains only enough data for eight time-steps plus an initialization dataset). Significantly, in some time-steps no documents associated with novel classes are introduced to the algorithm, so an ROC curve is not generated.

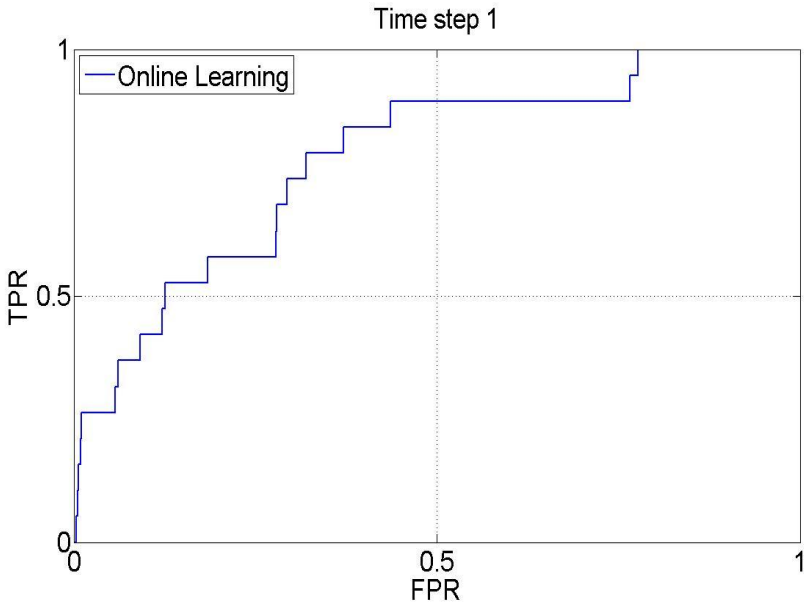

Figure 3. Time-step 1 for samples 1001-2000

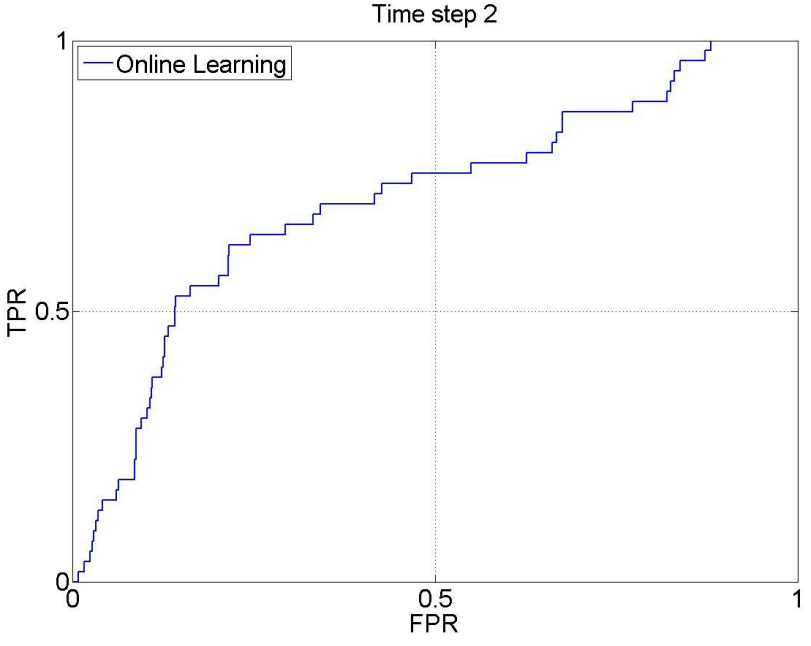

Figure 4. Time-step 2 for samples 2001-3000

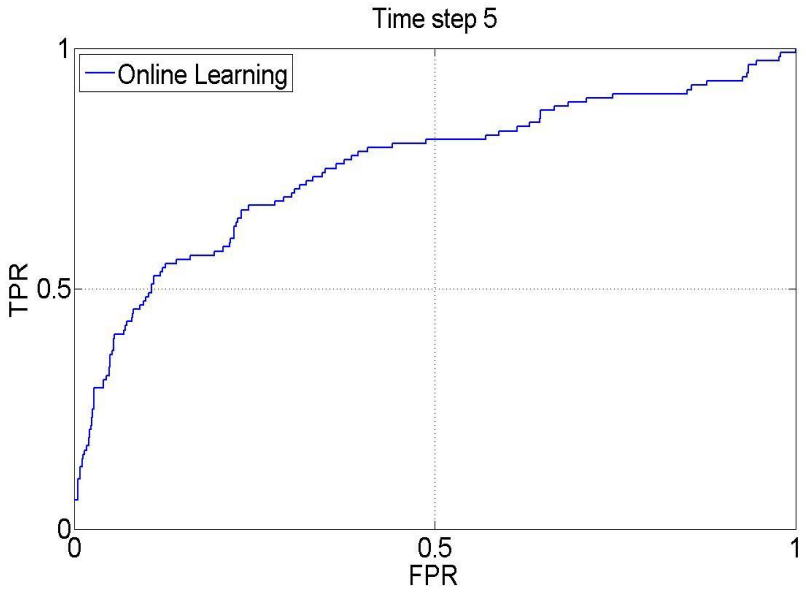

Figure 5. Time-step 5 for samples 5001-6000

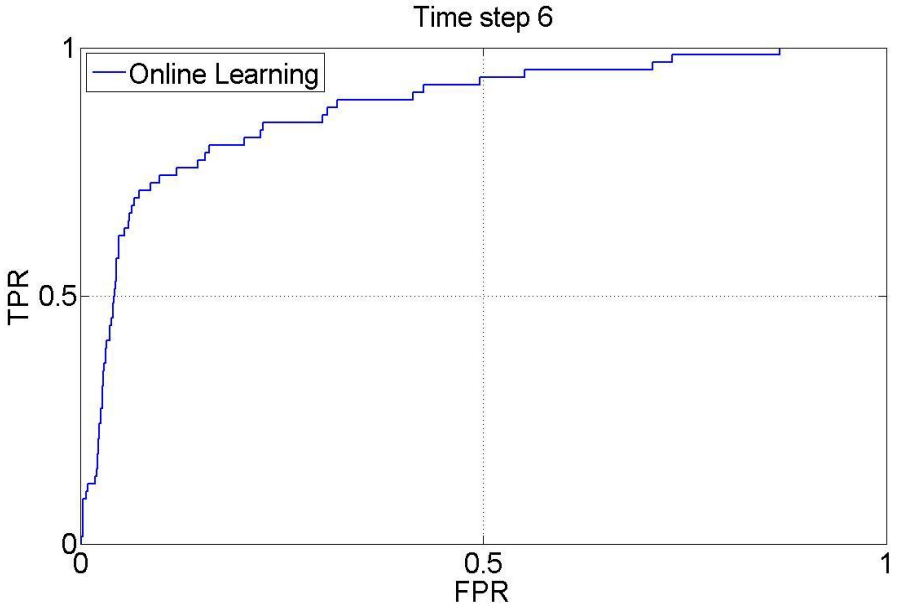

Figure 6. Time-step 6 for samples 6001-7000

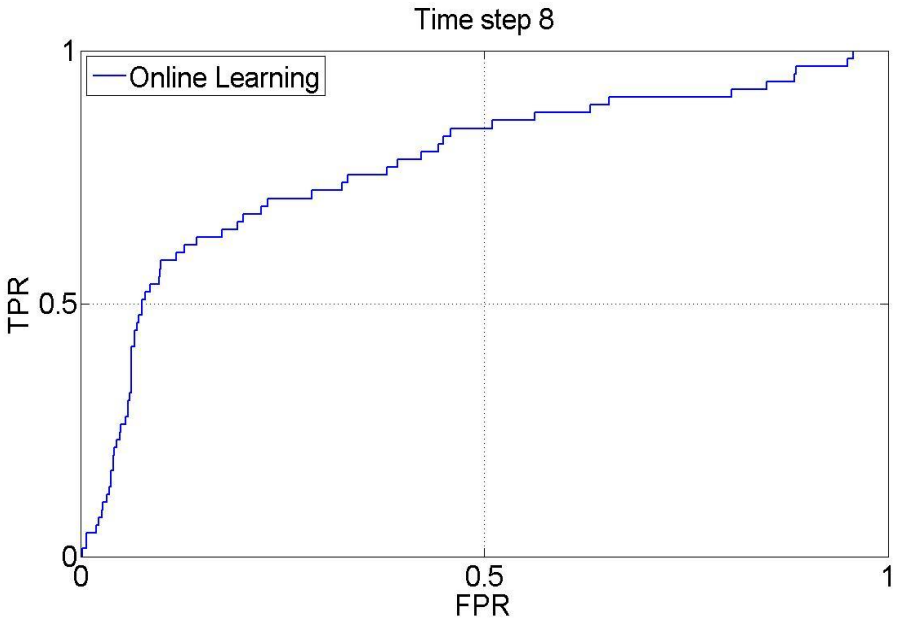

Figure 7. Time-step 8 for samples 8001-9000

These curves represent the ROC associated with each time-step. The x-axis represents the probability of false Positive Rate (FPR), while the y-axis represents the probability of detection or True Positive Rate (TPR). The setups for these simulations are the same as in [4], except that here the dictionary starts with only 100 dictionary atoms and then100 additional atoms is added after each time-step. The area under each curve is listed in Table 1.

Time-step	# Of Novel Docs.	AUC l_1 −Online
1	19	0.793
2	53	0.688
5	116	0.704
6	66	0.881
8	65	0.773

Table 1. AUC Numbers for ROC Plots

Chapter 5

Conclusion

In this project, the dictionary learning problem was studied to expose the key areas where research concerning soft computing methods is being undertaken. Quantifying the degree of success associated with the different research approaches was also attempted. This project demonstrated the stochastic online algorithm for learning dictionaries adapted to sparse coding tasks and its convergence was proven. Experiments proved that this proposed algorithm is significantly faster than batch alternatives on large datasets that may contain millions of training examples.

One can imagine numerous directions for future work based on the proposed framework. While the current work uses dictionaries of a fixed size, it may be more desirable in certain applications to use adaptive dictionaries, whose size changes based on the set of active topics. Furthermore, from an optimization perspective, one may be able to use accelerated gradient descent and related proximal methods to speed up the alternating directions method. Similar ideas can also be developed for other domains, such as healthcare, climate sciences, where detection of novel signal streams is of interest.

Finally, further research is needed to develop generic guidelines for a variety of different data and types of problems, which are commonly faced by financial markets and new researchers in the area of document detection.

REFERENCES

[1] Pang, Bo, Lee, Lillian, and Vaithyanathan, Shivakumar. “Thumbs up? Sentiment classification using Machine Learning techniques.” In *Proceedings of Conference on Empirical Methods in NLP*, 79-86. 2002.

[2] A. Saha and V. Sindhwani. “Learning Evolving and Emerging Topics in Social Media: A Dynamic NMF Approach with Temporal Regularization.” In *WSDM*, pages 693–702, 2012.

[3] Brett Drury, LuísTorgo, and José João Almeida. “Classifying news stories with a constrained learning strategy to estimate the direction of a market index.” .*IJCSA*, 9(1):1–22, 2012.

[4] S. P. Kasiviswanathan, H. Wang, A. Banerjee, and P. Melville, “Online L1-Dictionary Learning with Application to Novel Document Detection,” in *NIPS*, 2012.

[5] S. Shalev-Shwartz. “Online Learning and Online Convex Optimization.” *Foundations and Trends in Machine Learning*, 4(2), 2012.

[6] K. Mizumoto, H. Yanagimoto, and M. Yoshioka, “Sentiment analysis of stock market news with semi-supervised learning,” in *Proceedings Of The 2012 IEEE/ACIS 11th International Conference On Computer And Information Science*. IEEE Computer Society, 2012, pp. 325–328.

[7] M. Aharon, M. Elad, and A. M. Bruckstein. “The K-SVD: An algorithm for designing of over complete dictionaries for sparse representations.” *IEEE Trans. SP*, 54(11):4311–4322, November 2006.

[8] B. A. Olshausen and D. J. Field. “Sparse coding with an overcomplete basis set: A strategy employed by v1?.” *Vision Research*, 37:3311–3325, 1997.

[9] R. Raina, A. Battle, H. Lee, B. Packer, and A. Y. Ng. “Self-taught learning: transfer learning from unlabeled data.” *In ICML*, 2007.

[10] E. Amaldi and V. Kann. “On the Approximation of Minimizing Nonzero Variables or Unsatisfied Relations in Linear Systems.” *Elsevier TCS*, 209(1-2):237–260, 1998.

[11] E. Candes and P. Randall. “Highly Robust Error Correction by Convex Programming.” *IEEE TIT*, 54(7):2829–2840, 2008.

[12] D. Donoho. “For most Large Underdetermined Systems of Equations, the Minimal L1-norm Near-solution Approximates the Sparsest Near-solution.” *Communications on Pure and Applied Mathematics*, 59(7):907–934, 2006.

[13] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. “Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers.” *Foundations and Trends in Machine Learning*, 2011.

[14] J. Wright, A. Yang, A. Ganesh, S. Sastry, and Y. Ma. “Robust Face Recognition via Sparse Representation.” *IEEE TPAMI*, 2008.

[15] A. Y. Yang, S. S. Sastry, A. Ganesh, and Y. Ma. “Fast L1-minimization Algorithms and an Application in Robust Face Recognition: A Review.” *In ICIP*, 2010.

[16] J. Yang and Y. Zhang. “Alternating Direction Algorithms for L1-Problems in Compressive Sensing.” .*Arxiv*, 2009.

[17] S. Shalev-Shwartz. “Online Learning and Online Convex Optimization.” *Foundations and Trends in Machine Learning*, 4(2), 2012.

[18] P. Tseng. “Approximation Accuracy, Gradient Methods, and Error Bound for Structured Convex Optimization.” *Mathematical Programming, Series B*, 125:263–295, 2010.

[19] J. Yang and Y. Zhang. “Alternating Direction Algorithms for L1-Problems in Compressive Sensing.” *SIAM Journal of Scientific Computing*, 33(1):250–278, 2011.

[20] V. Chenthamarakshan, P. Melville, V. Sindhwani, and R. D. Lawrence. “Concept Labeling: Building Text Classifiers with Minimal Supervision.” *In IJCAI,* pages 1225–1230, 2011.

[21] A. W. Lo and A. C. MacKinlay, “Stock market prices do not follow random walks: Evidence from asimple specification test.,” *Review of Financial Studies,* vol. 1(1), no. http://press.princeton.edu/books/lo/chapt2.pdf 2012, pp. 41-66, 1988.

[22] E. Guresen, G. Kayakutlu and T. U. Daim, “Using artificial neural network models in stock marketindex prediction,” *Expert Systems with Applications,* vol. 38, no. 8, p. 10389–10397, August 2011.

[23] K. -J. Kim, “Artificial Neural Networks With Feature Transformation Based on Domain KnowledgeFor the Prediction of Stock Index Futures,” *Intelligent System in Accounting Management, Financeand Management,* vol. 12, p. 167–176), 2004.

[24] T. B. Fakhreldin and M. Fakhreldin, "Prediction of Stock Market Indices using Hybrid GeneticAlgorithm/ Particle Swarm Optimization with Perturbation Term," in *International Conference onswarm intelligence*, http://icsi11.eisti.fr/papers/paper_30.pdf 2013, 2011.

[25] N. L. Suanu, G. Kabari and P. Asagba, "Nigerian Stock Market Investment using a Fuzzy Strategy,"*Journal of Information Engineering and Applications*, vol. 2(8), 2012.

[26] P. Ou and H. Wang, "Prediction of Stock Market Index Movement by Ten Data MiningTechniques," *Modern Applied Science*, vol. 3(12),http://ccsenet.org/journal/index.php/mas/article/viewFile/4586/3925%20rel=%5C%27nofollo%5C%27, December 2009.

[27] L. Mitra and G. Mitra, "Applications of news analytics in finance: A review," *CARISMA* (Centre forAnalysis of Risk and Optimisation Modelling Applications), http://optirisksystems. com/papers/Opt0014.pdf, June 17, 2010.

[28] D. Leinweber, *Nerds on Wall Street*, New Jersey: John Wiley, 2009.

[29] "1998 TDT-2 Evaluation Specification Version 3.7" http://www.nist.gov/speech/tdt98/tdt98.htm

[30] Fiscus, Jon, et al. "NIST's 1998 Topic Detection and Tracking evaluation (TDT2)." *Proceedings of the 1999 DARPA Broadcast News Workshop*. 1999.

[31] C. Manning, P. Raghavan, and H. Schutze. *Introduction to Information Retrieval*. Cambridge UniversityPress, 2008.

[32] J. Friedman, T. Hastie, H. Hfling, and R. Tibshirani. “Path wise Coordinate Optimization.” *The Annals of Applied Statistics*, 1(2):302–332, 2007.

[33] P. O. Hoyer. “Non-Negative Sparse Coding.” *In IEEE Workshop on Neural Networks for Signal Processing*, pages 557–565, 2002.