Applied Finance Project

PREDICTING MARKET REACTIONS TO BAD NEWS

March 2018

Advisor

Dr. David Sraer

Professor at Haas School of Business

Dr. Eric Reiner

Professor at Haas School of Business

Submitted by

Liangliang Chen

Hang Sun Kim

Xin Xin

Xiaowen Yu

This page is left blank intentionally.

Abstract

Our Applied Finance Project aims to develop a framework to predict short-term and medium-term market reactions to bad news shocks. The study is based on a sample of 18,497 bad news articles and time series of 1,008 Russell 3000 stocks returns during the period 2005 to 2017. Our research proposes a three-stage model for the analysis. Firstly, given a dataset of bad news events and stock prices, we employ time series clustering techniques on cumulative abnormal returns of stocks, by which the news articles related to those stocks are grouped into different clusters. Secondly, we apply Natural Language Processing and multi-class classification algorithms on relevant news articles to extract features of each cluster. Then, by applying Support Vector Machine model, whenever specific bad news is released, we can predict the subsequent short-term, and medium-term market reactions post negative news. Finally, we develop long/short trading strategy for both short-term and medium-term horizons that asset managers in the real world can apply every day.

This page is left blank intentionally.

Contents

1. Introduction

From the efficient market hypothesis proposed by Fama (1991), new information is absorbed by the market instantly and entirely, hence the news is fully reflected in the asset prices within at most a day or two. A vast number of studies focus on the short-term effect of news shocks. Recently, more and more researchers show that the impact on the market does not typically subdue within such a short time frame after the news is released (Werner and Murray, 2006). One argument is that investors usually overreact to negative environmental and social story, resulting in a short-term downward movement and a long-term reversal (Lansilahti, 2012).

But is this general for all adverse news events? We believe that different types of adverse news events cause different market reaction patterns. We are interested in short-term and medium-term market reactions to bad news since in long-term the impact of news on the market is very subtle and the long-term market movement is more likely affected by other factors besides the news. Then the resulting question is how to differentiate these news events causing distinct market reactions, for both short-term and medium-term periods.

To answer this question, we propose a method combining time series clustering on stock returns after bad news is released and Natural Language Processing(NLP) techniques on news contents. In contrast to most of the current literature, which tests market reactions to a particular group of news, we start with the stock returns data and then analyze the news contents. More

specifically, for short-term market reactions to bad news, we label the next day abnormal return after news as two states, i.e., upward and downward. For medium-term market reactions, we employ time series clustering (Aghabozorgi et al., 2015) on cumulative abnormal returns of stocks for the subsequent 63 trading days after the news events. This clustering method will give several clusters and categorize them by distinct medium-term market reaction patterns to the news shocks. We then eventually label each news story with the cluster number.

For both short-term and medium-term periods, we conduct NLP analysis on news articles and generate the features of each cluster. With the features of each cluster given, for a news update, by producing its features and classifying it into one of the labels or clusters we obtained in the previous step, we can predict the subsequent market reaction pattern to this news update. Finally, we develop short-term and medium-term trading strategies based on our prediction of subsequent market reactions.

1.1 Objectives

This project aims to solve the following questions:

(1) Does bad news entail same market reaction patterns?

(2) If not, what are the patterns of market reactions after bad news is released, and which type of news content causes which pattern?

(3) How to predict market reactions to bad news update for both short-term and medium-term horizons?

(4) How to generate a trading strategy based on our research?

1.2 Innovation

Most of the previous relevant research uses NLP to pre-define the topics of news articles and then checks the subsequent market reactions for different topics or groups of news. Our study takes the opposite direction. We think pre-defining the topics would lose information. We do not look at the contents of news first, but instead, we begin by analyzing the market reactions and clustering the news based on the subsequent stock returns. Then we examine the contents of news for each cluster and try to find the features of each cluster.

To the best of our knowledge, very few researchers have approached the problem from this perspective, and that is why we think that our research idea is creative and original. Liu et al. (2014)'s research is most relevant to our project. They also start with time series clustering on returns data and then apply NLP on different clusters of news. However, they just generate features for each cluster and neither predict market reactions for a news update nor develop trading strategies. We will extend their research in both directions.

1.3 Hypothesis

ESG refers to the environmental, social and governance factors in measuring the sustainability of an investment in a company[1]. More and more research shows that integrating ESG factors into investment analysis can provide investors with a long-term positive return. Like ESG factors, adverse ESG incidents also have a relatively long-term impact on the market. In this project, we use ESG incidents as bad news so that the news in our dataset will have a relatively long-term impact on the mark, which led to our decision to set the time horizon to be short-term and mid-term in our research.

Many studies examine stock returns after either specific news events or comprehensive news sets. Some present evidence supporting an overreaction hypothesis while some studies support an underreaction hypothesis. There are also studies that find no significant return patterns at all. One possible reason for these mixed and contradictory results is that researchers use comprehensive news set as a single dataset or rely on specific topics of news. While it is possible that news events in the dataset or with the same pre-defined topics cause different market behaviors. Thus, we are led to propose the hypothesis of our research as below.

Hypothesis: Negative news with different contents will cause different patterns of market reaction.

[1] From Wikipedia, https://en.wikipedia.org/wiki/Environmental,_social_and_corporate_governance

1.4 Research Framework

As shown in Figure 1, our research framework is integrated and systematic.

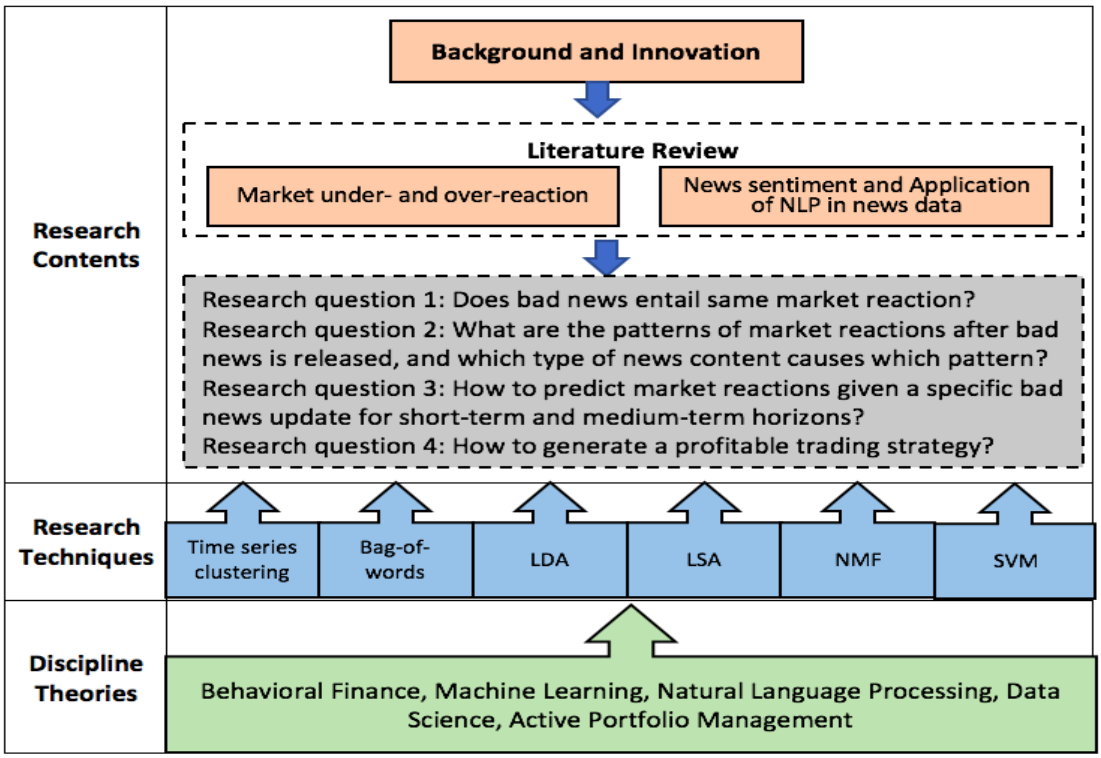


Figure 1. Research Framework

The theoretical foundations of our project are built from multiple disciplines, including Behavioral Finance, Machine Learning, Natural Language Process, Data Science and Active Portfolio Management. We apply various techniques and models from these disciplines to solve the research questions and achieve the research goals of our project.

2. Literature Review

There are two streams of relevant literature, i.e., one is market under- and over-reaction after the news, the other is the Natural Language Processing (NLP) for stock market prediction.

2.1 Market under- and over-reaction to news

2.1.1 Theories

Researchers hold different views on explaining market reactions of stock prices to the public news. We review the three main theories one by one.

The most popular theory argues that investors' behavioral bias is the main driving force of market reactions to the news. For example, Daniel et al. (1998) claim that investors tend to be self-assured about their newly acquired secret information, so they over-react when their information is consistent with the news and ignore the news information when their private information is different from the news. Barberies et al. (1998) explain this phenomenon by conservatism bias and representative bias. The former leads to investors' under-reaction while the latter leads to investors' over-reaction.

The second theory advocates that interactions among investors lead to distinct types of reactions. Hong and Stein (1999) argue that asymmetric information among investors leads to underreaction while the interaction between myopic traders and trend-based traders causes overreaction.

Different from the above two theories, the third theory asserts that there is no under-reaction and over-reaction. Fama (1998) claims that the changes in stock prices would not be obvious if people properly take risk factors into account. This theory proposes that the market is efficient and that deviations from expected stock price are caused by misspecification of the asset pricing models.

2.1.2 Empirical studies

Some studies examine the stock returns after either specific news event or comprehensive news sets to test the theoretical models on market reactions to the news. Early research focused on a particular news event studies to examine the extent to which investors overreacted/ underreacted (Brown and Harlow, 1998). Later, researchers began to take advantage of big data in the news, and test market reactions to a large set of news (Werner and Murray, 2006; Naderi and Mekanik, 2012; Sinha, 2016). Recently, researchers have applied NLP techniques to group news according to their topics or features, and then test market reactions for each group of news (Xie et al., 2013; Ding et al. 2015; Chew et al., 2017).

Among the published empirical studies, many studies present evidence supporting an overreaction hypothesis. For example, Werner and Murray (2006) use a dataset of 250,000 corporate news stories published on Wall Street Journal over 30 years from 1973 and find that pre-event and post-event abnormal

returns have the opposite sign, which proves the existence of market overreaction. Tetlock (2010) finds reversals after news shocks when he studies 10-day post-shock return patterns using a comprehensive news set. Manela (2014) finds over-reaction to drug approvals with high media exposure. However, some studies support an underreaction hypothesis. For example, Sinha (2016) constructs a measure of information that predicts returns over the next 13 weeks and identifies the existence of market underreaction. There also exist other studies that do not identify any significant return patterns. For example, Naderi and Mekanik (2012) find that investors in Tehran Stock exchange (TSE) did not over-react to information in the short term over the period 2005-2011.

These mixed and contradictory results, obtained over more than two decades in multiple markets, are puzzling and unsatisfactory. One important reason may be that researchers either use comprehensive news sets as a single dataset or group the news events based on their topics. The patterns of market reactions can be different for different sets or within a pre-defined group. Therefore, it might be improper to define the group of news before examining the market reaction patterns. We believe that pre-defining topics or groups of news may put news articles predicting different market reactions into one group, which will lead to information loss.

Therefore, unlike the state-of-the-art research using NLP to cluster news content before checking market returns (Herz et al., 2003; Xie et al., 2013; Ding

et al., 2015; Chew et al., 2017), our project proposes to examine the market reactions (or stock returns) first, cluster the news based on the subsequent market reaction patterns, instead of its contents, and then conduct NLP analysis on the news contents to generate the features of each cluster.

2.2 NLP for stock market prediction

With the development of computer technology and the capability to handle massive databases, it is now more feasible to apply more complex machine learning techniques in analyzing financial text. Natural Language Processing has been applied to predict stock price movements. Researchers use multiple NLP techniques to generate features for financial reports and news articles. We review this stream of relevant literature from the perspective of NLP methods, which will be used in our feature generation step.

2.2.1 Bag-of-words

Bag-of-words (BOW) document representation is one of the most used techniques. For example, Luss and d'Aspremont (2008) apply text classification to model stock price movements on a daily basis and predict the subsequent abnormal returns. Kogan et al. (2009) perform text regression and analyze 10K reports to predict stock return volatility. Ruiz et al. (2012) correlate text in financial reports with stock volume and price.

2.2.2 Latent Dirichlet Allocation

Although the bag-of-words model is known to be the most widely used model for text mining, the model does not take sentence structure and grammar into account. A popular extension of the bag-of-words model is Latent Dirichlet Allocation(LDA), which is a Bayesian probabilistic model for text corpus (Blei et al., 2003). LDA can handle disaggregate time periods with sparse data, and thus it is very efficient. Many researchers have developed the algorithm of LDA and applied it in different areas. For example, Blei et al. (2003) describe LDA as a multiple-level hierarchical Bayesian model, and each item of a set of text is a limited variety of underlying topic list. Hoffman et al. (2003) develop an online variational Bayes model for LDA and provide evidence that online LDA is effective in finding topic models. Teh et al. (2007) develop a collapsed variational Bayesian inference algorithm for LDA. Tirunillai and Tellis (2014) propose a unified framework using unsupervised LDA to deal with a large dataset on product reviews across 15 firms over four years.

2.2.3 Latent Semantic Analysis

Latent Semantic Analysis(LSA) is a research technique for analyzing documents to find the underlying meaning or concepts of those documents. It was patented in 1989 by Deerwester et al. (1989). LSA is today widely applicable to many different areas such as retrieving search results, grouping documents into clusters and patent searches. LSA was first used in information

retrieval. LSA solves the problem of synonymy and polysemy in information retrieval by finding the semantic relations between words (Deerwester et al., 1990; Landauer et al., 1998). Landauer and Dumais(1997) show how LSA can replicate how humans acquire knowledge. Most recently, Gálvez et al. (2017) provide a deep insight on how to efficiently mine on-line text data for new information to predict stock returns using LSA.

2.2.4 Non-negative Matrix Factorization

Non-negative Matrix Factorization(NMF), or non-negative matrix approximation, is an approximation technique that decomposes a matrix $\boldsymbol{V}$ into two non-negative factors $\boldsymbol{W}$ and $\boldsymbol{H}$. NMF is widely used today in various research areas like computer vision, audio signal processing, recommender systems and document clustering. Nielsen et al. (2005) study the primary functions of the posterior cingulate cortex (PCC), a part of the brain, by using hierarchical NMF on abstracts downloaded from PubMed. PubMed is a free search engine that accesses the MEDLINE (Medical Literature Analysis and Retrieval System Online) database of abstracts on life sciences and biomedical topics. Berry et al. (2005) apply NMF to cluster parts of the publicly released Enron electronic email collection (Cohen, William, 2005) into 50 clusters to extract the semantic features from the automated email messages.

3. Data

Our research is based on two sets of data: bad news articles and daily stock prices. The below sections describe the two datasets in detail.

3.1 Bad news articles

Bad news articles were scraped from the RepRisk platform. RepRisk daily screens about 80,000 public sources and external stakeholders, including international and local print and online media, news websites, newsletters, NGOs, governmental bodies, think tanks, blogs, and Twitter. This screening is done to identify companies and projects that have been linked to ESG-related risk event or incidents. RepRisk summarizes the company name, relevant news title, abstract, release date and original news link for each incident it identifies. We utilized *selenium* package in Python and ChromeDriver[2] to scrape all the information on RepRisk. Since RepRisk doesn't contain news articles, we employed *newspaper* package in Python to extract the full articles using the original news links. From RepRisk, we collected 18,497 news articles on story-level or 100,328 on company-level for global companies. The release dates of news range from 2005 to 2017.

[2] ChromeDriver is an open source tool that provides capabilities for navigating to web pages, JavaScript execution and user input across Chrome, developed by members of the Chromium and WebDriver teams. The latest version of ChromeDriver can be find at https://sites.google.com/a/chromium.org/chromedriver/.

3.2 Stock Prices

3.2.1. Stocks Universe

The stock prices were obtained from Bloomberg. We downloaded the daily adjusted close prices, taking dividends and stock splits into account, of the Russell 3000 stocks from the year 2005 to 2017. Before combining the stock prices data with the news data, we first tried to add stock tickers to the news dataset by matching company names using *Yahoo Ticker Symbols* table[3]. We then matched the stock tickers of the Russell 3000 with the labeled tickers of companies that appeared in the news dataset and listed the stocks that have a match. The total number of stocks with such match turned out to be 1008. For these 1008 companies, we had 8020 news articles in total.

3.2.2. Key Computations

With the 1008 stock prices, we calculated the daily returns for each stock, which were used as the main starting point of our analysis. Since the stock price movement could be affected by the entire market, we calculated the abnormal returns based on CAPM (Capital Asset Pricing Model). In the following part, we describe several key computations for our analysis.

1. Abnormal Returns

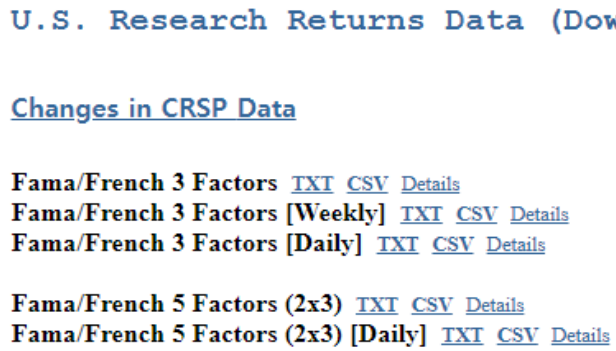
$$AR_{it} = r_{it} - \left(r_{ft} + \beta_{it}\left(r_{mt} - r_{ft}\right)\right)$$

[3] Downloaded from http://investexcel.net/.

- r_{it} : $stock\ i's\ return\ at\ date\ (t)$
- r_{ft} : $risk-free\ rate\ return\ at\ date\ (t)$
- r_{mt} : $market\ return\ at\ date\ (t)$
- β_i : $stock\ i's\ beta\ coefficient$

2. Risk-free returns

We use 1-month T-bill rate as the risk-free rate. Fama/French 3 Factors[Daily] data from the Ken French's website provides the daily risk-free rate returns, which are under the *Rf* column in the downloaded csv file.


U.S. Research Returns Data (Downloadable Files)

Changes in CRSP Data

Fama/French 3 Factors TXT CSV Details
Fama/French 3 Factors [Weekly] TXT CSV Details
Fama/French 3 Factors [Daily] TXT CSV Details

Fama/French 5 Factors (2x3) TXT CSV Details
Fama/French 5 Factors (2x3) [Daily] TXT CSV Details

Source: http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html

3. Market Returns

S&P 500 index is considered as the market average. The daily prices for S&P500 index were downloaded from Yahoo Finance, and daily returns were calculated from the prices.

4. Beta

Beta of $stock\ i$ is calculated on a rolling window of previous 30 trading days using the following OLS regression.

$$r_{it} = \beta_{it}\, r_{mt} + \varepsilon_{it}$$

$$\hat{\beta}_{it} = \frac{Cov(r_{it}, r_{mt})}{Var(r_{mt})}$$

For a stock with news update at time *t*, we calculate its beta at date *t* by regressing the daily stock excess returns on excess market returns over the previous 30 trading days. The coefficient from the above regression is the rolling window based beta.

4. Methodology

Our research can be divided into three main stages: Time Series Clustering, Text Mining and Model Prediction. This section will describe each stage of our research in detail, as shown in Figure 4.

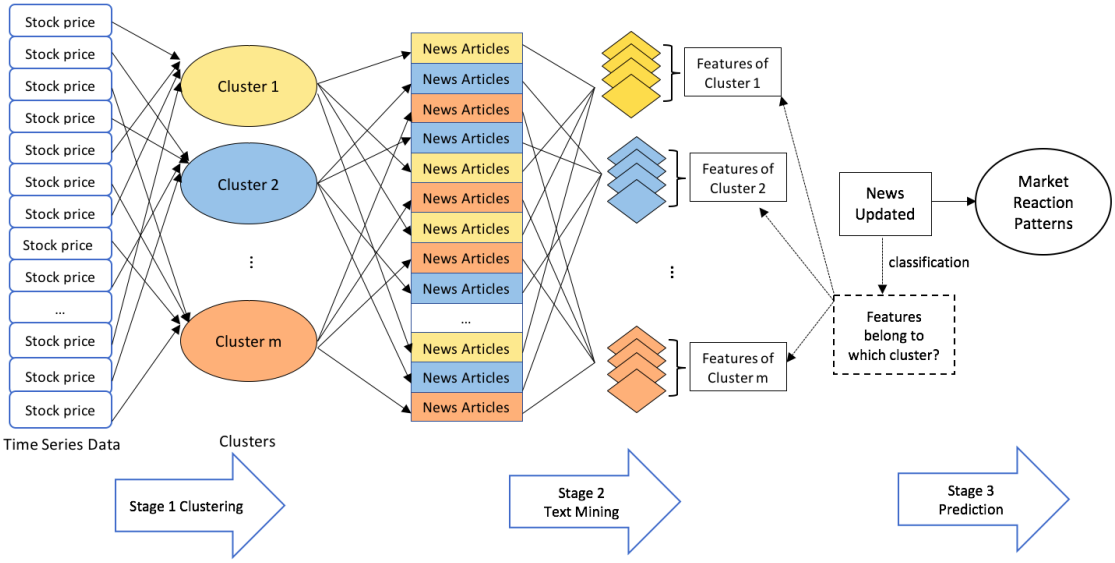


Figure 4. Research steps

Stage 1. Stock returns clustering

We implement the clustering technique into two fields: short-term market reaction and medium-term market reaction. For short-term horizon, we first calculate the abnormal return at date (t+1) for each stock after the news is released at date (t), then label the stock as "upward" if its abnormal return is larger than 1.02, and as "downward" if its abnormal return is smaller than 0.98[4]. For medium-term horizon, we apply time-series clustering on the cumulative abnormal returns for each stock over the subsequent 63 trading days after the

[4] Here we define the thresholds as 1.02 and 0.98, instead of 1, because in this way, we can better differentiate upward and downward movements. It is easy to understand that 1.01 and 0.99 are very close but will be clustered into two groups if we use 1 as threshold.

news is released at date (t). From the time series clustering, we get clusters of different market reactions, e.g., reversal and drift, and the stocks are each labeled as different behaviors. For both short-term and medium-term horizons, the labels we obtain from stock returns are used in feature generation as shown in Stage 2.

Stage 2. Text mining and Feature generalization

We employ three NLP techniques, i.e., Latent Dirichlet Allocation, Latent Semantic Analysis and Non-negative Matrix Factorization, on the news contents to generate relevant features for each cluster. We train these models on our dataset based on the labels we have obtained in Stage 1.

Stage 3. Predicting market reactions

Given a news update, we extract its features and match the features to one of the sets of clusters. Technically, we apply Support Vector Machine (SVM) classification model to classify the news into one of the labels, so that we can predict the subsequent market reactions to this news for both short-term and medium-term horizons.

4.1 Time Series Clustering

4.1.1 Overview

Time Series Clustering is a way of dividing a specific time series data into groups in which the time series in a given cluster are similar based on some metric. We implement time series clustering using the *dtwclust* package in R.

First of all, we determine an appropriate similarity metric, and then we employ a clustering algorithm, such as hierarchical clustering and partitional clustering, to find clustering structures.

4.1.2 Similarity Metric: Distance Functions

The crux of time series clustering is to calculate the distance between two time-series. There exist many different distance functions that can serve as the metric in measuring the degree of dissimilarity between different time sequences. One can easily think of using Euclidean distance, but this metric has a clear limitation when working with large datasets. In this project, we employ two widely-used distance measures in time series clustering: Shape-based Distance (SBD) and Dynamic Time Warping (DTW) distance.

4.1.2.1 Dynamic Time Warping (DTW)

DTW is a dynamic algorithm to find the optimum warping path between two time-series under certain restrictions. Figure 4.1.2.1 below shows how DTW works graphically, which is the alignment between two time-series. Here, the very beginning and end points of the two time-series must match, while other intermediate points are warped to find better matches (Alexis Sarda-Espinosa, 2017).

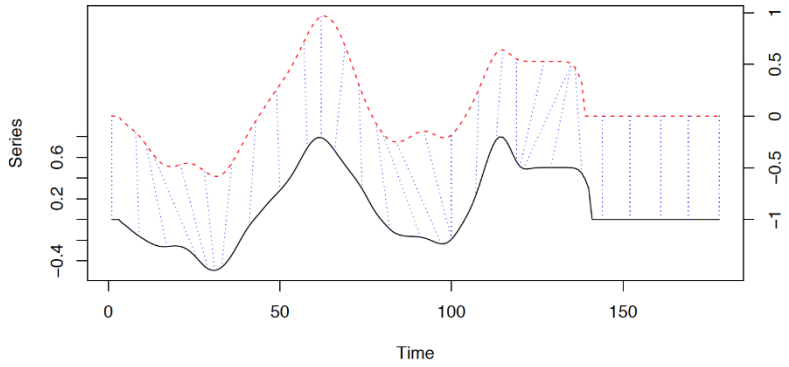


Figure 4.1.2.1 Alignment by DTW between two time-series

Mathematically, for two time-series x_1 and x_2, DTW is defined as

$$DTW_p(x_1, x_2) = \left(\sum \frac{n_\emptyset lcm(k)^p}{N_\emptyset}\right)^{1/p}, \forall k \in \emptyset$$

where $\emptyset = \{(1,1), \dots (n,m)\}$ is the set of all the points that fall on optimum path, lcm is local cost matrix and measured as $lcm(i,j) = (\sum_v |x_{1i}^v - x_{2i}^v|)^{1/p}$, $n_\emptyset$ is a per-step weighting coefficient and $N_\emptyset$ is the corresponding normalization constant (Giorgino 2009).

DTW is used to improve the effectiveness of Euclidean distance, but it is computationally expensive and its calculations are very slow. In our project, DTW is implemented in R using the *dtw_basic* function in *dtwclust* package.

4.1.2.2 Shape-based Distance (SBD)

As part of K-Shape clustering, SBD is developed from the normalized coefficient cross-correlation, or NCCc sequence, between two time-series. SBD is sensitive to scale and is usually used with z-normalization (Paparrizos and Gravano, 2015). The NCCc sequence is calculated by convolving two time-series and does not require point-wise warpings in time.

Mathematically, for two time-series, SBD is defined as

$$SBD(x_1, x_2) = 1 - \frac{\max\left(NCCc(x_1, x_2)\right)}{\|x_1\|_2 \|x_2\|_2}$$

where $\|\cdot\|_2$ is the l_2 norm of the time series. NCCc sequence is obtained by using Fast Fourier Transform (FFT).

SBD is much faster than DTW, and it supports time-series with different lengths directly. We implement SBD in R using the *sbd* function in *dtwclust* package.

4.1.3 Clustering Algorithms

4.1.3.1 Hierarchical clustering

Hierarchical clustering attempts to build a hierarchy of clusters by merging clusters from lower level in the hierarchy (Hastie et al, 2009). We do not need to specify the number of clusters for the hierarchy to be created. Since the procedure of hierarchical clustering is deterministic, it yields the same outcome for a given similarity measure. Two algorithms, i.e., agglomerative and divisive, can be used in hierarchical clustering and agglomerative clustering is more widely used.

For hierarchical clustering, we must calculate the whole distance matrix for the dataset, so the complexity will be huge if the number of objects is large. Therefore, hierarchical clustering is more appropriate for small datasets. Hierarchical clustering is implemented in R using the *hierarchical control* function in *dtwclust* package.

4.1.3.2 Partitional clustering

Different from hierarchical clustering, partitional clustering tries to divide

the dataset into several non-overlapping subsets by putting each object into only one subset. Therefore, we need to specify the number of clusters beforehand. Cluster Validity Indices (CVI) is used to choose the optimal number of clusters. The procedures of partitional clustering are stochastic. K-means and K-medoids are two popular partitional algorithms in practice.

The complexity of partitional clustering is much lower than that of hierarchical clustering. Thus, it is often preferred when we deal with large datasets. Partitional clustering is implemented in R using *norm*, *window.size* and *trace* functions in dtwclust package.

4.2 Text Mining

When working with textual data, we are mainly concerned with finding algorithms that could summarize a large amount of text data into a certain number of topics that we human beings can easily understand and interpret. In this section, we explore different NLP algorithms to extract key features from our clustered news articles and compare their outcome.

4.2.1 Data cleaning and preprocessing

To facilitate the detailed systematic text analysis of the news data, we implement the following steps to clean and preprocess the news data. We first use the regular expression package to identify the garbage characters in the news (for example, OMG, LOL, WOW, etc.) and remove them to eliminate them from further consideration. Certain stop words, such as "the, is, are, that

and which", are frequently used in casual and official expressions, but under most circumstances, they are irrelevant in being associated with sentiment or helping classify other words into sentiment categories. The *nltk* package in Python can download the stop-words set in English and remove those words in addition to numbers, punctuations and special characters. For the preprocessing work, we used the word tokenizer to mark each word in the news and extracted all the nouns, verbs and adjectives that are associated with sentimental information and expected to have predictive power on the asset value or volatility. The following sections detail the different NLP techniques that are used in our research.

4.2.2 Bag-of-words

An important aspect when dealing with categorical data is to convert text or words into numerical values before applying the machine learning algorithm. The bag-of-words technique allows us to represent text as numerical feature vectors.

There are two main ideas behind this model. First, a vocabulary of unique tokens—for example, words—from the entire set of news is created. For example, an array of text data is taken, and by implementing the *fit_transform* method on *CountVectorizer* class, sentences in the text can be changed into sparse feature vectors. Second, the feature vector is constructed from each news that contains the counts of how often each word occurs in a particular news

article.

When processing the text into tokens, there are many different algorithms that are useful. Before applying the most effective algorithm, one can first think of splitting the cleaned news text into individual words at its whitespace characters. A simple example would be that when a text "students at UC Berkeley like reading books and playing soccer" is brought in, a simple code will return a tokenized text in the form of an array ['students', 'at', 'UC', 'Berkeley', 'like', 'reading', 'books', 'and', 'playing', 'soccer'].

In this project, we apply the bag-of-words method to find out the differences in words, n-grams and their frequencies in the news for different clusters.

4.2.3 Latent Dirichlet Allocation (LDA)

LDA is a probabilistic algorithm used to identify and analyze the topics from a corpus of documents under the assumption that each document is generated by a generative process. The overall framework of the generative model is distribution over distribution. The reason why the name has “latent” in it is the details of the topics and statistical distributions are unknown and hence latent. As a starting point, only word counts/frequencies are directly observable. Under the generative assumption, a document is generated with the following steps. A topic is first randomly selected from a list of topics that collectively follow the Dirichlet distribution associated with that document and then a word is randomly selected from a word vocabulary whose component words

collectively follow the Dirichlet distribution associated with that specific topic. The above two steps are repeated until the document has obtained all the words needed.

The LDA algorithm has already been implemented in Python *Scikit Learn* package. Here's a brief description of how it works. The algorithm takes as input a word matrix with word counts as entries and then decomposes that matrix into two smaller matrices, with one governing the topic composition of each document and the other governing the word composition of each topic. Since the algorithm can't automatically determine the number of topics, an initial value will need to be provided to initialize the algorithm. With the two matrices returned by the LDA algorithm, further analysis can be performed to justify the classification or propose another approach otherwise.

4.2.4 Latent Semantic Analysis (LSA)

LSA is a distribution algorithm used in natural language processing. The algorithm is established under the assumption that words with similar meanings tend to appear in similar documents. LSA seeks to reduce the number of words while preserving the difference between the documents. This goal is reached by tokenizing the documents with the term frequency inverse document frequency (TF-IDF) matrix and then performing singular value decomposition (SVD) to reduce the rank of the matrix. The low-rank matrix can then be used to identify similarities or differences between documents. LSA algorithm has been

implemented in Python's Scikit Learn package.

4.2.5 Non-negative Matrix Factorization (NMF)

Similar to other NLP algorithms, like LDA and LSA, NMF is also a rank-reduction algorithm. LDA is a probabilistic algorithm, and it represents the text corpus with latent topics and words following Dirichlet distribution. However, NMF is a deterministic algorithm. The rank-reduction is achieved through factorizing the non-negative TF-IDF matrix **V** into two non-negative matrices **W** and **H**. There are several types of non-negative matrix factorizations, and this arises from using different cost functions for measuring the gap between **V** and **WH**. NMF algorithm is also available in Python's *Scikit Learn* package. The algorithm calculates the two non-negative matrices by minimizing the Frobenius norm under the non-negative constraint.

4.3 Classification and Prediction

After features are generated for each cluster, for a news update, we can analyze its features using the same NLP model and classify it into one of the clusters. This procedure is the same for short-term and medium-term market reactions prediction. Here, we apply a popular supervised learning model, i.e., Support Vector Machine (SVM) to do the classification.

To perform classification, SVM tries to find the hyperplane that maximizes the margin between two categories. There are three steps in the SVM algorithm.

Firstly, we define an optimal hyperplane by maximizing the margin. The margin is the distance between the hyperplane and the training samples that are nearest to this dividing plane, which are called the support vectors. Secondly, we add a penalty term for misclassifications so that the hyperplane is extended for non-linear classification problems. Finally, since classifying with linear decision surfaces is much easier, we reformulate the problem to map data into higher dimensional space implicitly.

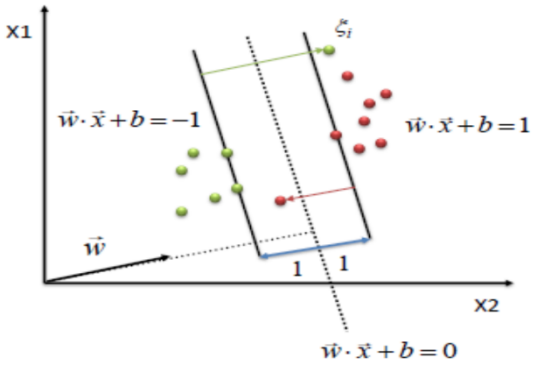


Figure 4.3 SVM Classification

As Figure 4.3 shows, the objective function of SVM penalizes for misclassified instances and those within the margin. With a slack variable, the model allows some instances to fall off the margin, but at the same time penalizes these instances. Mathematically, SVM model is defined as:

$$\min \frac{1}{2}\|w\|^2 + C\sum_i \xi_i$$

$$s.t.\ \ y_i(w \cdot x_i + b) \geq 1 - \xi_i, \forall x_i$$

$$\xi_i \geq 0$$

where ξ_i is the slack variable, C trades-off margin width and

misclassifications.

Support Vector Machine is very effective in high dimensional spaces, even when the number of dimensions is greater than the number of samples. SVM is memory efficient because it uses a subset of points in the objective function. It is also versatile since it can use different Kernel functions in the objective function. We implement SVM classification in Python using *SVC* function in *Scikit Learn* package.

4.4 Trading Strategy

Based on our prediction of short-term and medium-term market reactions to a certain news update, we can build trading strategies for both short-term and medium-term horizons.

4.4.1 Short-term trading strategy

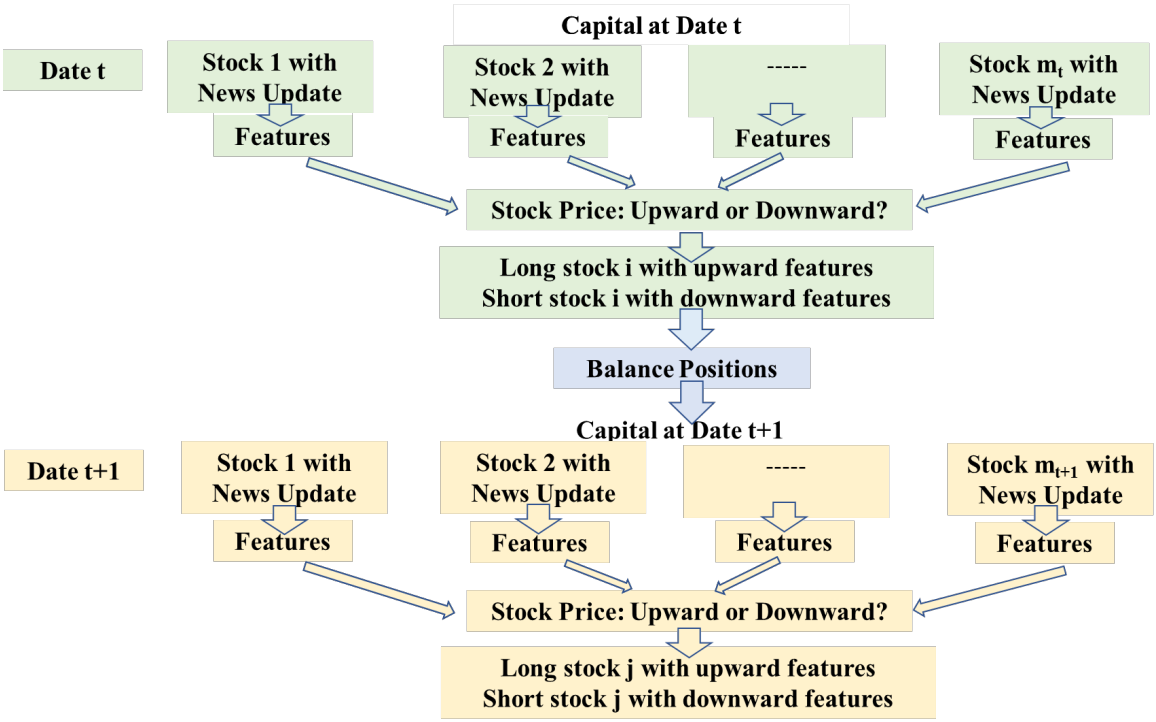


Figure 4.4.1 Short-Term Trading Strategy

Figure 4.4.1 demonstrates the flowchart for our short-term trading strategy. The details are as follows. Given a news update at date *t*, we extract its features and classify it into one of the two labels for short-term period, i.e., upward and downward. If we predict the next day market reaction to this news is upward (or downward) movement, we will long (or short) this stock at date t, and then balance the position the next day. If there are multiple news events on the same day, we assume our capital on date t is evenly distributed to those stocks so that the portfolio includes all the stocks that have a news update. Then we calculate cumulative abnormal returns for our portfolio and mark-to-market daily. We implement this strategy for both the training set (in-sample) and the test set (out-of-sample).

4.4.2 Medium-term trading strategy

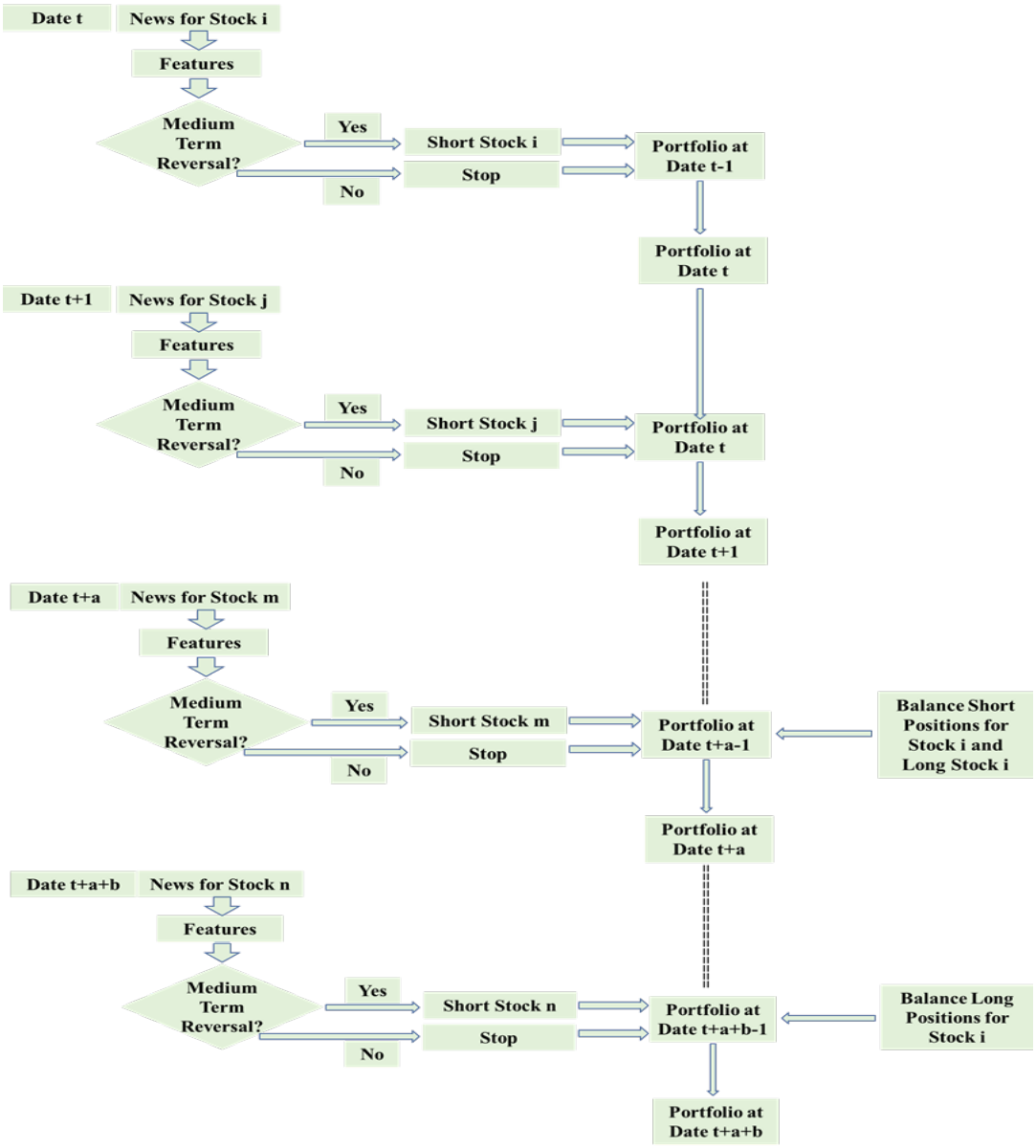


Figure 4.4.2 Mid-Term Trading Strategy

Figure 4.4.2 demonstrates the flowchart of our mid-term trading strategy. For medium-term investment, we aim to find out which news leads to market overreaction and bet on the reversal. Therefore, given a news update at date *t*, we extract its features and classify it into one of the labels we obtain using time series clustering. If the features of this news match to the features of reversal cluster, we will trade on the stock relevant to this news.

From time series clustering, besides the clustering results, we can also plot

the centroid of the cumulative abnormal return time series for each cluster. Then the plot of the centroid in reversal cluster gives us the estimated date when reversal shows up. Therefore, dynamically, our trading strategy is that at date t, we short the stock and hold the position until ($t+a$) days, and then we balance the short position and long the stock until ($t+a+b$) days. Here, a is the number of trading days before the reversal shows up, and b is the number of trading days after the reversal appears and before the upward movement disappears.

Our strategy is based on the assumption that we have enough capital to long or short the stock when there is a relevant news update. Therefore, the portfolio is constructed dynamically, and a new stock is added to the portfolio whenever its relevant news is classified into one reversal cluster. We calculate cumulative abnormal returns and mark-to-market daily. Similar to our short-term strategy, we implement this dynamic medium-term strategy for both the training set (in-sample) and the test set (out-of-sample).

5. Results

5.1 Time Series Clustering

Time series clustering is conducted using cumulative abnormal returns for the stocks over subsequent 63 trading days after the news is released. It is important that we choose the most optimal clustering algorithm, distance function and the number of clusters.

5.1.1 Clustering algorithm and distance function

For clustering algorithm, we employ hierarchical clustering and partitional clustering, and for each algorithm, we use two different distance functions, i.e., Shape-based Distance (SBD) and Dynamic Time Warping (DTW). To select the clustering algorithm, we pre-define the number of cluster to be 10 and run the hierarchical and partitional clustering algorithms. We compare these two algorithms and two distance functions from the cluster sizes and the average intra-cluster distance. In the next step, we will select the optimal number of the cluster using Cluster Validity Indices (CVI).

From table 5.1.1, we can see that the intra-cluster distance is larger for hierarchical clustering than that for the partitional cluster, while the cluster size is more evenly distributed for the partitional cluster. In this project, we prefer partitional clustering algorithm, because it differentiates distinct behaviors much better and evenly distributed cluster sizes are better for training NLP models. As to the distance functions, we can see that Shape-based Distance and Dynamic Time Warping give similar results.

Table 5.1.1 Cluster size with average intra-cluster distance for k=10

	Hierarchical clustering				Partitional clustering			
	DTW		SBD		DTW		SBD	
Cluster	Size	av_dist	Size	av_dist	Size	av_dist	Size	av_dist
1	5540	3.0277	6198	0.4259	544	0.1556	626	0.2012
2	676	3.7429	18	0.3070	485	0.2296	610	0.1689
3	166	3.9824	142	0.4935	243	0.5967	738	0.1439
4	13	3.7558	29	0.5651	762	0.4032	841	0.1424
5	4	4.0601	2	0.1130	703	0.1102	468	0.3096
6	1	0.0000	11	0.3194	850	0.1559	253	0.1504
7	12	4.4678	6	0.3431	177	0.3506	919	0.1876

8	1	0.0000	3	0.6879	1108	0.1605	754	0.1601
9	1	0.0000	2	0.3937	892	0.1316	486	0.2975
10	1	0.0000	4	0.2084	651	0.1297	720	0.1199

5.1.2 Optimal number of clusters

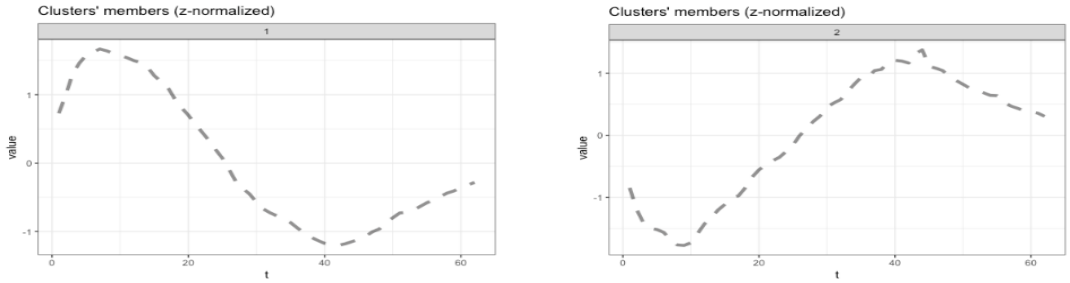
To choose the optimal number of clusters, we use Cluster Validity Indices (CVI), which measure the "goodness" of a clustering result by comparing to the other ones created by other algorithms, or by using different parameter values. Larger value of CVI means a better cluster result. For partitional clustering with SBD distance, the CVI test results for varying number of clusters are shown as follows in Table 5.1.2.

Table 5.1.2 CVI for different number(k) of clusters (Partitional Clustering with SBD)

CVI	k=2	k=3	k=4	k=5	k=6	k=7	k=8	k=9	k=10
Sil	**0.4477**	0.3188	0.3969	0.3097	0.2588	0.2310	0.2539	0.1727	0.1452
SF	**0.4262**	0.3435	0.2709	0.1910	0.1768	0.1674	0.0987	0.1261	0.1104
CH	**6047.9**	4067.3	3213.9	2827.2	2428.2	2233.9	2174.5	1729.7	1484.5
DB	0.7925	0.8837	**0.9314**	0.8570	0.12057	0.1271	0.1085	0.1614	0.1886
DBstar	0.7925	1.0327	**4.1059**	1.4748	2.0726	2.4929	2.1315	3.1800	1.3235
D	0.0012	0.0008	**0.0020**	0.0010	0.0014	0.0007	0.0013	0.0008	0.0008
COP	**0.0469**	0.0405	0.0289	0.0256	0.0263	0.0239	0.0208	0.0229	0.0219

Note: Max value of each row(CVI) across different number of clusters is highlighted as bold.

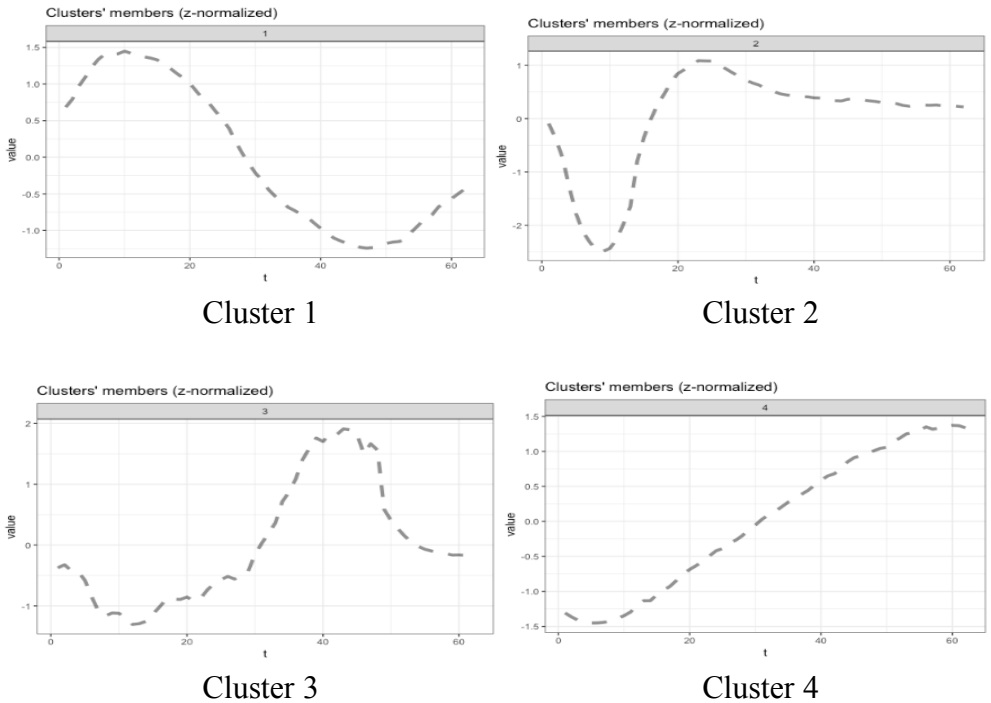
The majority of indices suggest that we use k=2, and then use k=4. We consider two clusters are too limited and may not differentiate distinct behaviors. To further compare the cluster results of k=2 and k=4, we plot the centroid of each cluster for both cases. When the number of clusters is set to be 2, the plots of the centroids of each cluster are shown in Fig. 5.1. We can observe that Cluster 1 shows downward drift for subsequent two months with a slight reversal afterward and that Cluster 2 shows a short-term reversal.



Cluster 1 Cluster 2

Fig. 5.1 Plots of centroids of each cluster for k=2 (Partitional Clustering with SBD)

When the number of clusters is pre-defined as 4, the plots of centroids for each cluster are shown in Fig. 5.2. Cluster 1 and 3 are similar to the clusters obtained when the number of cluster was 2. But the patterns in Cluster 2 and Cluster 4 are new. Cluster 2 here is also a reversal, but the upward trend which appears after reversal is relatively short-lived, and Cluster 4 shows upward drift right after the news date.



Cluster 1 Cluster 2

Cluster 3 Cluster 4

Fig. 5.2 Plots of centroids of each cluster for k=4 (Partitional Clustering with SBD)

Comparing the plots for k=2 and k=4, we can see that pre-defining the number of clustering as two might hide the real reversal trend and misclassify the long-term upward drift as reversal cluster. Thus, in this project, we define the number of clusters to be 4[5].

5.1.3 Clustering results

We employ partitional clustering with Shape-based Distance(SBD) measure to conduct the time series clustering with a number of clusters pre-defined as 4. The cluster size with intra-cluster distance is shown in Table 5.1.3. The total sum of the cluster size equals to 8020, which is the total number of news articles as mentioned in Part 3.1. From Fig. 5.2, we can observe the trend of the centroid for each cluster. Intuitively, cluster 1 is a downward drift, cluster 2 is a short-term reversal, cluster 3 is a medium-term reversal and cluster 4 is an upward drift.

Table 5.1.3 Cluster size with intra-cluster distance for k=4 (Partitional Clustering with SBD)

Cluster	Size	av_dist
1-Downward drift	2992	0.2231
2-Short-term reversal	1483	0.3342
3-Medium-term reversal	1721	0.2463
4-Upward drift	1824	0.1558

From the time series clustering, all the news events are clustered into one of

[5] Actually we tried both cases in our following NLP and SVM analysis, we did find that k=4 works better for our dataset.

the four clusters. The cluster numbers are used as labels in the following NLP analysis.

5.2 NLP Analysis

5.2.1 Bag-of-words

As the first step for NLP analysis, the bag-of-words method is employed to test whether the most frequent words of the news contents in different clusters are distinctive. We first combine all the contents of news in the same cluster, and then tokenize the combined strings. The number of tokens contained in each cluster is shown in Table 5.2.1-1.

Table 5.2.1-1 Number of tokens in each cluster

Cluster	Number of tokens
1-Downward drift	3,192,821
2-Short-term reversal	1,540,870
3-Medium-term reversal	1,607,372
4-Upward drift	1,754,433

Since the verbs used in the negative news article can reflect the emotions of the press to the news and the related company, different emotions delivered will lead to different subsequent market reactions. Thus, we compare the most frequent negative and positive verbs in the news contents for the four clusters in Word Clouds, a text data visualization which allows the user to see at a glance what words in text appear more frequently than others. As shown in Fig. 5.2.1-1, negative verbs for each cluster are displayed. The more a specific word appears in the text, the bigger and bolder it appears in the word cloud.

Cluster 1. Downward Drift

Cluster 2. Short-term Reversal

Cluster 3. Medium-term reversal

Cluster 4. Upward Drift

Fig. 5.2.1 -1 Word Clouds of negative verbs in each cluster

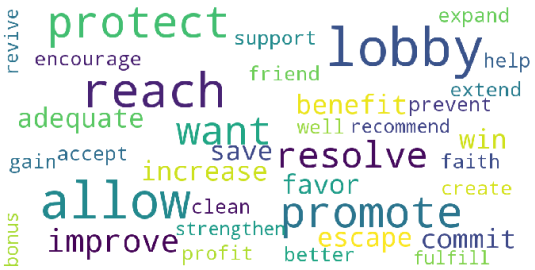
We can see that the most frequent negative verbs in the first three clusters are quite similar, while cluster 4 (Upward Drift) has fewer negative verbs. Intuitively, since the market reactions in the first three clusters are downward movement right after the negative news, while the market reaction in cluster 4 is upward after the news, we conclude that bad news with fewer negative words would predict a subsequent upward movement. Otherwise, the stock price will drop down for a while.

The Word Clouds for positive verbs in each of the clusters are shown below in Fig. 5.2.1 -2. It is reasonable to think that the Word Clouds of positive verbs are sparser than those of negative verbs, since the positive verbs are less frequent in bad news. Interestingly, we observe that positive verbs are much

more frequent in cluster 4 (Upward Drift) than in the other three clusters. Thus, we can conclude that bad news with more positive verbs would predict a subsequent upward movement in stock prices.

Cluster 1. Downward Drift

Cluster 2. Short-term Reversal



Cluster 3. Medium-term reversal

Cluster 4. Upward Drift

Fig. 5.2.1-2 Word Clouds of positive verbs in each cluster

The frequency of appearance shows how many times one word appears in the document. Another dimension to analyze the words used in the news is to look at how many negative or positive words exist in the news contents for each cluster. Therefore, besides using Word Clouds to display the appearance frequency of positive and negative verbs, we also calculate the ratios of negative and positive words to total tokens, as well as the ratio of negative verbs to all verbs, and the ratio of positive verbs to all verbs for each cluster.

Table 5.2.1-2 Ratio of negative and positive words/verbs in each cluster

Cluster	Negative Words / Total Tokens	Positive Words / Total Tokens	Negative Verbs / Total Verbs	Positive Verbs / Total Verbs
1-Downward drift	32.1440%	21.3723%	11.3771%	3.04450%
2-Short-term reversal	32.4099%	21.3123%	11.7695%	3.1257%
3-Medium-term reversal	32.9175%	21.8527%	12.3889%	2.9144%
4-Upward drift	31.6095%	21.7165%	10.6396%	3.5257%

Table 5.2.1-2 shows that the news in Cluster 4 (Upward Drift) has the lowest percentage of negative words and negative verbs but has the highest percentage of positive verbs. This result further demonstrates our finding that bad news articles with fewer negative words and more positive words included are related to subsequent upward movement.

From the above analysis, we examine that bad news with fewer negative words and more positive words tends to predict an upward movement. In such case, it is possible that before the news is released, private new information is leaked to some investors leading to a drop down of the relevant stock prices. The company concurrently starts to take actions to deal with the incidents. When the news is released to the public, even though it is still bad news for the company, the press tends to report the news with a more positive sentiment besides just describing the adverse incidents.

The bag-of-words method is explicitly used to show the difference in news contents for different clusters. In the following part, we will employ NLP models to generate features for each cluster.

5.2.2 Topic generation

We apply three NLP models, i.e., LDA, LSA, and NMF, on news abstracts to generate features for all the clusters. The models are used to extract topics from texts. Firstly, we split the dataset into two exclusive sets, i.e., training set and test set. We fit the NLP models on the training set, and then in the next step, we will evaluate the models using the test set. To avoid using future information, the training set is all data before January 1, 2016, and the test set is the data from January 1, 2016 to end of 2017.

The three NLP models are used to generate topics on the training set. Fig. 5.2.2-1 shows the words in the first topic generated by the three models. From this figure, we can see that the words in the first topic generated by the three models are quite similar. For example, "billion", "federal", "corporate", "reportedly" are all included in the topics generated by the three models. There also exists difference, which is caused by the different algorithms in each model. Fig. 5.2.2-2 and Fig 5.2.2-3 show the words in the second and third topics generated by the three models respectively. We can also find the similarity in words in the second and third topics.

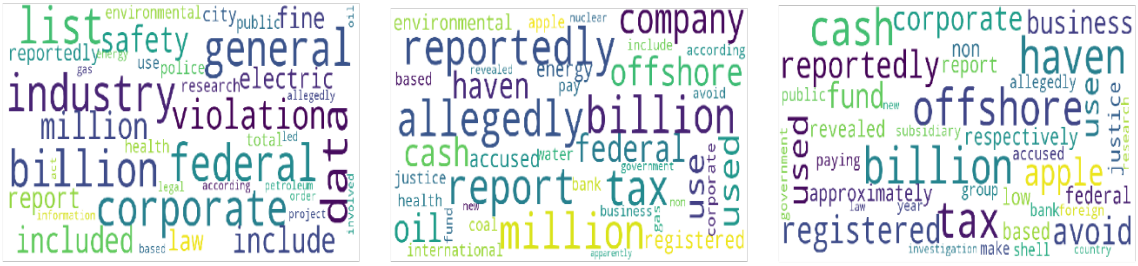


Fig. 5.2.2-1 Words in the first topic generated by LDA, LSA and NMF

Fig.5.2.2-2 Words in the second topic generated by LDA, LSA and NMF

Fig.5.2.2-3 Words in the second topic generated by LDA, LSA and NMF

To generate features, we train the three models to fit the training set. Since the number of topics is a hyper parameter, we tried different numbers and found that ten topics give us better fitting and classification result. The ten topics are generated from the news abstracts for each of the clusters. Fig. 5.2.2-4 shows the distribution of topics within each cluster when LDA is employed. It is obvious that the topics are distributed in a very different way for each cluster, which means the features generated for different clusters are quite different. This result also demonstrates that LDA works pretty good for our dataset.

Cluster 1 (downward drift) is characterized mainly by Topic 10, cluster 2 (short-term reversal) is characterized by Topic 6, cluster 3 (medium-term reversal) is characterized by the combination of Topic 6 and Topic 9, and cluster 4 (upward drift) is characterized by Topic 7.

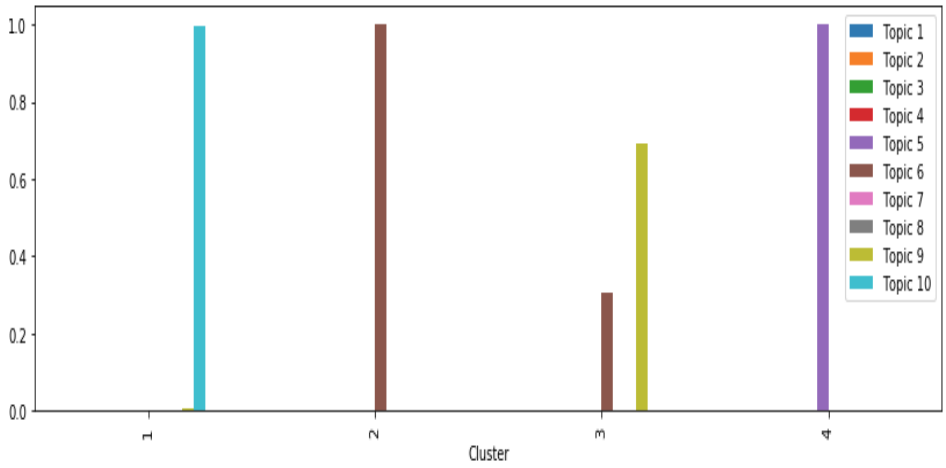
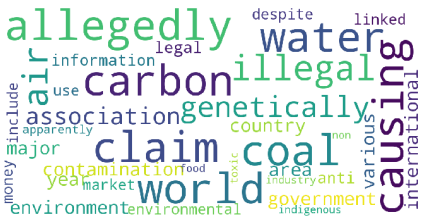


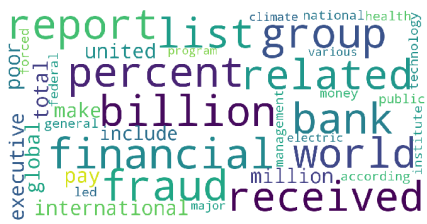
Fig. 5.2.2-4 Topics distribution within each cluster (LDA model)

The words in Topic 10, Topic 6, Topic 9 and Topic 5 are shown in Fig. 5.2.2-5. We can observe that the top words in Topic 10 are “carbon”, “water”, “coal”, “air”, “claim”, which present environmental issues, top words in Topic 6 are “human”, “health”, “government”, “people”, “country”, “supply”, which present society and community issues. Obviously, Topic 9 is financial issue, which includes key words like “billion”, “bank”, “fraud”, “percent”, “financial” et al. Topic 5 is more like governance issue, which is consisted by key words like “union”, “safety”, “administration”, “program”, “safety”, “investigation”.



Topic 10

Topic 6



Topic 9

Topic 5

Fig. 5.2.2-5 Words in Topic 10, 6, 9 and 5 (LDA model)

Therefore, we conclude that environmental issues tend to predict downward drift in stock price, society and community issues predict short-term reversal, governance issues predict upward drift, while the combination of society and finance issues predict a medium-term reversal.

The top 40 words in each topic generated by LDA are shown below in Table 5.2.2. The information on topics generated in the first ten topics from LSA and NMF is summarized in the Appendix.

Table 5.2.2 Top 40 words in each Topic generated by LDA

	Top 40 words in each topic
Topic 1	general federal industry corporate list billion data violation million include safety included report fine electric law reportedly city environmental health total use police research public according petroleum gas energy project based involved oil legal allegedly like act information led order
Topic 2	energy report climate coal change fuel palm public power shell electric institute oil based include corporate allegedly global corp reportedly research according industry clean management number environment river petroleum revealed government union received gas tobacco exchange association chevron subsidiary involved
Topic 3	million department pay agreed settlement justice settle company act billion antitrust

	reportedly agreement commission investigation anti fine agency similar case exchange protection according air government fraud new research allegedly money subsidiary market financial total illegal following received related approximately institute
Topic 4	drug action class price millions company pay legal food investigation illegal antitrust settlement apparently paying competition billion similar cause damages revealed law various wages involved fraud federal oil administration international led association spill district center anti allegedly order facing supply
Topic 5	company safety reportedly administration allegedly health union year program potential following industry work food investigation new failing despite received like number percent commission according management led similar plant accused order genetically make energy pay local said control executive united international
Topic 6	allegedly accused reportedly report use government used company human high health include people country international according apparently group linked illegal police project supply local land public water million law environmental failing industry new state revealed led order major cause despite
Topic 7	commission food international bribery exchange accused foreign technology investigation competition trade market said allegedly corp department reportedly anti law human business health act justice apparently control federal government related public percent paying subsidiary include global center union world according shell
Topic 8	people indigenous land risk chevron oil reportedly environmental report allegedly new total canada face pollution mining toxic include company law waste based group financial according palm research information apparently gas government shell recently protection city united genetically food public following
Topic 9	group financial percent bank coal report fraud billion carbon received world related list international poor total executive pay million united global make include according money general health public management technology institute national led federal forced program electric major various climate
Topic 10	world allegedly reportedly claim causing illegal air genetically association contamination environment major international government area various health year country people information anti legal environmental use linked market despite include money apparently indigenous non water industry food toxic recently united following

5.3 SVM Classification

To evaluate the models fitted using the training set, we apply SVM to

classify the news in the test set into one of the clusters based on their features. This is a procedure of predicting subsequent market reaction patterns for updated news. We conduct SVM classification model for both short-term and medium-term horizons. For short-term horizon, SVM is used to classify the news in the test set to be one of the two labels, i.e., upward and downward. While for medium-term horizon, SVM is used to classify the news to be one of the four labels we obtained through time series clustering.

5.3.1 Accuracy for short-term prediction

For short-term prediction, the accuracy scores on the training set and the test set when features are generated by LDA, LSA and NMF models are shown in Table 5.3.1. The accuracy score is the ratio of correctly classified observation by SVM, compared with the upward/downward labels we made based on next-day stock price movement, to the total observations. It shows that LDA gives the best in-sample classification accuracy, while NMF gives best out-of-sample classification accuracy. The overall accuracy scores are not high, most of which are just slightly over 0.5. We think this is mainly caused by the size of the final dataset for SVM model. However, as we will prove in the trading strategy part, even though the prediction accuracy is not very high, we still can obtain relatively good returns using our short-term trading strategy.

Table 5.3.1 Accuracy scores for short-term prediction

Dataset	LDA	LSA	NMF
Training Set (in-sample)	0.5553	0.5464	0.5296
Test Set (out-of-sample)	0.5031	0.4847	0.6012

5.3.2 Accuracy for medium-term prediction

Table 5.3.2 shows the classification accuracy scores on both the training set and the test set for medium-term horizon. Here, the accuracy score is the ratio of correctly classified observation, compared with the labels we obtained from time series clustering, to the total observations. We can observe that NMF gives the best in-sample and out-of-sample accuracy scores. Compared with the prediction accuracy for the short-term horizon, the prediction accuracy for medium-term horizon is much higher. Thus, in our dataset, SVM classification model works better for medium-term market reaction prediction.

Table 5.3.2 Classification accuracy scores for medium-term prediction

Dataset	LDA	LSA	NMF
Training Set (in-sample)	0.6826	0.6713	0.6873
Test Set (out-of-sample)	0.7392	0.7136	0.7513

From Table 5.3.1 and Table 5.3.2, we observe that among the three NLP models, NMF gives us the best prediction accuracy, thus, in the following

trading strategy part, we will use the features obtained from NMF model.

5.4 Performance of Trading Strategy

Using the labels predicted by SVM classification, we implement the short-term and medium-term strategies for both the training set (in-sample) and the test set (out-of-sample). The detailed information of trading strategies is described in Part 4.4. In the following part, we show the performance of our trading strategies for both short-term and medium-term horizons.

5.4.1 Short-term trading strategy

Based on the daily and cumulative returns of the in-sample analysis displayed in Figure 5.4.1-1 and 5.4.1-2 respectively, we observe that when executed in the training set, the strategy gives us acceptable results. Even though the daily return is not quite stable, the cumulative return grows during 2012 and the mid of 2014, and obviously beats the market. The cumulative return reaches its peak in the mid of 2014, after which the cumulative return drops down until 2015. The flat trend in cumulative return from 2005 to 2010 is caused by the fact that the number of news during this period is relatively small compared to the following periods.

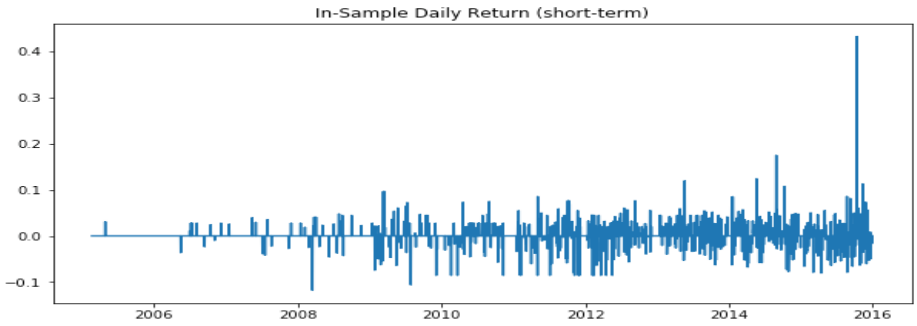


Fig. 5.4.1-1 In-sample daily returns for short-term strategy

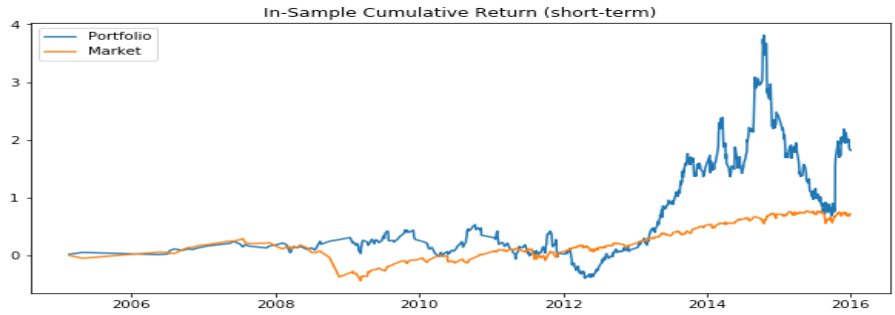


Fig. 5.4.1-2 In-sample cumulative returns for short-term strategy

Table 5.4.1-1 shows the annualized Information Ratio (IR) and its t-statistics. We observe that in most of the years in the training set, i.e., from 2006 to 2015, annualized IR is larger than zero, except in 2009, 2010 and 2015. We can also examine that the IR in 2006, 2007, 2008 and 2013 is relatively good. However, the t-statistics shows only the IR in 2013 is statistically significant. This is mainly caused by the high volatility in daily returns.

Table 5.4.1-1 In-sample annualized IR and t-Stats

Year	IR	t-stats
2006	**2.4710**	1.4009
2007	**0.4686**	0.2657
2008	**0.5871**	0.7026
2009	-0.1578	-0.4075
2010	0.0510	0.1445
2011	-0.0753	-0.3462

2012	0.0351	0.2413
2013	**0.4123**	**2.6233**
2014	0.0681	0.5194
2015	-0.0227	-0.1831

The out-of-sample daily and cumulative results are shown in Figure 5.4.1-3 and 5.4.1-4, respectively. We observe that the daily returns are much more stable compared to training set, and this strategy performs fairly well in the test set. The maximum of cumulative return reaches 1.5 in July 2017.

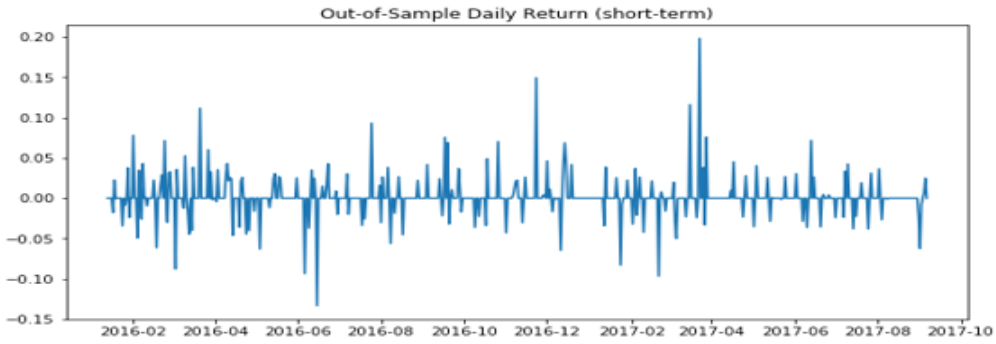


Fig. 5.4.1-3 Out-of-sample daily returns for short-term strategy

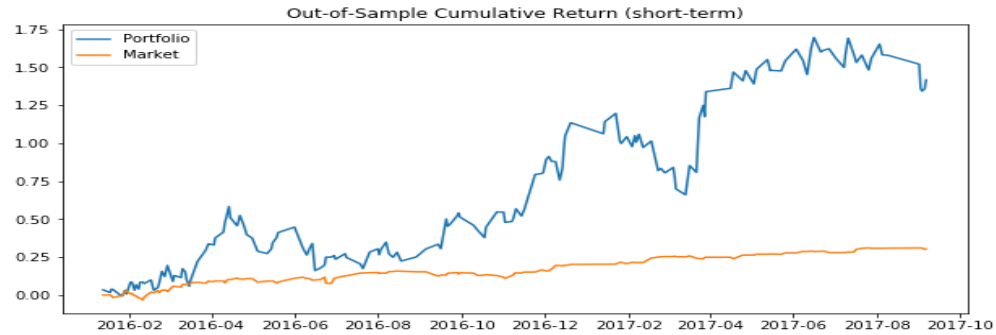


Fig. 5.4.1-4 Out-of-sample cumulative returns for short-term strategy

The annualized IR in the test set, i.e., from 2016 to 2017, is shown in Table 5.4.1-2. We observe that IR is positive for both years, while the t-statistics shows the IR is not statistically significant, which is also caused by the high volatility in daily returns.

Table 5.4.1-2 Out-of-sample annualized IR and T-stats

Year	IR	t-stats
2016	0.2223	1.5961
2017	0.0649	0.2495

5.4.2 Medium-term trading strategy

Medium-term trading strategy is designed to bet on the medium-term reversal. Fig.5.4.2-1 and 5.4.2-2 show the in-sample daily and cumulative returns for this strategy. Fig.5.4.2-3 and 5.4.2-4 show the out-of-sample daily and cumulative returns for this strategy. We observe that even though the cumulative returns of medium-term trading strategy are smaller than those of short-term strategy, the volatility of daily returns is much lower, and the upward trend of cumulative returns of medium-term strategy is better. The maximum of cumulative return for in-sample and out-of-sample is 1.25 and 0.7, respectively.

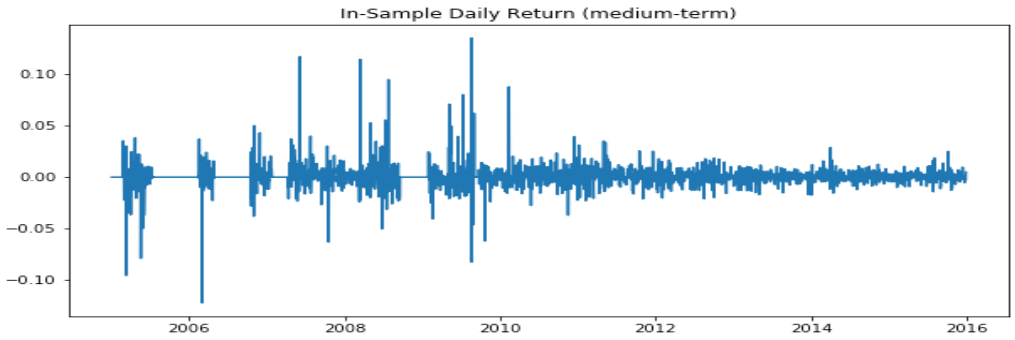


Fig. 5.4.2-1 In-sample daily returns for medium-term strategy

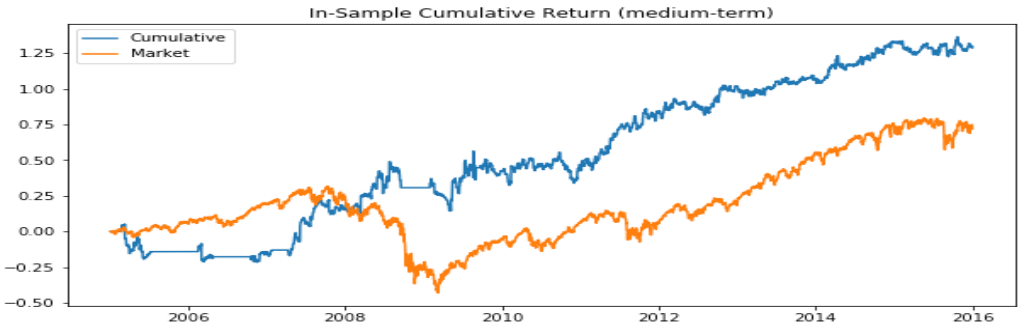


Fig. 5.4.2-2 In-sample cumulative returns for medium-term strategy

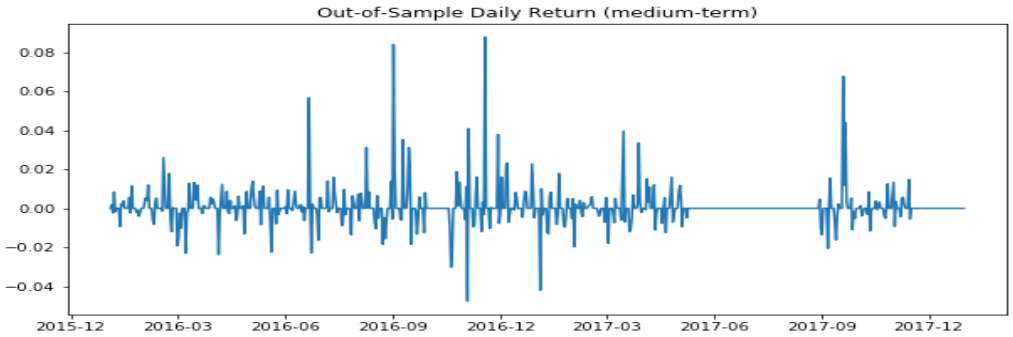


Fig. 5.4.2-3 Out-of-sample daily returns for medium-term strategy

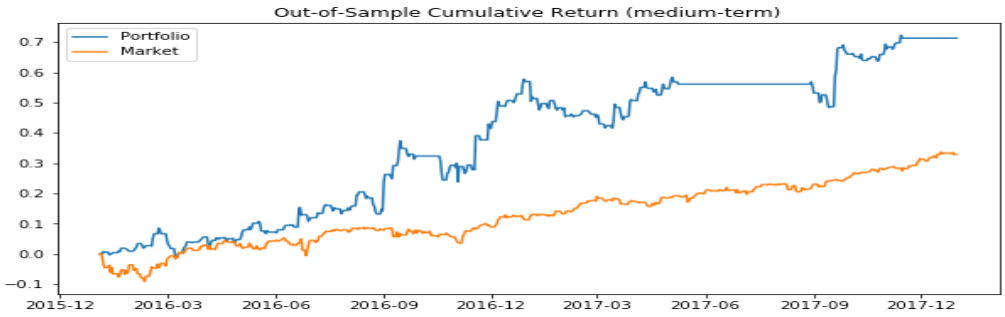


Fig. 5.4.2-4 Out-of-sample cumulative returns for medium-term strategy

Table 5.4.2-1 and 5.4.2-2 show the annualized IR and t-statistics for the strategy executed in the training set (in-sample) and the test set (out-of-sample). It is obvious that the IR we obtain from medium-term strategy is lower for both in-sample and out-of-sample than that from short-term strategy. This means that our medium-term strategy to generate excess returns relative to the market is not satisfactory for a single year. The low IR and t-statistics are mainly caused by

the relatively low return mean, which is different from the case in short-term strategy when high volatility plays major cause. But if we look at the cumulative returns over time, the medium-term strategy beats the market.

Table 5.4.2-1 In-sample annualized IR and t-Stats

Year	IR	t-stats
2005	-0.1281	-0.6212
2006	-0.0581	-0.2926
2007	0.1335	1.2703
2008	0.1036	0.9264
2009	-0.0293	-0.3175
2010	0.0016	0.0206
2011	0.0526	0.6590
2012	0.0081	0.0987
2013	-0.1291	-1.5937
2014	-0.0105	-0.1300
2015	-0.0085	-0.1055

Table 5.4.2-2 Out-of-sample annualized IR and t-stats

Year	IR	t-stats
2016	0.1319	1.5615
2017	-0.0010	-0.0074

It is interesting that the prediction accuracy for short-term horizon is lower, but the performance of short-term trading strategy is much better. At first glance, this seems contradictory. However, if we realize how the labels are generated for short-term and medium-term horizons, we find the results reasonable.

For short-term prediction, the upward/downward label is generated by the value of next-day return, and thus, the labels are 100% correct, which means all the news with an upward movement in the relevant stock price are labeled as one group, and all the other news are labeled as the other group. Then the SVM

classification predicts the labels in the test set, which brings some prediction error. However, for medium-term prediction, the labels are generated through time series clustering, which is an unsupervised model and brings some clustering error. Thus, the labels are already not perfect before the SVM classification, which will bring some prediction errors.

In short, the short-term prediction has only one source of error, while medium-term prediction has two different sources of error. Thus, even though the prediction error from SVM is smaller for the medium-term horizon, its total error may be still larger than that of the short-term horizon. Therefore, it is possible that the short-term trading strategy performs better than the medium-term strategy.

6. Conclusion

In this paper, we have explored the possibility of the market reacting differently to various types of negative public news. Our study relies on the assumption that there are certain distinct features incorporated in the news which could lead to a specific behavior in the movement of stock prices. For our analysis, we have developed a systematic framework combining time series clustering on stock returns after an update of negative news and Natural Language Processing(NLP) on the news contents. In addition, we have created trading strategies for both short-term and medium-term horizons based on our analytical scheme. We will discuss below the significant findings from our research, mainly focusing on the objectives stated in Part 1 Introduction.

Through the implementation of the time series clustering method, we can derive four different kinds of market reaction patterns over the subsequent three months after the news becomes public: downward drift, short-term reversal, medium-term reversal, upward drift. Based on the CVI which returns the accurate measurement of "goodness" of clustering results, we believe the foundation of our research is solid with all the news events grouped into four clusters exhibiting four distinct market patterns.

This observation answers our question that we have raised at the beginning. We can conclude that negative news tends to cause different market reaction patterns, depending upon the content of the news released. From using the bag-of-words model on the news in each cluster, we find that bad news with

fewer negative words and more positive words tends to predict an upward movement.

Furthermore, we have developed a machine learning based framework for predicting subsequent market reactions to negative news update for both short-term and medium-term horizons. We have implemented different NLP models to extract the features for each cluster and applied the SVM classifier on the news in the test set to label each news based on its features. The accuracy scores for the short-term prediction are not as good as we have anticipated, but the accuracy scores for the medium-term prediction are relatively good.

Lastly, we have moved our analysis further by developing trading strategies for both short-term and medium-term horizons that are viable from the framework we developed. The short-term strategy we have designed to bet on the next-day stock price movement gives us good returns, for both in-sample and out-of-sample implementation, even though the daily returns are not very stable over time, which is common when the signal is generated from the news. The medium-term strategy we have developed to bet on the medium-term reversal also provides acceptable performance, with its cumulative return beating the market. The back-test results over ten years demonstrate the effectiveness of our trading strategies.

We would like to conclude by sharing our plans in the future. It is always important to remember that having a good source of data is essential in any academic research. We believe that our research can have more depth in the

analysis if we had all the different global stocks included in the RepRisk platform. The 8,020 stocks with news events from the 1,008 stocks initially extracted from the Russell 3000 data could not quite match with the plethora of news articles in the news data. Hence, the first task in the future is to plan a robust method in labeling all the tickers of the stocks in the news data platform. This would allow us to come up with a much larger number of stocks that are matched with those labeled tickers. Then the question is if we have had more stocks in our data, how would this expand and improve our research? We could get more clusters, but do not know exactly how many clusters we can get, implying that we can explore more different market reaction patterns. However, is it always better to get more clusters for this research? Does the market show many more different patterns? Our analysis will be more robust in the future if we also think about these issues because we believe that a stronger model in the assessment of market reactions to negative news event will lay a firm groundwork before building a good trading strategy.

7. References

[1] Tetlock, P.C. (2010). Does public financial news resolve asymmetric information? *Review of Financial Studies,* 23, 3520-3557.

[2] Manela, A. (2014). The value of diffusing information. *Journal of Financial Economics,* 111, 181-199.

[3] Luss, R. & D'Aspremont, A. (2012). Predicting abnormal returns from news using text classification, *Quantitative Finance*, 15:6, 999-1012.

[4] Kogan, S., Levin, D., Routledge, B.R., Sagi, J.S., Smith, N.A. (2009). Predicting risk from financial reports with regression. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, NAACL '09, 272–280, Stroudsburg, PA, USA. Association for Computational Linguistics.

[5] Ruiz, E.J., Hristidis, V., Castillo, C., Gionis A., Jaimes, A. (2012). Correlating financial time series with micro-blogging activity. In *Proceedings of the fifth ACM international conference on Web search and data mining*, WSDM '12, pages 513–522, New York, NY, USA. ACM.

[6] Tetlock, P.C., Saar-Tsechansky, M., Macskassy, S. (2008). More than Words: Quantifying Language to Measure Firms' Fundamentals. *The Journal of Finance*.

[7] Agarwal, A. Xie, B., Vovsha, I., Rambow, O., Passonneau, R. (2011). Sentiment analysis of twitter data. In *Proceedings of the Work- shop on Languages in Social Media*, LSM '11, pages 30–38. Association for Computational Linguistics.

[8] Kim, S.M., & Hovy, E. (2006). Extracting opinions, opinion holders, and topics expressed in online news media text. In *Proceedings of the Work- shop on Sentiment and Subjectivity in Text*, SST '06, pages 1–8, Stroudsburg, PA, USA. Association for Computational Linguistics.

[9] Bollen, J., Mao, H., & Zeng, X. (2010). Twitter mood predicts the stock market. ArXiv e-prints. Retrieved from https://arxiv.org/abs/1010.3003.

[10] Bollen, J., Pepe, A., Mao, H. (2011). Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena, in: *Proceedings of the International AAAI Conference on Weblogs and Social Media (ICWSM)*.

[11] Antweiler, W., Frank, M. (2006). Do U.S. Stock Markets Typically Overreact to Corporate News Stories, *SSRN Electronic Journal*, Retrieved from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=878091

[12] Shengle, L., Stephen, R. (2012). Are under- and over-reaction the same matter, *Journal of Economic Behavior & Organization,* 84, 39-61.

[13] Liu, E., Ahluwalia, V., Datta, D., Zhang, D. (2014). Identifying and Predicting Market Reactions to Information Shocks in Commodity Markets. Retrieved from https://pdfs.semanticscholar.org/523e/a43eab945bbe395e4b343a592fe76e6f3787.pdf

[14] Kraussll, R., & Mirgorodskaya, E. (2017). Media, Sentiment and Market Performance in the Long Run. *The European Journal of Finance*, V23(11).

[15] Boubaker, S., Farag, H., Nguyen, D.K., (2015). Short-term overreaction to specific events: Evidence from an emerging market. *Research in International Business and Finance*, 35, 153-165.

[16] Baule, R. & Tallau, C. (2016). Stock Returns Following Large Price Changes and News Release- Evidence from Germany. *Credit and Capital Markets*, 49 (1), 57–91.

[17] Sinha, N.R. (2016). Underreaction to News in the US Stock Market. *Quarterly Journal of Finance*, 6(2).

[18] Chari, S.G., Desai, P.H., Borde, N. (2017). A review of literature on short term overreaction generated by news sentiment in stock market. Retrieved from http://irgu.unigoa.ac.in/drs/bitstream/handle/unigoa/4828/Anushandhan_7%281%29_2017_12-21.pdf?sequence=1&isAllowed=y.

[19] Loughran, T., McDonald, B., & Pragidis, I. (2018). Assimilation of Oil News Into Prices. SSRN Electronic Journal.
Retrieved from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3074808

[20] Piccoli, P., Chaudhury, M., Souza, A. (2017). How do stocks react to extreme market events? Evidence from Brazil. *Research in International Business and Finance*, 42, 275-284.

[21] Gálvez, R.H., Gravano, A. (2017). Assessing the usefulness of online message board mining in automatic stock prediction systems. *Journal of Computational Science*. 19,1877–7503.

[22] Paparrizos, J., Gravano, L. (2015). k-Shape: Efficient and Accurate Clustering of Time Series. In *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*, 15, 1855-1870. ACM, New York, NY, USA.

[23] Giorgino T. (2009). Computing and Visualizing Dynamic Time Warping Alignments in R: The dtw Package. *Journal of Statistical Software*, 31(7), 1-24.

[24] Sarda-Espinosa, A. (2017). Comparing Time-Series Clustering Algorithms in R using the dtwclust Package. technical report, Retrieved from https://cran.r-project. org/ web/packages/dtwclust/vignettes/dtwclust.pdf.

[25] Hastie, T., Tibshirani, R., Friedman, J. (2009). The Elements of Statistical Learning. 2nd Edition, chapter 14.3. New York: Springer-Verlag.

[26] Nielsen, F,Å., Balslev, D., Hansen, L.K. (2005). Mining the posterior cingulate: segregation between memory and pain components. *NeuroImage*. 27 (3),520–522.

[27] Berry, M.W., Browne,M. (2005). Email Surveillance Using Non-negative Matrix Factorization. *Computational and Mathematical Organization Theory.* 11 (3), 249–264.

[28] Cohen, W. (2005). Enron Email Dataset. Retrieved 2008-08-26.

[29] Deerwester, S. T., Furnas, G.W., Harshman, R.A., Landauer, T.K., Lochbaum, K.E., & Streeter, L.A. (1989). Computer information retrieval using latent semantic structure.
Retrieved from https://patents.google.com/patent/US4839853A/en

[30] Deerwester, S.T., Furnas, G.W., Landauer, T.K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41, 321–407.

[31] Landauer, T.K., Foltz, P.W., & Laham, D. (1998). Introduction to latent semantic analysis. *Discourse Processes*, 25, 259–284.

[32] Tonta, Y., Darvish, H.R. (2010). Diffusion of latent semantic analysis as a research tool: A social network analysis approach. *Journal of Informetrics*. 4(2),

166-174.

[33] Raschka, S. (2016). Python Machine Learning. 1st Edition. Chapter 8. 236-243, 249.

[34] Fama, E.F. (1991). Efficient Capital Markets: II. *The Journal of Finance*. 46(5), 1575–1617.

[35] Aghabozorgi, S., Shirkhorshidi, A.S., Wah, T.Y. (2015). Time-series clustering – A decade review. *Information Systems*.53, 16-38.

[36] Hong, Harrison, and Jeremy C. Stein, 1999. A Unified Theory of Under-reaction, Momentum Trading and Over-reaction in Asset Markets. *Journal of Finance*, 54, 2143–2184.

[37] Barberis, Nicholas C., Andrei Shleifer, and Robert W.Vishny, 1998, A model of investor sentiment, *Journal of Financial Economics,* 49, 307-343.

[38] Daniel, Kent D., David A. Hirshleifer and Avanidhar Subrahmanyam. 1998, A theory of overconfidence, self-attribution, and security market under- and over-reactions, *Journal of Finance, 53, 1839–1886.*

[39] Fama, Eugene F, 1998, Market efficiency, long-term returns and behavioral finance, *Journal of Financial Economics, 49, 283-306.*

[40] Brown, K.C., Harlow, W.V., & Tinic, S.M. (1988). Risk aversion, uncertain information and market efficiency. *Journal of Financial Economics, 22, 355-385.*

[41] Naderi, M. & Mekanik, S. (2012). An Analysis of Short-Term Overreaction to Stock Market News: Iranian Evidence. SSRN Electronic Journal, Retrieved from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2260070.

[42] Xie, B., Passonneau, R.J., Wu, L., Creamer, G.G. (2013). Semantic frames to predict stock price movement. *51st Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Sofia, Bulgaria.

[43] Herz,R.S., Eliassen, J.C., Beland, S., & Souza, T. (2003). Neuro imaging evidence for the emotional potency of odor evoked memory. *Neuropsychologia*, 42, 371–378.

[44] Blei, D.M., Ng, A.Y., Jordan, M.I. (2003). Lafferty, J., ed. Latent Dirichlet

Allocation. *Journal of Machine Learning Research.* 3 (4–5), 993–1022.
[45] Hoffman, M., Blei, D.M., & Bach, F (2010). Online learning for latent Dirichlet allocation. in *NIPS Proceedings.*
[46] Teh, Y.W., Newman, D., Welling, M. (2007). A Collapsed Variational Bayesian Inference Algorithm for Latent Dirichlet Allocation. Retrieved from https://papers.nips.cc/paper/3113-a-collapsed-variational-bayesian-inference-algorithm-for-latent-dirichlet-allocation.pdf
[47] Tirunillai, S., & Tellis, G.T. (*2014*). Mining Marketing Meaning from Online Chatter: Strategic Brand Analysis of Big Data Using Latent Dirichlet Allocation. *Journal of Marketing Research,* 51(4), 463-479.
[48] Landauer, T.K. & Dumais, S.T. (1997). A solution to plato's problem: The latent semantic analysis theory of the acquisition, induction, and representation of knowledge. *Psychological Review,* 211-240.
[49] Landauer, T.K., Foltz, P.W., Laham, D. (1998). Introduction to Latent Semantic Analysis. *Discourse Processes*. 25 (2–3), 259–28.
[50] Ding, X., Zhang, Y., Liu, T., & Duan, J. (2015). Deep Learning for Event-Driven Stock Prediction. In *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence (IJCAI 2015)*, Retrieved from https://www.ijcai.org/Proceedings/15/Papers/329.pdf.
[51] Chew, M., Puri, S., Sood, A., Wearne, A. (2017). Using Natural Language Processing Techniques for Stock Return Predictions. SSRN Electronic Journal, Retrieved from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2940564.
[52] Länsilahti, S. (2012). Market reactions to Environmental, Social, and Governance (ESG)-news: evidence from European markets. SSRN Electronic Journal. Retrieved fromhttps://aaltodoc.aalto.fi/bitstream/handle/123456789/3021/hse_ethesis_12744.pdf?sequence=1target=.

8. Appendix

	Top 40 words in first 10 topics generated by LSA
Topic 1	reportedly tax billion report allegedly million company offshore used oil cash federal use haven registered accused energy environmental international health coal justice corporate pay based water apple bank according fund gas include business avoid government revealed nuclear non new apparently
Topic 2	tax offshore billion haven registered cash reportedly used avoid apple low approximately respectively revealed non corporate business fund based justice use federal paying competition antitrust foreign make wages research police facing face technology genetically factory number country money trade tobacco
Topic 3	news information criticism latest severe source refer tax corp offshore police bribery reg istered investigation haven cash used apple business avoid low justice illegal world appr oximately exchange respectively revealed foreign non data antitrust subsidiary commissio n market following executive company paying competition
Topic 4	million company pay settlement lawsuit agreed court case department fraud settle drug j ustice accused commission federal investigation related district class exchange similar pri ce act agreement illegal bribery damages billion antitrust subsidiary foreign executive co mpensation paying administration recently apple corp fine
Topic 5	report nuclear bank cluster munitions financial million international billion china funds p ension pay group fraud include involved association percent world according united mak e human risk general total respectively justice major received global business various ex ecutive despite included millions anti violation
Topic 6	million coal energy billion oil gas settlement carbon pay power climate fraud total agree d list fuel bank percent change petroleum industry include electric settle federal fine rel ated spill corporate shell fund received respectively general public natural news offshore price refer
Topic 7	oil palm cluster munitions bank nuclear financial international million spill billion labor land indigenous report forced production petroleum involved funds canada shell human g eneral group total accused food various make linked compensation pension fund area fra ud china supply despite offshore
Topic 8	cluster munitions company billion coal report percent group pay allegedly china financial violation use production according energy list tax agreed technology industry action sett lement fraud number executive fine corporate total fund data carbon compensation relate d cause offshore settle apparently potential
Topic 9	water environmental million health toxic indigenous report chemical protection people wa ste pollution agency food land use mining air project plant research clean human safety cancer settlement department risk linked site center local labor chevron environment co ntamination river according causing used
Topic 10	reportedly coal allegedly china cluster million labor munitions climate carbon food group bank change factory government human people work global accused genetically world i nternational action working palm fund linked mining wages pension forced production n ational bribery use said land corp

	Top 40 words in first 10 topics generated by NMF
Topic 1	tax billion offshore haven cash registered used reportedly apple avoid corporate fund use business approximately respectively justice low based revealed non federal report allegedly paying shell accused public make group bank subsidiary government year research foreign investigation country new law
Topic 2	environmental plant health protection agency toxic air chemical pollution waste clean safety power department new cancer site contamination according environment state river center apparently research public linked cause causing note city analyst act electric institute control included data federal national
Topic 3	news information criticism severe latest source refer corp police world bribery international investigation data people illegal public exchange new subsidiary following act executive related technology group analyst apple foreign security number government led high market management commission note used business
Topic 4	million pay settlement agreed fraud settle department related justice bribery exchange illegal price compensation according agreement case paying act bank drug health received money total subsidiary respectively foreign executive similar corp damages note analyst high commission investigation government year financial
Topic 5	report include according based compensation pay executive toxic corporate health research apparently risk fuel global major climate despite avoid data use electric gold forced industry institute public production human natural canada shell total price security chemical international program new high
Topic 6	energy climate power change data electric according based public use center apple management petroleum general fuel high corporate state subsidiary number environment union control like received include commission technology institute year revealed cause pollution clean apparently exchange river human federal
Topic 7	oil palm spill land new production climate petroleum linked shell fund action fuel change involved environmental local claim area offshore forced canada act include pollution environment number facing related river food industry high compensation make analyst state despite management causing
Topic 8	nuclear bank international financial funds pension involved world plant association state china following millions various area campaign accused make united respectively business despite included report canada general justice global court federal risk include reportedly major total human billion million power
Topic 9	company commission apparently accused according investigation pay executive new following despite subsidiary compensation exchange union similar gold drug administration city information safety data competition security act based trade millions management foreign local illegal said year agency public order market recently
Topic 10	reportedly used compensation climate based use non apparently people following despite world china revealed executive change genetically high facing work low cash order factory working area business justice face foreign carbon exchange corp said bribery money similar linked case state