

A THEORETIC FRAMEWORK INTEGRATING TEXT MINING AND ENERGY DEMAND FORECASTING

Wen-Bin Yu^{1*}, Bih-Ru Lea² and Balasubramania Guruswamy¹

¹ *Department of Information Science and Technology*

² *Department of Business Administration*

School of Management and Information Systems

University of Missouri-Rolla

Rolla, MO (65409), USA

ABSTRACT

News articles are an important source providing information about society. The analysis of news articles helps to measure the social importance of many events and to give an understanding about current interests. In this research, a theoretical framework of text mining enhanced approach is proposed to accommodate short-term variations caused by special events, such as severe weather conditions. A sentiment analysis approach for extracting sentiments associated with positive or negative polarities from a series of news reports is utilized to illustrate impact on energy demand from a special event. The magnitudes of the sentiments from the series of news articles are used to compose a time-series pattern to represent the event that translated into the causes of short-term demand or price variation. The proposed approach for sentiment analysis is demonstrated with experimental results. In particular, the development of an event pattern using text mining is illustrated. The implications of the proposed framework and the future research direction are discussed.

Keywords: Text Mining, Forecasting, Sentiment Analysis

1. INTRODUCTION

Demand and price forecast is essential for energy producers and consumers when planning strategies to maximize their benefits. In particular, in a competitive market, accurate forecasting is key to an efficient energy supply network. In today's information era, consumers are usually influenced by environmental dynamics such as weather conditions, news reports of disruption of pipelines, economic trends, etc. Consequently, fluctuation of demand and consumers' expectations of price changes would take place because of the impact of those events.

Information sources such as news articles provide information about society. People read news to understand what is happening and perceive what might happen in the future. Hence, the analysis of news helps to measure the social importance of events and to give an understanding about current interests. In a sense, the analysis of news can be used to predict readers' potential reaction to various events and, thereby, provide a useful mechanism to forecast possible variations for energy-related business

operations such as demand and price fluctuation. For example, events, such as severe weather, fire in a power plant, promotion or positive scientific review of a new alternative energy source would direct/redirect consumers' attention about their preference of use of energy source and, therefore, could have a significant impact on the demand of related products. In a sense, information retrieved from news reports can provide an indication of an important demand driver, i.e. latest consumer interests, beyond historical forecasts.

The objective of this study is to develop a framework for forecasting short-term demand variations with a focus on a sentiment analysis approach for extracting sentiments in news articles as demand drivers. The sentiment analysis methodology is based on a text-mining technique that captures important key terms from unstructured data such as news, distinguishes the key terms into positive or negative lexicons, ranks those key terms based on the frequency of occurrences, and measures the development and duration of an event by capturing the magnitudes of sentiment revealed in a series of related news reports. The results of the proposed study lay a foundation for the analysis of impacts of a special event on demand and/or price variations.

* Corresponding author: yuwen@umr.edu

Section two of this paper presents a survey of the relevant literature on demand forecasting, data/text mining and news analysis using sentiment analysis. Section three presents the framework of the proposed architecture and details the procedures for the text mining process. Section four presents a case study that shows the development of event pattern for Hurricane Katrina in 2005. Finally, the conclusions and future research directions are addressed in section five.

2. LITERATURE REVIEW

2.1 Energy Demand Forecasting

Traditional methods in demand or price forecasting for energy products are mostly time-series forecasting or causal forecasting methods. Basically, time-series forecasting methods are extrapolative models based on the price or demand series. Examples of such models include: moving average, autoregressive models and their variations or integrations (ARIMA), dynamic regression (DR), transfer function (TF) and artificial neural network (ANN) [2][19][30]. Contreras and Santos [2] observed that time-series models perform better than other methods, such as wavelet-transform or neural network techniques. Among time-series techniques, the dynamic regression and the transfer function algorithms are more effective than ARIMA models. Nogales et al. [30] applied both DR and TF approaches to conduct electricity price predictions for the Spanish and Californian electricity markets. Differences between both markets have been observed even with the same forecasting methods.

Causal methods relate possible factors to develop the cause and effect relationships for energy prices. Examples of causal models include: linear regression (LR), artificial neural network (ANN), neuro-fuzzy systems and support vector machine (SVM) [23] [24] [35]. Most of the ANN models were proposed to accommodate possible nonlinear cause-effect relationship. Khotanzad and Elragal [23] studied the nonlinear relationship between gas demand with recent demand values and weather data by using a non-adaptive ANN-based gas load forecaster. Liu et al. [24] applied SVM on natural gas load forecasting based on the principle of structure risk minimization.

Traditional forecasting methods have been criticized because they do not learn from new data as they arrive. Artificial neural network (ANN) models have been introduced to enhance the learning capabilities of both time series and causal forecasting methods. In most of the ANN time-series forecasting models, data from previous time periods ($t, t-1, \dots, t-n+1$) are used as input nodes to predict the future value of the next time period ($t+1$). Hill et al. [15] compared ANN models with statistical time-series

methods as well as the judgmental-based method and found that the ANN model performed more accurately than or comparing to the traditional methods. In an earlier study by Tang and colleagues [37], they concluded that a traditional Box-Jenkins (ARIMA) model performed better for short-term forecasting, while ANN was better for long term forecasting. However, for a short memory time series (short input data), ANN also performed better than the Box-Jenkins model. Kao [21] proposed a forecasting system based on a fuzzy neural network with initial weights generated by genetic algorithm (GA). Results from his study showed the system provided more reliable forecasts by introducing GA, and also the ANN had a similar result when it was compared to ARIMA.

Even ANN models, which were claimed to have a learning capability, are still limited to the same principles of using historical data (maybe more recent history) to create extrapolative models. The assumption for extrapolative methods is that the past is a predictor of the future. However, information that is not found in the historical data may also have an impact on the forecasts. For example, items such as industry trends, new competitors, and new products, which might have a significant impact on future demand, would not be accounted for past history [11]. Such events may not all be random in nature. A systematic way to collect such information would certainly improve forecasting accuracy.

2.2 Data/Text Mining

The phenomenal growth of databases in almost every area of human activity has created a huge potential for new powerful tools to extract knowledge from all types of information sources. The problem of converting data into useful knowledge by uncovering relationships from large databases involves data manipulation and retrieval by applying traditional mathematical and statistical techniques [6] [16]. The knowledge discovery in databases (KDD) is a non-trivial, iterative, and multi-step process for developing a model with a uniform description of data and patterns for manipulating the data to gain useful knowledge [10]. Data mining is an important step in the KDD process consisting of enumerating models over the data by applying efficient algorithms and discovering useful knowledge from the given data sources [12]. The additional KDD processes such as data preparation, data cleaning, and interpretation of the mining results ensure that useful knowledge is interpreted from the data [25]. KDD incorporates methods including statistical pattern recognition, applied statistics, machine learning, and neural networks to find patterns from data in the data mining step of the KDD process [6].

The available information can be divided into two forms: structured and unstructured data. Structured data have a schema to represent the data, and are handled by querying and reporting against predetermined data types and by understanding relationships. For example, web pages in the Amazon website (www.amazon.com) are dynamically generated from an underlying structured relational database and, thus, have the same schema containing the title, authors, price, and other information for each book page. These kinds of data for which schema and relationships can be established are termed as *structured data*.

On the other hand, unstructured data can be defined as data that do not have any proper structure. Everyone uses unstructured data in their day-to-day lives. The unstructured content does not have a data type definition or conceptual definitions; i.e., in textual documents, a word is just a word. For example, news, notes, customer requests, e-mail, reviews, reports, spreadsheets, and other types of documents are all unstructured data having no predetermined forms.

Most of the work in knowledge discovery in databases has been concerned with structured data despite the tremendous amount of online information that appears only in collections of unstructured text [7]. However, a lot of information is now available in the form of natural language texts. Text mining has emerged as a new area of text processing that attempts to fill the demand for a tool to turn data into useful knowledge [5] [8].

Text mining can be defined as data mining applied to textual data, i.e., this process involves the discovery of new facts and knowledge from large collections of texts that do not explicitly contain the knowledge to be discovered [14]. Text mining uses unstructured textual information and examines it in an attempt to discover the structure and implicit meanings “hidden” within the text. Most approaches to text mining apply mining algorithms on labels or keywords associated with each document [22].

Text mining involves various stages such as document collection, preprocessing, and knowledge discovery [23] [36]. In the first step, relevant documents are collected by identifying the right source. The documents are then preprocessed and transformed into the required format. The last step involves the text mining operations where patterns and relationships are discovered from the transformed documents.

Various techniques for text mining are discussed in the literature such as information extraction, clustering, and categorization [4] [9] [22] [32]. Feldman et al. [9] performed text mining via information extraction. In this approach, the information is extracted from each document to find meaningful events, facts, and entities in the given

domain, and then the data mining operations are performed on the extracted information. Karanikas et al. [22] introduced a number of linguistic pre-processing techniques such as tokenization, part of speech tagging, and lemmatization in order to feed the information extraction system. The terms and events are extracted to construct a table for feeding the clustering algorithm.

2.3 News Analysis Using Sentiment Analysis

Text mining techniques discussed in this paper help to uncover the sentiment present in the unstructured information such as news articles. Analyzing news articles using text mining techniques is a growing research area. Montes-y-Gomez et al. [27] followed the idea of a structured representation of the contents of a news report using a list of keywords or topics with their respective frequencies. Montes-y-Gomez et al. [28] used simple statistical representations of the news reports (frequencies and probability distribution of topics) and statistical measures (the average of the median, the standard deviation, and the correlation coefficient) for analysis and discovery of some interesting facts, but mainly trends, associations, and deviations. The trend was then analyzed to study the behavior of society’s interests and to determine which topics contributed most significantly to the trend. One of their ideas, ephemeral association discovery, focuses on the analysis of a very common phenomenon in news, that is, the influence of the peak news topic on other topics. Another idea, deviation detection focuses on irregularities, mainly on detecting news reports that differ from the typical case in their topics, as well as on detecting the specific sources of news flows.

Among the different types of information present in a news article, one useful type is the sentiment, or the opinion expressed towards the subject in the news article. The core issue in sentiment analysis is to identify how sentiments are expressed in texts and whether the expressions indicate a positive or negative opinion towards the subject.

The literature provides various approaches to extract sentiments from documents. An approach based on directional criteria is proposed by Hearst [13] (e.g., these include sentiments in favor of, neutral to, or opposed to an event). Hearst proposed a sentence interpretation model called direction-based text interpretation (DTI). This model is inspired by cognitive linguistics. It isolates the portion of a text that can be interpreted within the framework of a general, domain independent metamorphic model such that semantic interpretation is done and the directionality of the sentence is determined. Within this restricted model, lexical items are assigned a value to only

one semantic attribute, thus circumventing the need for large complex knowledge bases required by full text understanding systems. Integrating this method with an information retrieval system yields an incremental improvement in a text classification system. Another approach includes machine learning techniques (naïve Bayes, maximum entropy classification, and support vector machines) that were applied to analyze the sentiment classification of the documents [31]. Using movie reviews as data, Pang et al. found that standard machine learning techniques definitely outperform human produced baselines.

Instead of classifying the sentiment of an entire document about a subject, sentiment can be analyzed and detected from all references to the given subject, and sentiment in each of the references can be determined using natural language processing (NLP) techniques [39]. Natural language processing is a modern computational technology that helps to evaluate, understand, and translate unstructured texts into useful information. The sentiment analyzer developed by Yi et al. [39] consists of a topic-specific feature term extraction, sentiment extraction, and sentiment association by relationship analysis. The sentiment analyzer uses two linguistic resources for the analysis: the sentiment lexicon and the sentiment pattern database.

Most of the literature on sentiment analysis concentrates on extracting sentiments from reviews such as movie reviews [31], customer reviews [17] [39], and stock message boards [3] [34]. Sentiment

can be extracted from customer reviews by mining the features of a product on which the customers have expressed their opinions and determining whether the opinions are positive or negative [17]. This task can be performed by mining the product features commented on by the customers, identifying the opinion in each sentence of every review and deciding whether each opinion sentence is positive or negative. To decide the opinion orientation of each sentence (positive or negative), three subtasks are performed. First, a set of adjectives (which are normally used to express opinions) is defined using a natural language processing method. Second, semantic orientation is determined for each opinion word using a bootstrapping technique from WordNet [26]. Finally, the opinion orientation of each sentence is decided, and the final results are summarized.

The technique to identify positive and negative sentiments from news articles offers enormous opportunities for various applications. It would provide powerful functionality for competitive analysis, stock market analysis, technology change forecasts, and detection of opinions in product or movie reviews. The sentiment analysis results could also be used for forecasting and decision making. It would help the users to make decisions without having to read through all the articles.

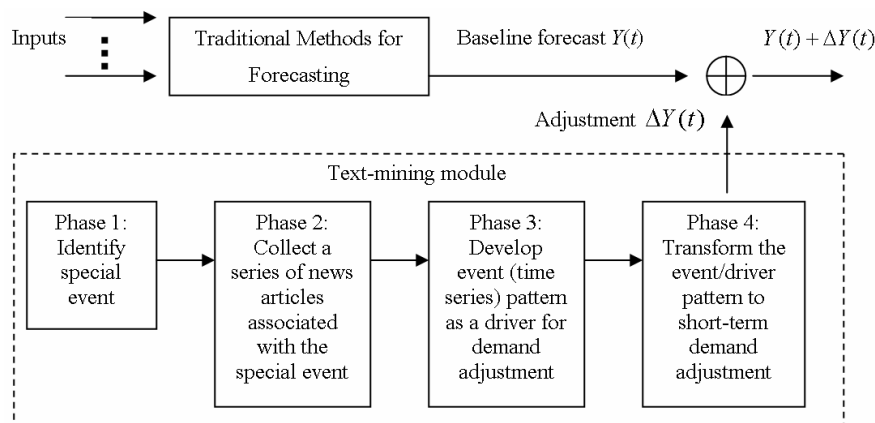


Figure 1: The proposed framework using text-mining to aid traditional demand forecasting

3. THE FRAMEWORK

This paper proposes a framework using the information revealed in news articles to enhance short-term demand forecasting. This framework calls for integration of traditional forecasting methods (e.g., ANN) and text-mining techniques, as illustrated in Figure 1. The output using traditional forecasting methods alone is called *baseline forecast* (say, time series $Y(t)$) in this paper. As the authors have argued

in the introduction section of this paper, text-mining techniques can provide additional insights that cannot be considered in traditional forecasting methods. The proposed text-mining module starts with identifying special events that may have an impact on demand. The module will output an adjustment (say, time series $\Delta Y(t)$) to be added to the baseline $Y(t)$.

The detailed steps of the text-mining module are explained as follows:

Procedure of the text-mining module

Phase 1: Special event identification

The purpose at this phase is to identify special events that would have an impact on the demand. Any short-term demand variations would most likely be caused by a special or unexpected event. An example of an unexpected event would be severe weather conditions. Such special events that occur randomly and cannot be incorporated in a traditional forecasting would have a delayed effect on the demand.

Phase 2: News collection

The purpose at this phase is to collect news reports addressing the progress of the special events identified at Phase 1. Normally, the activity involved in this step would be searching news with appropriate search key words.

Phase 3: Event pattern development

From the news reports, a time-series pattern can be developed to describe the progress of the event based on the positive/negative sentiment found from the news articles. In the sequel, a “time-series pattern” may simply be called a “pattern” for short.

Phase 4: Demand impact assessment

A statistical relationship between the event patterns obtained in Phase 3 and energy demand patterns is established. The impact of the events on the actual energy demand is assessed in terms of an adjustment term to be added to the baseline forecast.

Both Phases 1 and 2 are straightforward and are demonstrated in Section 4, where a case study is presented. Phase 3, the main focus of this paper, involves a module for event pattern development. Detailed activities in Phase 3 are described in Section 3.1. Phase 4 is sketched in Section 3.2.

3.1 Phase 3: Event Pattern Development Using Sentiment Analysis on News articles

The basic principle for developing a time-series event pattern using sentiment analysis is to extract sentiments associated with polarities of positive or negative for the keywords and to classify the whole document as either positive or negative. The structured representation of the contents of a news article is presented as a list of keywords with their corresponding frequency of occurrences and weights. The magnitude each news article is then decided by such frequencies and weights of the keywords.

Analysis of news collections is an interesting challenge because news reports have many characteristics different from the texts in other domains. For instance, news topics have a high correlation with society interests and behavior, they are very diverse and constantly changing, and they also interact and influence each other [28]. Most of the literature concentrates on extracting opinions only from stock message boards and customer or movie

reviews where the sentiment towards a subject is often explicitly mentioned by the users. However, in this paper we take a different approach by extracting sentiments from news articles using keywords rather than frequently occurring phrases. Unlike stock message boards or reviews, news articles do not always contain key phrases based on which sentiments can be decided. Rather, news articles possess keywords that are often repeated whenever a positive (good) or negative (bad) news article is published. The keywords thus repeated are given weights based on the term frequency–inverse document frequency (TF-IDF) algorithm. The sentiment of the whole news article is decided based on the weights for positive and negative key words present in the article.

Procedure of the Event Pattern Development

Step 1: Pre-process the collected documents

This step pre-processes the retrieved articles for any kind of required transformation. The keywords are extracted, and separate lexicons are created for positive and negative wordlists.

Step 2: Sentiment analysis

In this step, high-level information such as sentiments, magnitude and patterns are discovered and extracted from the pre-processed document.

The methodology is described in detail in the following sections.

3.1.1 Document Preprocessing

The news articles could be collected from reliable websites such as Google news, CNN Yahoo news, and BBC based on some inputted search terms. After the news articles are identified, each article is then represented by a set of terms that characterize its content using a “String Tokenizer”. The extracted terms are stemmed using a stemmer and validated using WordNet [26]. The preprocessing process is shown in Figure 2.

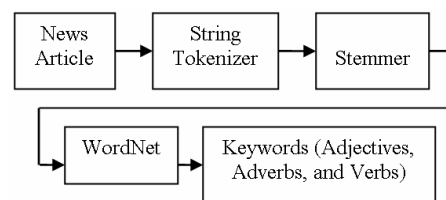


Figure 2: Document Preprocessing

String Tokenizer

The Tokenizer function segments the unstructured text into broadly conceived ‘words’ identified by indexing tokens, which are then often normalized to lowercase. The normalization of the text into lowercase helps to identify repeating keywords irrespective of their case. This function then breaks the text into sentences and divides those

sentences into tokens based on a set of pre-defined delimiters (the characters that separate the tokens).

The application of tokenization rules or token boundaries helps to segment the text into tokens. The most common form of the rules is the popularly known as "black and white tokens" [38]. In this rule, the contiguous regions of non-white space characters are extracted whenever the white space characters separating the tokens occur in the text. This kind of tokenization works well in some cases, but when punctuation characters are involved, this method fails to give proper results. Hence, in addition to white space characters, the punctuation characters are also used as delimiters for tokenization.

The Tokenizer function identifies each word in the article by using delimiters such as period, spaces, commas, and semi-colons. The importance of tokenization stems from the fact that the output from this process determines the keywords for the document collections to be used in the subsequent processes. The Tokenizer function thus identifies those keywords and returns it for the stemming process.

Stemming Process

The stemmer function parses the text by eliminating all the commonly occurring words that do not carry any weight. The commonly occurring words such as articles, prepositions, and conjunctions are identified and cleaned by the stemmer process. The keywords thus extracted are passed into another stemmer process developed by Porter [33]. The Porter stemming algorithm is a process for removing the common morphological and inflectional endings such as prefixes and suffixes from words in English.

Stemming by the Porter algorithm reduces all the variant word forms into their common root assuming that if two words have a common root, then they represent a similar concept or meaning. Thus, stemming allows a text mining system to match document terms related to a same meaning but occurring in different morphological variants [1]. This algorithm thus normalizes the terms for the text mining process. The stemmed keywords are then passed to WordNet [26] for validation. Only those keywords returned as verbs or adjectives will be used for further processing.

Term extraction

The primary objective of the term extraction operation is to identify facts and relations in text by identifying the appropriate verbs and adjectives present in the document. Because the news articles extracted were based on search keywords, which are mostly nouns, the nouns present in the documents were disregarded. Also, using noun phrases tends to produce too many low-precision terms [17]. Thus, the verbs and adjectives are extracted. The extracted

terms were then stored in separate flat files for positive and negative lexicons.

The term extraction process identifies the morphosyntactic categories (or alignments) (verb, adjective, adverb, etc.) of words in the documents. The words are classified into verbs and adjectives using WordNet. WordNet [26] is an on-line lexical reference system whose design is inspired by current psycholinguistic theories of human lexical memory. It organizes English nouns, verbs, and adjectives into synonym sets, with each set representing one underlying lexical concept. Verbs are one of the most important lexical and syntactic parts of a language, and most of the sentences in English contain at least one verb in it. An adjective modifies nouns, whereas modification is not the primary function of noun, verb, and prepositional phrases. Adjectives have particular semantic properties that are not shared by other modifiers. This process extracts verbs and adjectives and filters non-significant words on the basis of their morphosyntactic category.

3.1.2 Sentiment Analysis

Filtering process

The filtering process extracts only the most frequent terms present in news articles. The selection is based on the frequency of the keywords, i.e., only those words with a document frequency equal to or above a given threshold θ will be considered. The keywords from the document are extracted using a proposed triple filter. The triple filter verifies and validates a particular term if it had enough occurrences in the past. The filter ensures that only if a particular term had three or more occurrences in the past would it be eligible for addition into the positive or negative lexicons. In this way, most of the redundant words occurring in the document are filtered to a considerable extent. The triple filter helps to reduce unnecessary noise present in the final results.

Keywords

Apart from adjectives, other content words such as verbs and adverbs are also used to express sentiments. A sentence using an adjective indicates the sentiment towards the subject present in the sentence. For example, a phrase such as *bad person*, which uses an adjective *bad*, denotes the negative sentiment present in the sentence against the subject. Thus, the whole phrase itself becomes a negative sentiment expression with the same polarity as the sentiment adjective *bad*. Similarly, in a phrase such as *play beautifully*, the adverb *beautifully* indicates the positive sentiment towards the verb *play* and the polarity of the sentiment is inherited by the verb. Thus, sentiment expressions using adjectives, adverbs, and verbs can be

classified as either positive or negative based on their polarities.

Nasukawa and Yi [29] classified sentiment-related verbs into two types:

- (1) Sentiment verbs that direct either positive or negative sentiment toward their arguments.
- (2) Sentiment transfer verbs that influence sentiment among their arguments.

Thus, sentiments can be defined as expressions in a sentiment lexicon that consists of information on polarity such as positive (good) or negative (bad).

Keyword ranking

The terms or keywords are accompanied by their frequencies in the document, and a weight signifying their importance is assigned by the text mining process. Keywords in a text are the terms that represent a document, and the candidate keywords are extracted from the analysis results of the document. The keyword-based ranking method is based on the keyword importance factors in a document. It is an analytic approach that analyzes the contents of a document to retrieve a keyword list from the document [20].

The keyword ranking gives more flexibility in determining weights for the most frequent keywords. The keywords are ranked dynamically based on the frequency of occurrences and the higher the number of occurrences, the higher would be the ranking of the keywords. This word ranking can provide sufficient weights for the keywords and will improve the accuracy of the sentiment analysis algorithm.

Documents can be represented in many ways using different feature descriptions. The most straightforward description of documents is based on document vectors. The documents represented by document vectors consist of <keyword, weight> pairs. A document vector specifies the keyword and its corresponding weight calculated based on how often the term occurs in that document. The document vectors are constructed from the term frequency (TF) and the inverse document frequency (IDF) [18]. In a document, vector term frequency and document frequency are the most important features to calculate the weight of the term. Detailed information regarding term frequency (TF) and the inverse document frequency (IDF) can be found in Appendix A.

Sentiment and magnitude

The sentiment analysis process uses an algorithm based on the naïve classifier in determining the sentiment of a news article. The naïve classifier algorithm is based on a word count of positive or negative connotation words. It is the simplest and most intuitive of the classifiers [3].

The mining process analyses the news articles and compares the terms extracted with those stored in

positive and negative lexicons. Each term in the news article is checked against a lexicon and assigned a value based on its word ranking. The algorithm behind the sentiment analysis process is described below.

The positive keyword set $p = \{p_1, p_2, p_3, \dots, p_i\}$ where p_i is the positive term, and w_{p_i} is the weight associated with the positive term. The negative keyword set $n = \{n_1, n_2, n_3, \dots, n_j\}$, where n_j is the negative term. w_{n_j} is the weight associated with the negative term. The news article document D_k is defined as a mixture of both positive p and negative n keyword sets. The (naïve) sentiments, $S(D_k)$, of the document D_k is defined as shown in Equation (1):

$$S(D_k) = \max\left[\sum_i w_{p_i}(D_k \cap p_i), \sum_j w_{n_j}(D_k \cap n_j)\right] \quad (1)$$

The sentiment of the document $S(D_k)$ is termed as positive when the sum of the positive weights, is more than the sum of the negative weights and termed as negative when the sum of the negative weights is more than the sum of the positive weights. When the sums of the positive and negative weights are equal, then the sentiment of the document $S(D_k)$ is classified as a neutral article. Each news article is, thus, associated with a sentiment magnitude. The magnitude of the news article is calculated by finding the difference between the sum of the positive weights and the sum of the negative weights associated with the news article as shown in Equation (2).

$$\text{Magnitude} = \sum_i w_k(D_k \cap p_i) - \sum_j w_k(D_k \cap n_j) \quad (2)$$

An example of computing the sentiment of an article can be found in Appendix B. The accumulated magnitudes of a series of new articles that reports a specific event in a certain period simply represent the development of the event and hence, define the event pattern used in this study.

3.2 Phase 4: Demand Impact Assessment

Some special events can impact future energy demand of interest. For example, a more severe aftermath of a tornado would incur higher energy demand and cause a significant spike on the prices for a longer period of time than a mild damage caused by a thunderstorm. It may not be easy to quantify individual impact of a natural disaster simply based on its characteristics, such as wind speed and diameter. However, the event patterns developed by news reports regarding the damages

caused by the tornados or thunderstorms can be used to assess the significance of the events.

The procedure for assessing demand impact of some special events is presented as follows.

Procedure of Assessing Demand Impact

Step 1: Model building from past experiences

The underline assumption in this step is that the demand adjustment (from the baseline to be forecasted in Figure 1) caused by the special event is a simple transformation from the event pattern. Three parameters need to be decided from past experiences: lag, duration factor, and magnitude factor. (Figure 3) Lag (l) defines the delay between the beginning of the event and the beginning of the demand adjustment. Duration factor is defined as the ratio of the duration of the demand adjustment (d_2) to the duration of the event (d_1). And finally the magnitude factor is defined as the maximum of the demand adjustment (m_2) divided by the maximum of the magnitude from the event pattern (m_1).

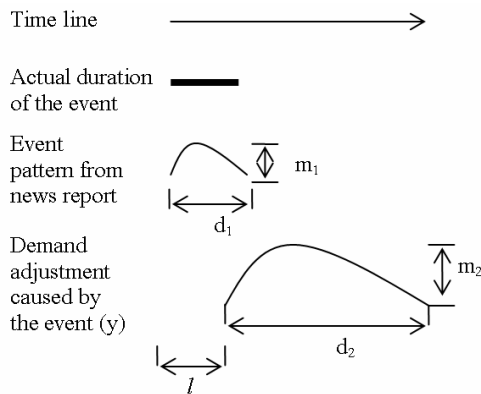


Figure 3: Transformation from event pattern to demand adjustment

Step 2: Pattern transformation

Once the three parameters described in the previous step are decided from historical data, the transformation of event pattern to actual demand adjustment can be computed by adding the lag and applying both factors to the duration and magnitudes of the event pattern.

Assume the time index sets for durations of pattern and demand are $i \in \{1, 2, \dots, d_1\}$ and $j \in \{1, 2, \dots, d_2\}$, respectively. Given the time series for the event pattern denoted as $\{x_i\}$, the time series for the demand adjustment is $\{y_j\}$, where each y_i can be computed by the following equation:

$$y_j = \frac{m_2}{m_1} \times [x_i + (x_{i+1} - x_i) \times (\frac{d_1}{d_2} j - i)] \quad (2)$$

with i satisfies $\frac{d_2}{d_1} i \leq j < \frac{d_2}{d_1} (i + 1)$.

Equation (3) is depicted in Figure 4.

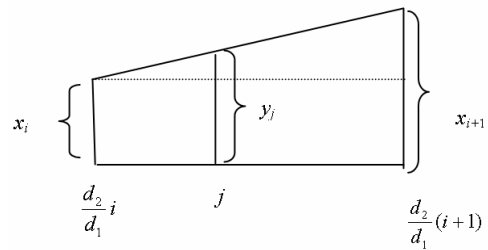


Figure 4: The correspondence between x_i and y_j

The actual adjustment (ΔY) on the demand time series will be made based on the time lag (l) and the demand deviation y_j transformed from the event pattern (x_i).

4. CASE STUDY

A case study has been conducted to shows how sentiment analysis can be applied for analyzing cumulative magnitude of news articles related to a particular event. This case study will focus on development of event patterns, corresponding to Phases 1 to 3 in Figure 1. As for Phase 4 that bridges text mining and demand forecasting, its implementation will not be included in this case study due to availability of a data bank that connects useful news information and demand data. This task, however, is sketched in this section.

4.1 News Collection for a Specific Event: Hurricane Katrina in 2005

In this experiment, a collection of news articles related to Hurricane Katrina was retrieved from the same news website. The news collection included about sixty news articles collected over a time period of Aug. 24, 2005 to Sept. 9, 2005, when Hurricane Katrina created huge destruction to life and property. Because of the large-scale devastation caused by the hurricane, most news articles collected during this period have a negative sentiment with varying magnitude.

4.2 Keyword Ranking

Keyword ranking is done by associating a weight with each keyword based on the TF-IDF method. The weights are then used for building document vectors containing keywords with their associated weights. A suitable weight is assigned to the keyword based on the frequency of occurrences from the trained articles.

The positive and negative lexicons shown in Figures 5 and 6 are built by training a mixture of various news articles published from a pre-selected Web Site. The lexicon is built by training 200 news articles. The positive lexicon contains 329 keywords, and the negative lexicon contains 336 keywords. Each keyword in the lexicon is represented as a document vector containing the

keyword and its frequency of occurrence. The lexicon building is an evolving process, and the keywords in the lexicons keep growing as more news articles are recorded.

```
rank02
good09
seed03
try02
great07
high05
through03
end06
appear03
Cooperate04
extend03
began04
show08
dominate03
head08
score01
different05
decide06
rule02
enable03
mark02
gain04
complete02
train02
unite06
remain03
unbeaten02
play04
coach02
plan02
continue02
democratic05
libery05
freedom05
start04
```

Figure 5: Positive lexicon

```
bad07
downward07
down05
fail06
disgust04
disappoint06
fear04
fearful03
negative04
attack06
pull02
terrible10
end06
force04
later10
appear03
dismiss02
early04
extend03
dominate03
head08
decide06
rule02
enable03
mark02
train02
remain03
want04
unseed02
plan02
difficult06
continue02
start04
charge02
punch02
```

Figure 6: Negative lexicon

4.3 Event Pattern

As mentioned in section 4.1, a collection of news articles related to Hurricane Katrina was collected over the time period of Aug. 24, 2005 to Sept. 9, 2005 to show the negative sentiment with varying magnitude during the period of interest. The cumulative magnitude of a news article is calculated by finding the difference between the sum of the

positive weights and the sum of the negative weights associated with the news article. The cumulative magnitude of each news article is plotted according to the time interval as shown in Figure 7.

Figure 7 shows the variations of the cumulative negative magnitude present in the news articles due to the impact of Hurricane Katrina. The hurricane lashed various cities in US starting from Aug. 26 to Aug. 30, 2005. The hurricane forecast news published from Aug. 24 to Aug. 26 had low cumulative magnitude for the negative sentiment associated with the news articles. However, after Aug. 26, 2005, the cumulative magnitude for negative sentiment gradually increased with the severity of the hurricane. This can be seen from the progress of the cumulated magnitudes in the Figure 7. During the period from Aug. 28 to Aug. 30, the cumulative magnitudes averaged around -300 and the news articles showed very high cumulative magnitude for negative sentiment associated with it. This was the period when reports of massive destruction caused by the Hurricane started appearing predominantly in the news articles.

The aftermath of the Hurricane Katrina is captured during the period starting from Sept. 1 to Sept. 5. The cumulative magnitudes during this period were reduced and became less intense when compared with that of the earlier period of hurricane fury. However, the negative sentiment from the cumulative magnitudes still averaged at about -170, which was considerably higher than that of the starting period when the magnitudes averaged only -50. From Sept. 5 to Sept. 9, most news articles addressed the recovery and rescue efforts, which reduced the negative magnitude considerably. During this period, the focus of the news articles shifted from addressing destruction to recovery, and the cumulative magnitude averaged at about -60.

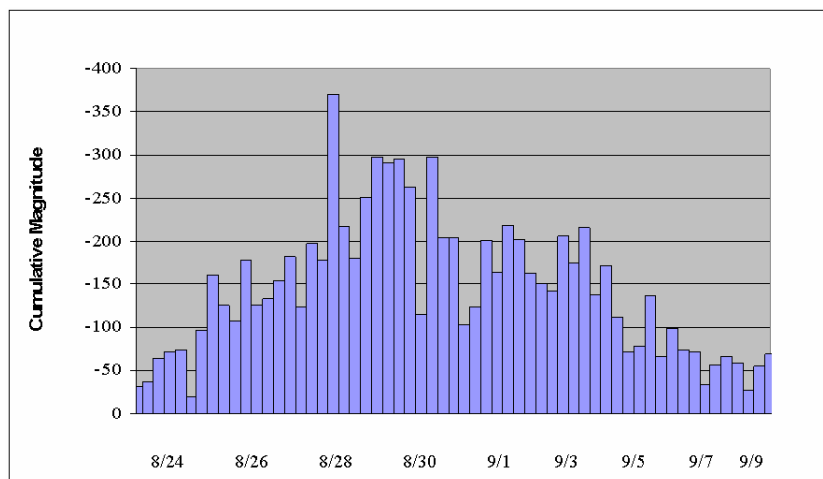


Figure 7: Hurricane Katrina’s impact on news articles

The hurricane's impact analyzed during a two-week period showed that the intensity or cumulative magnitude of the negative sentiment increases and then decreases according to the aftermath of the hurricane. This experiment shows that the cumulative magnitude of sentiment for news articles with related topics can be tracked effectively to provide a quantitative representation of the event, both the duration and degree of the negative impact caused by the hurricane.

Once the event patterns as shown in Figure 7 have been identified, the actual forecasting for any demand time series of interest can be done by transforming this event pattern to demand time series adjustment with appropriate lag, duration and magnitude factors as described in Section 3.2. This task requires a data bank that contains news information for special events and demand data. The establishment of the data bank often requires time and effort to collect enough news events so that useful transformation information described in Section 3.2 can be extracted. Since the purpose of the present paper is to propose a theoretical framework for integrating text mining and load forecasting, detailed implementation of demand forecasting task will be tackled in a future paper.

5. CONCLUSION AND FUTURE DIRECTION

The continuous publication of news articles on the Internet has made the task of monitoring the news articles virtually impossible. This research focused on the development of special event pattern with a sentiment analysis approach for monitoring the news articles by extracting the polarity of positive or negative sentiments. Because sentiments can be conveyed in different ways including indirect expressions that require common sense reasoning for recognition, it has been a challenge to demonstrate the feasibility of the proposed system. However, the experimental results show that useful information on sentiments can be extracted from news articles with this implementation.

The experiment for magnitude of the sentiments illustrated how magnitudes obtained from sentiment analysis can be used for tracking a series of news articles over a period of time. The forecast and the aftermaths of Hurricane Katrina were captured in this study, and the cumulative negative magnitudes were analyzed. The magnitudes during the peak of the hurricane fury were considerably higher than that of the initial period. The news articles that addressed the recovery and rescue efforts had a less negative magnitude than the news articles that covered the massive destruction caused by the hurricane. The hurricane's impact analyzed during the two-week

period showed that the magnitude of the negative sentiment fluctuated according to the progress of the hurricane and the damage caused by the hurricane. Based on the magnitude, the sentiment analyzer accumulated all the related news articles that will have a cascading impact on the future.

This research contributes by providing a mechanism for analyzing and categorizing news that can be a useful input for demand forecasting applications. This study differentiates itself from the traditional method of applying weather forecasts or reports as the indicators of the impact of a natural disaster. Usually people would determine the impact of a hurricane based on its size, speed and/or route. However, there are other factors that can contribute to actual impact of a hurricane. For example, had the levees in New Orleans not broken, the damages caused by hurricane Katrina might have been less significant. Hurricanes having exactly the same characteristics and routes as Katrina may not necessary cause the same energy prices fluctuations. But similar aftermath caused by different natural disasters might have similar patterns of impact on, for example, gas demands. News reports captured the actual development of the event, which would not be revealed simply based on the characteristics of the hurricane.

The results from the case study can be extended to forecasting energy demand fluctuation due to a special event such as the aftermath of a hurricane. Additional studies that link the event patterns to actual demand variations as described in Phase 4 of the proposed framework would need to be carried out in order to complete the entire forecasting system. The development of an event can be quantified based on the proposed study and, hence, provide an excellent starting point for provide more accurate and flexible prediction of future energy demand adjusted to the ever changing environments.

REFERENCES

1. Bacchin, M., Ferro, N. and Melucci, M., 2002, "The effectiveness of a graph-based algorithm for stemming," *Proceedings of the 5th International Conference on Asian Digital Libraries: Digital Libraries: People, Knowledge, and Technology*, pp. 117-128.
2. Contreras, J. and Santos, J. R., 2006, "Short-term demand and energy price forecasting," *Proceedings of 2006 IEEE Mediterranean Electrotechnical Conference*, pp. 924-927.
3. Das, S. R., and Chen, M. Y., 2001, "Yahoo! for Amazon: Sentiment parsing from small talk on the Web," *Proceedings of the 8th Asia Pacific Finance Association Annual Conference*.

4. Dörre, J., Gerstl, P. and Seiffert, R., 1999, "Text mining: Finding nuggets in mountains of textual data," *Proceedings of the fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 398-401.
5. Dunja, M., 2000, "Workshop on text mining," *Proceedings of the Sixth International Conference on Knowledge Discovery and Data Mining*, pp. 534-542.
6. Fayyad, U., Piatetsky-Shapiro, G. and Smyth, P., 1996, "The KDD process for extracting useful knowledge from volumes of data," *Communications of the ACM*, Vol. 39, No. 11, pp. 27-34.
7. Feldman, R., Fresko, M., Kinar, Y., Lindell, Y., Liphstat, O., Rajman, M., Schler, Y. and Zamir, O., 1997, "Text mining at the term level," *Springer Lecture Notes in Computer Science*, pp. 65-73.
8. Feldman, R., 1999, "Workshop on text mining: foundations, techniques and applications," *Proceedings of the 16th International Joint Conference on Artificial Intelligence*.
9. Feldman, R., Aumann, Y., Liberzon, Y., Ankori, K., Schler, J. and Rosenfeld, B., 2001, "Information retrieval and text mining: A domain independent environment for creating information extraction modules," *Proceedings of the Tenth International Conference on Information and Knowledge Management*, pp. 586-588.
10. Geist, I., 2002, "Declarative data mining: A framework for data mining and KDD," *Proceedings of the 2002 ACM Symposium on Applied Computing*, pp. 508-513.
11. Gilliland, M. and Prince, D., 2001, "New approaches to "unforecastable" demand," *Journal of Business Forecasting Methods and Systems*, Vol. 20, No. 2, pp. 9-13.
12. Han, J. and Kamber, M., 2001, *Data Mining: Concepts and Techniques*, Academic Press.
13. Hearst, M. A., 1992, "Direction-based text interpretation as an information access refinement," *Text-Based Intelligent Systems*, Laurence Erlbaum Associates.
14. Hearst, M. A., 1999, "Untangling text data mining," *Proceedings of the ACL'99: the 37th Annual Meeting of the Association for Computational Linguistics*, pp. 3-10.
15. Hill, T., O'Connor, M. and Remus, W., 1996, "Neural network models for time series forecasts," *Management Science*, Vol. 42, No. 7, pp. 1082-92.
16. Hirji, K. K., 2001, "Exploring data mining implementation," *Communications of the ACM*, Vol. 44, No. 7, pp. 87-93.
17. Hu, M. and Liu, B., 2004, "Mining and summarizing customer reviews," *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 168-177.
18. Jing, L., Huang, H. and Shi, H., 2002, "Improved feature selection approach TFIDF in text mining," *Proceedings of the 1st International Conference on Machine Learning and Cybernetics*.
19. Kaboudan, M. A., 2001, "Computometric forecasting of crude oil prices," *Proceedings of the 2001 Congress on Evolutionary Computation*, pp. 283-287.
20. Kang, S., 2003, "Keyword-based document clustering," *Proceedings of the 6th International Workshop on Information Retrieval with Asian Languages*, pp. 132-137.
21. Kao, R. J., 2001, "A sale forecasting system based on fuzzy neural network with initial weights generated by genetic algorithm," *European Journal of Operational Research*, Vol. 129, pp. 496-517.
22. Karanikas, H., Tjortjis, C. and Theodoulidis, B., 2000, "An approach to text mining using information extraction," *Principle and Practices of Knowledge Discovery in Databases, Knowledge Management: Theory and Applications*.
23. Khotanzad, A. and Elragal, H., 1999, "Natural gas load forecasting with combination of adaptive neural networks," *Proceedings of 1999 International Joint Conference on Neural Networks*, Vol. 6, pp. 4069-4072.
24. Liu H., Liu, D., Liang, Y. M. and Zheng, G., 2004, "Research on natural gas load forecasting based on least squares support vector machine," *Proceedings of 2004 International Conference on Machine Learning and Cybernetics*, Vol. 5, pp. 3124-3128.
25. Loh, S., Wives, L. K. and Oliveira, J. P. M., 2000, "Concept-based knowledge discovery in texts extracted from the Web," *ACM SIGKDD Explorations Newsletter*, Vol. 2, No. 1, pp. 29-39.
26. Miller, G., Beckwith, R., Fellbaum, C., Gross, D. and Miller, K., 1990, "Introduction to WordNet: An online lexical database," *International Journal of Lexicography*, Vol. 3, No. 4, pp. 235-312.
27. Montes-Y-Gomez, M., Gelbukh, A., Lopez-Lopez, A. and Baeza-Yates, R., 2001, "Text mining with conceptual graphs," *Proceedings of the IEEE Conference on Systems, Man, and Cybernetics*, Vol. 2, pp. 898-903.
28. Montes-y-Gomez, M., Gelbukh, A. and Lopez-Lopez, A., 2001, "Mining the news: Trends, associations, and deviations," *Revista Iberoamericana de Computación*, Vol. 5, No. 1, pp. 14-24.

29. Nasukawa, T. and Yi, J., 2003, "Sentiment analysis: Capturing favorability using natural language processing," *Proceedings of the 2nd International Conferences on Knowledge Capture*, pp. 70-77.
30. Nogales, F. J., Contreras, J., Conejo, A. J. and Espinola, R., 2002, "Forecasting next-day electricity prices by time series models," *IEEE Transactions on Power Systems*, Vol. 17, No. 2, pp. 342-348.
31. Pang, B., Lee, L. and Vaithyanathan, S., 2002, "Thumbs up? Sentiment classification using machine learning techniques," *Proceedings of the Conference of Empirical Methods in Natural Language Processing*.
32. Pierre, J. M., 2002, "Mining knowledge from text collections using automatically generated metadata," *Proceedings of the 4th International Conference on Practical Aspects of Knowledge Management*.
33. Porter, M. F., 1997, "An algorithm for suffix stripping," *Morgan Kaufmann Multimedia Information and Systems Series*, pp. 313-316.
34. Rajagopalan, B., Konana, P., Lee, C. and Wimple, M., 2004, "Extracting relevance from virtual investing-related community postings," *Proceedings of AMCIS2004 Conference on Text and Data Mining for Decision Support*.
35. Rodriguez, C. P. and Anders, G. J., 2004, "Energy price forecasting in the Ontario competitive power system market," *IEEE Transactions on Power Systems*, Vol. 19, No. 1, pp. 366-374.
36. Tan, A., 1999, "Text mining: The state of the art and the challenges," *Proceedings of PAKDD'99 Workshop on Knowledge Discovery from Advanced Databases*, pp. 71-76.
37. Tang, Z., de Almeida, C. and Fishwick, P. A., 1991, "Time series forecasting using neural networks v.s. Box-Jenkins methodology," *Simulation*, Vol. 57, No. 5, pp. 303-10.
38. Witten I. H., Bray, Z., Mahoui, M. and Teahen, W. J., 1999, "Text mining: A new frontier for lossless compression," *Proceedings of the Data Compression Conference*, pp. 198-207.
39. Yi, J., Nasukawa, T., Bunesco, R. and Niblack, W., 2003, "Sentiment analyzer: Extracting sentiments about a given topic using natural language processing techniques," *Proceedings of the Third IEEE International Conference on Data Mining*, pp. 427- 434.

ABOUT THE AUTHORS

Wen-Bin Yu is an assistant professor of Information Science and Technology at the School of Management and Information Systems, University of Missouri at Rolla. Dr. Yu received his PhD in Computer Science and Engineering from University of Louisville. His research interests are in the fields

of data/text mining, business process simulation, software agent applications, and demand forecasting especially in a supply-chain environment.

Bih-Ru Lea is an assistant professor of Business Administration at the University of Missouri - Rolla. Dr. Lea conducts interdisciplinary research in the areas of Supply Chain Management, Enterprise Resource Planning, Management Accounting Systems, and Business Application of Software Intelligent Agents. Dr Lea has published in numerous journals including International Journal of Production Research, International Journal of Production Economics, Industrial Management and Data Systems, Technovation, and Journal of International Technology and Information Management.

Balasubramania Guruswamy received his MS degree in Information Science and Technology from University of Missouri at Rolla.

(Received April 2007, revised June 2007, accepted August 2007)

APPENDIX

A. Definitions of Term Frequency (TF) and IDF (Inverse Document Frequency)

The term weighting method based on the information retrieval measure TF-IDF is used to construct a document vector. The keyword ranking values are calculated by a combination of the statistics $TF(k,d)$ and $DF(k)$. The term frequency $TF(k,d)$ is the number of times the keyword k occurs in document d . The document frequency $DF(k)$ is the number of documents in which the keyword k appears at least once. The inverse document frequency $IDF(k)$ for the total number of documents $|D|$ can be calculated from the document frequency $DF(k)$ (Jing et al., 2002) as shown in Equation (A1).

$$IDF(k) = \log\left(\frac{|D|}{DF(k)}\right) \quad (A1)$$

The weight w_i of the keyword k_i in document d can be calculated using $TF(k_i,d)$ and $IDF(k_i)$ as shown in Equation (A2).

$$w_i = TF(k_i,d) \times IDF(k_i) \quad (A2)$$

This keyword ranking indicates that a keyword k_i is an important indexing term for document d if it appears frequently in the document. *TF-IDF* weighs the frequency of a term

in a document with a factor that would reduce its significance when it occurs across a collection of documents. The words that appear in many documents are ranked as less important indexing terms because of their low inverse document frequency. Therefore, the keywords that occur very rarely or very often are ranked lower than the keywords that hold the balance. Thus, the inverse document frequency $IDF(k_i)$ contributes to sentiment analysis by serving as an adjusting function to modulate the term frequency.

B. Sentiment Analysis-An Example

The sentiment of each news article is identified by subjecting the news article to keyword analysis. As an example, consider the following news article for sentiment analysis.

A majority of Americans believe the city of New Orleans will never completely recover from the effects of Hurricane Katrina and the resulting flooding, according to results of a CNN/USA Today/Gallup poll released Tuesday. Fifty-six percent of 609 adults polled by telephone September 5-6 said they believe the Hurricane devastated the city beyond repair. And 93 percent of poll respondents said they believe Katrina is the worst natural disaster to strike the United States in their lifetime. But a majority of respondents -- 63 percent -- said they believe the city should rebuild. And 66 percent said they believe all New Orleans residents should evacuate the city. Opinions varied widely, however, on the response of federal, state and local officials regarding Katrina. Forty-two percent of respondents characterized President Bush's response to the disaster as "bad" or "terrible," while 35 percent said it was "good" or "great." Federal government agencies' response was described as "bad" or "terrible" by 42 percent, and "good" or "great" by 35 percent. State and local officials' response was described as "bad" or "terrible" by 35 percent and "good" or "great" by 37 percent. Respondents also disagreed widely on who is to blame for the problems in the city following the Hurricane -- 13 percent said Bush, 18 percent said federal agencies, 25 percent blamed state or local officials and 38 percent said no one is to blame. And 63 percent said they do not believe anyone at federal agencies responsible for handling emergencies should be fired as a result.

First, keywords from the news article are identified using the stemmer. The keywords are then matched with positive and negative lexicons for identifying positive and negative keywords present in the news article. For every match, a suitable weight is assigned to the keyword based on the frequency of occurrences from the trained articles. The weights are calculated using the TF-IDF algorithm. Table B-1 shows a detailed analysis of the keywords and their corresponding weights for positive keywords.

Table B-1: Keywords analysis for positive weights

Positive Words	Doc. Freq.	Term Freq.	Inverse Doc. Freq.	Weights
Good	9	3	1.204	3.612
Great	7	3	1.322	3.967
Unite	6	1	1.398	1.398
Release	3	1	1.699	1.699
Believe	3	6	1.699	10.194
Recover	2	1	1.875	1.875
Rebuild	3	1	1.699	1.699
Responsible	2	1	1.875	1.875
Sum of weights for positive words, $\sum P$				26.319

The positive keyword analysis for the news article shows that words such as *good*, *great*, and *believe* carry more weight than other keywords. This is because these keywords had a higher term frequency i.e., these keywords had a higher number of occurrences in the news article. The higher the number of occurrences of a keyword in a news article, the higher will be the weight associated with the keyword. This is because the weight is a product of the term frequency and the inverse document frequency. Another interesting observation in Table B-1 is the inverse document frequency for each keyword. The keyword *good* has the highest document frequency i.e., this keyword has the highest number of occurrences across all the observed documents. However, it has the lowest inverse document frequency. This is because the TF-IDF algorithm ensures that no particular keyword gets any undue weight based on its past occurrences alone. The past occurrences along with the term frequency of keywords make up for its weight. Table B-2 shows a detailed analysis of the keywords and their corresponding weights for negative keywords.

The negative keyword analysis for the news article shows that words such as *bad*, *terrible*, *disaster*, and *blame* carry more weight than other keywords. The reason for this is these keywords had a higher term frequency in the news article, which contributed towards their weights. Table B-2 shows that the keyword *bad* has the highest document frequency, but it has the lowest inverse document frequency. This is due to the TF-IDF algorithm, which ensures that no particular keyword dominates other keywords. Keywords such as *release*, *describe*, and *effects* are not necessarily negative in context. However, these words were given a small negative weight because they were found in most of the trained news articles with negative sentiment, and hence they were stored in the negative lexicon.

Table B-2: Keywords analysis for negative weights

Negative Words	Doc. Freq.	Term Freq.	Inverse Doc. Freq.	Weights
Bad	7	3	1.322	3.967
Terrible	10	3	1.176	3.528
Release	3	1	1.699	1.699
Describe	2	2	1.875	3.750
Never	4	1	1.568	1.568
Effects	7	1	1.322	1.322
Flooding	5	1	1.477	1.477
Devastate	9	1	1.204	1.204
Repair	8	1	1.255	1.255
Worst	7	1	1.322	1.322
Disaster	9	2	1.204	2.408
Strike	2	1	1.875	1.875
Evacuate	6	1	1.399	1.399
Disagree	5	1	1.477	1.477
Problem	4	1	1.568	1.568
Blame	3	4	1.699	6.796
Emergency	2	1	1.875	1.875
Fire	5	1	1.477	1.477
Sum of weights for negative words, $\sum N$				39.968

The keywords analysis shows that $\sum N > \sum P$; i.e., the sum of the weights for negative words is more than that of the positive words. Thus, the news article is termed as negative in context. The magnitude of the news article is calculated as $26.32 - 39.97 = -13.65$