International Journal of Multimedia Data Engineering and Management
Volume 7 • Issue 4 • October-December 2016

Discovering News Frames:
An Approach for Exploring Text, Content, and Concepts in Online News Sources

Loretta H. Cheeks, Arizona State University, Tempe, AZ, USA
Tracy L. Stepien, Arizona State University, Tempe, AZ, USA
Dara M. Wald, Iowa State University, Ames, IA, USA
Ashraf Gaffar, Arizona State University, Mesa, AZ, USA

ABSTRACT

The Internet is a major source of online news content. Current efforts to evaluate online news content including text, storyline, and sources is limited by the use of small-scale manual techniques that are time consuming and dependent on human judgments. This article explores the use of machine learning algorithms and mathematical techniques for Internet-scale data mining and semantic discovery of news content that will enable researchers to mine, analyze, and visualize large-scale datasets. This research has the potential to inform the integration and application of data mining to address real-world socio-environmental issues, including water insecurity in the Southwestern United States. This paper establishes a formal definition of framing and proposes an approach for the discovery of distinct patterns that characterize prominent frames. The authors' experimental evaluation shows the proposed process is an effective approach for advancing semi-supervised machine learning and may assist in advancing tools for making sense of unstructured text.

KEYWORDS

Clustering, Content Analysis, Data Mining, Machine Learning, News Frames, Non-Negative Matrix Factorization, Text Mining

INTRODUCTION

Approximately 80% of all data today exists in digital form as unstructured text (e.g., news, e-mails, social media feeds, contracts, memos, clinical notes, and legal briefs) (Raghavan et al., 2004). The Internet is the premier digital platform for online news content and unstructured text (Chung, 2008). The ability to unlock hidden structure and latent meanings in unstructured text is an important area of research because text is a fundamental device of communication and human interaction for expressing real-world issues.

"The formation and transmittal of group standards, values, attitudes, and skills are accomplished largely by means of verbal communication" (Cartwright, 1953). As a consequence, efforts to understand social interaction, cooperation and influence require the study of text. Given the widespread use of unstructured text for individual and mass communication, such as email, online news and social media feeds, it is particularly important to understand how social influence and information about complex socio-environmental issues spread through online content.

DOI: 10.4018/IJMDEM.2016100103

Copyright © 2016, IGI Global. Copying or distributing in print or electronic forms without written permission of IGI Global is prohibited.

The use of strategic devices for presenting salient aspects and perspectives about an issue while using certain keywords, as well as stereotyped images and sentences, for the purpose of conveying latent meanings about an issue, it is called framing (Entman, 1993). The framing of news stories can shape public interpretation of social and environmental news (Druckman & Bolsen, 2002; Sheufele & Tewksbury, 2007; Tewksbury & Scheufele, 2009). Public opinion, attitudes, beliefs, and behaviors can be influenced by how an issue is framed, particularly when framing comes from elites (Druckman & Bolsen, 2002).

The causal effect of media communication, specifically the influence of the words and frames the media uses to influence public perceptions of social and environmental issues, has been studied extensively. However, it traditionally has been examined using content analysis, which was developed specifically to aid in the interpretation of social discourse or text for communication research (Holsti, 1969; Krippendorff, 1980). Content analysis involves methodical evaluation and categorization of text (Riffe et al., 2005). Current approaches to content analysis, however, require scholars and researchers to thoroughly examine documents in search of patterns in the text. This approach to text analysis is dependent on humans and limits the applicability of the analysis of large-scale unstructured text. Thus, a more effective tool for unlocking latent meanings found in unstructured text could enhance our understanding of online behaviors, responses to online advertising, and media influence on public perceptions.

Text mining, also known as text data mining (Hearst, 1997), is a multidisciplinary field involving information retrieval, text analysis, information extraction, clustering, categorization, visualization, database technology, and machine learning. Text mining, coupled with an interest in understanding social influence and information diffusion for document summarization (i.e., topic modeling, sentiment analysis, and opinion mining), is an active area of research (Bindela et al., 2015; Kempe et al., 2003). Text mining, when combined with machine learning algorithms, techniques and methodologies, offers an added value to data integration tasks: they highlight the similarities between heterogeneous sources and text features, which reduces uncertainty and risk exposure when performing the integration tasks.

The widespread availability of large data repositories, like those found in online news articles, creates an opportunity to develop new methods of text mining. These methods use machine learning algorithms and mathematical techniques to select, organize, and evaluate large quantitates of unstructured text. While research on the use of frames and public attitudes in traditional news venues has been widely explored, the identification and analysis of frames in unstructured online news has received minimal attention.

Therefore, the purpose of this research is to test a new approach for discovering distinct frame patterns using a process that advances semi-supervised machine learning for document clustering and classification. This research aligns favorably with concurrent efforts in machine learning and data mining that seek to discover novel patterns and latent relationships in unstructured text for deep learning (Blei et al., 2003). This method has several improvements over traditional methods of text mining: First, unlike the traditional approach to processing unstructured texts that focus on the characterization of form (syntactic) and its meaning (semantic meaning) (Stavrianou et al., 2007), this method views news texts as organized symbolic devices (frames) that act as carriers or or cues influencing perceptions about an issue that will interact with individuals' existing beliefs. Second, by placing framing within computer science, this approach allows rigorous processing of text representations that has the potential to extend bodies of research beyond bag-of-words, such as exposing pathways for bridging bag-of-words, clustering techniques, concept mapping, and linked data. Third, this technique enhances current methods of text analysis and mining by using machine learning to explore larger datasets and many more topics than human coders can currently evaluate. Moreover, this new approach will allow for the automatic emergence of the frame from the text at the end of the analysis (Murphy, 2001; Shah et al., 2002), using a combination of inductive and deductive iterative techniques.

We seek to answer the following questions in this research study: How is framing being produced? Which dominant frames emerge from the unstructured text data mining process? To what extent are scientists, policy makers, or business leaders being used as sources?

To address the aforementioned questions, as presented in this paper: a) we formulate a formal computer science definition for framing, b) we define an approach for the discovery of five distinct patterns (referred to as signatures) that characterize prominent frames, and c) we propose a process that advances machine learning for document classification that takes advantage of content analysis on a small number of documents for scaling up to meet the demands of large datasets.

PREVIOUS WORKS

Scholarly research in computer science on latent meanings in association of terms and documents to reveal relationships is found in literature related to text summarization, information retrieval, and text data mining. The earliest paper on text summarization is that of Luhn (1958) that describes work done at IBM in the 1950s. In his work, Luhn (1958) proposed that the frequency of a particular word provides a useful measure of its significance and identified the concept *term frequency* (TF), which states it is possible to identify significant terms solely based on the term calculated frequency of the term within that document. It relates to average information, or entropy, of a term or group of terms ranking in relationship to each other.

In the same year that Luhn (1958) was published, Baxendale (1958) published work done at IBM that provides early insight on a particular feature helpful in finding salient parts of documents: the sentence position, which assumes that important sentences are located at the beginning or end of paragraphs. The following year, Maron & Kuhns (1960) published a paper that was the first to be based on ranked retrieval and ranking of documents by their computed values of probability of relevance. This meant that two-valued indexing of documents could be replaced by weighted indexing where the weights were to be interpreted as probabilities.

Edmundson (1969) was the first to describe a system that produces document extracts. He developed a typical structure for an extractive summarization experiment that integrates features of word frequency and positional importance, borrowed from the works of Luhn (1958) and Baxendale (1958).

Critical to text summarization is the contribution of Salton et al. (1975), credited with developing the vector space model (VSM). Extending the work of Luhn (1958), similarities (and dissimilarities) were computed using both extracted words and cited data, which remains a principle concept upon which many scholarly research leverages. A retrieval model represents documents, description features (such as index terms), queries, and the relationships within and across those sets (Salton et al., 1975).

Meanwhile, Spärck Jones (1972) developed the measure of term specificity that later became known as *inverse document frequency*, or IDF, which proved to be a giant leap in text summarization and information retrieval. It is based on counting the number of documents in the collection being searched which contain (or are indexed by) the term in question. The intuition was that a query term that occurs in many documents is not a good discriminator and should be given less weight than one that occurs in few documents. The measure was a heuristic implementation of this intuition.

Dumais et a. (1988) improved the vector space model (VSM), known as Latent Semantic Indexing (LSI). The development of LSI marks the first scholarly research for transforming a high-dimensional VSM in association of terms with documents into a low-rank approximation to A_{mxn} to filter out noise and improve the detection of relevant documents. LSI uses singular value decomposition (SVD) to derive particular latent semantic structure models.

Other outstanding contributions to text data mining in the 1990s and beyond with the advent of machine learning include text representation and models construction (Salton et al., 1975; Steinheiser & Clifton, 1998; Tan et al., 2000; Tan et al., 2005; Ghanem et al., 2005); data dimensions reduction research in feature extraction (Karanikas et al., 2000; Hu et al., 2003); research on mining algorithm

of text classification (Jiang et al., 2010; Tan et al., 2005) and clustering (Hotho et al., 2003; Luo et al., 2009); and deep semantic mining based on natural language processing (Berendt et al., 2002; Dave et al., 2003; Kushal et al., 2003).

Frame discovery is motivated by framing theory, as known in communication research, which focuses on understanding the latent meanings of observable messages in their contexts (Scheufele, 1999), and can provide important insight into how the presentation or "framing" of an issue affects the choices people make. Other disciplines have focused on framing: in linguistics research, similar approaches are also described as "latent semantic analysis" (LSA) (Dumais et al., 1988). Social Network Analysis (SNA) focuses on the importance of relationships among interacting units (Fellbaum, 1998; Liu, & Singh, 2004).

CONTEXTUAL CASE

Water is a fundamental resource affecting all aspects of life on earth. Water is used for human consumption, industrial processes, production of food, sanitation, as well as other usages. The way water policy and water decisions are framed affects water rights allocations, policy decisions, human consumptions, emerging technologies, farming techniques, and agricultural outcomes (Lombard et al., 2002). The issue of water insecurity in the Southwest Region is particularly important in that the seven states that make up the Southwest Region (Arizona, California, Colorado, Nevada, New Mexico, Utah, and Wyoming) rely primarily on fresh groundwater flows deriving from the Colorado River. The Colorado River was an efficient source of water for decades, however, due to a decade of drought, this once flowing and plentiful source of water cannot meet the demand to sustain life as it exists in this region, contributing to water insecurity throughout the Southwest Region of the United States.

Water insecurity in the Southwest Region of the United States exemplifies an issue that is appropriate for study through the lens of framing in online news sources because this issue is context specific, complex and characterized by uncertainty. Newspapers in the Southwest regularly produce in-depth articles about drought, water and climate change that are published online. However, the general public outside of the Southwest has very little experience with water insecurity. Biased framing or terminology are more likely to influence uninformed respondents (Bishop et al., 1982) or respondents with reduced exposure to or interest in an issue (Krosnick, 1999). Therefore, citizens' attitudes and beliefs about water insecurity are likely influenced by the way reporters frame this issue. This source of unstructured text offers researchers a body of content to test our learning and machine-discovery approach on a relevant socio-environmental issue.

To the best of our knowledge, discovery of frames in online news content has not previously been addressed for this class of feature extraction and diffusion problems. The following sections include a description of our methods, experimental evaluation, and a discussion of our results.

METHODS

This section describes the data, the problem formulation, and the methods for discovering frames contained in online news article unstructured data.

Dataset Description

The news data for our study was collected from Google News, which is a news feed aggregator. Our feature selection comprises four different characteristics of a given article: a) the news articles being published by online news sources, b) the news sources that generated the articles, c) the frequency of news publications by sources over time, and d) the salient language contained in the article.

We restricted the time period from which the news articles published on the Internet by online news sources, which for this study was January 2006 to March 2015. A web data mining software program was developed using the Python programming language, which utilizes a universal web browser bot for traversing the online frontier for the purpose of gathering Uniform Resource Locator

(URL) seeds for collecting online news articles that contain pertinent features needed for this research study. The URL seeds were used as input to the crawler.

A depth-first search algorithm was applied for each URL seed node for this study. Articles were gathered by limiting the search for keywords associated with water or sub-topics of water consequences within Arizona, Colorado, California, and Nevada. 55,000 news articles were collected and the articles were then stored in a database for undergoing data pre-processing that included removing errors and inconsistencies in order to improve the quality of data. For example, date patterns were standardized, duplicates were removed, and articles with missing critical values were removed (except in cases where the author name was not identified). After data pre-processing, 30,000 articles were deemed relevant for our news dataset and 25,000 were deemed irrelevant. Our web data mining program applied string, regular expressions, and tree matching techniques to find patterns and perform alignments. The data records were segmented into an association list, called documents, where an alignment of data items, called properties, contained in the data records, were produced for storing in a table in the database.

Frame Discovery Methods

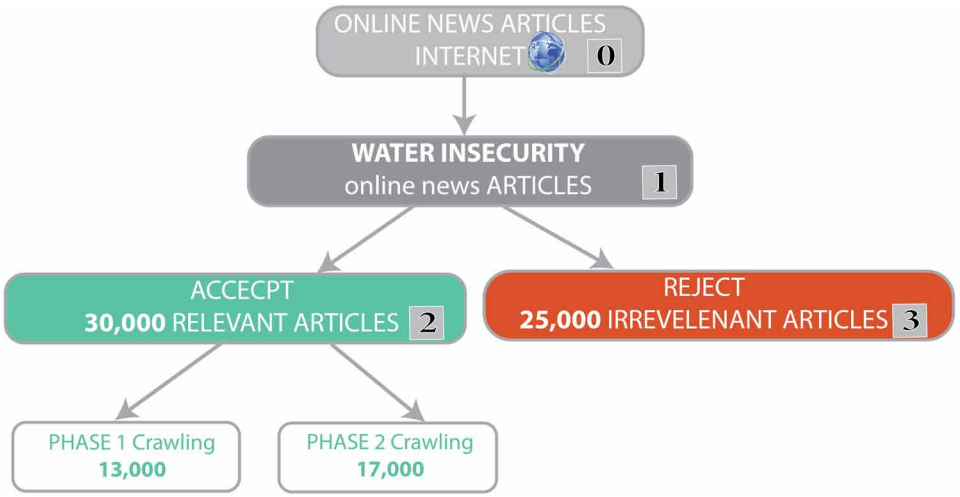
News frames are a central organizing idea that supplies a context and highlights salient features by selecting what to include, emphasizing or deemphasizing certain aspects of a news story, and elaborating or excluding particular perspectives, thereby shaping readers' interpretation of the issues (Severin & Tankard, 1997; Gamson & Modigliani, 1987). Using this definition, we experimented with existing techniques and algorithms for machine learning.

Scholarship exists for text mining and machine learning for novel knowledge discovery of patterns as mentioned above. However, this is the first known research to apply text mining and machine learning for discovering communication frames found in online news articles. Moreover, a discovery of frames may expose clues that illustrate how adversarial news frames influence public understanding of environmental issues, such as climate change effects, water insecurity, and agricultural insecurity (Karlberg, 1997).

Problem Formalization of Frame

Let U_t be a universe of online news documents at time t. Let $\{d_i\}, 1 \leq i \leq n,$ denote a finite subset of documents from the universe of documents U_t. Let $S_i = \{P_i, C_i\}$ be the set of document properties, where P_i is the source that produced the article and propagates a central organizing idea (i.e., framing),

Figure 1. A depiction of the process for the web data mining that resulted in the water insecurity dataset



and C_i is the set of terms (i.e., keywords or content-descriptors) from the news article with common words removed, which supplies a context and suggests what the issue is through the use of selection, emphasis, exclusion, cues, and elaboration (Tankard, 2001). For the issue of water insecurity, we use standard frame types (An & Gower, 2009), in particular, human interest, conflict, economic, managerial, and science. Let $f_1, f_2, f_3, f_4, \text{and } f_5$ correspond to each of these frames, where each f_i is associated with a pair $\{x_i, y_i\}$ where x_i is the feature vector and y_i is the vector with the corresponding weights.

The performance of the algorithm depends on the quality of the vector space model of documents $\{d_i\}$. Our study applied three methods for showing the quality of the vector space and evidencing machine-discovery accuracy: a) content analysis, b) machine learning techniques (TF-IDF, non-negative matrix factorization (NMF), L2 Norm for comparing between vectors, and classification), and c) linked data for feature extraction of elite cues. This paper provides treatment for a) and b) only to provide an approach for evaluating how salient features contained in online news articles are used for producing the frame about the issue covered in the article.

Content Analysis

In order to explore the relationship between online news source framing, we performed a content analysis on articles pertaining to the water insecurity in the Southwestern Region of the United States. Content analysis is a research technique that involves extracting information from text by identifying characteristics or categories of content within a text (Riffe et al., 2005). Quantitative content analysis using statistical methods is used as a tool for evaluating numerical patterns (e.g., most of the items described this concept) and identifying relationships among the categories (Riffe et al., 2005). Content analysis is used to measure dimensions (features) within a sample of representative text. In the following section, the term *coding* refers to the process of classifying the unstructured text and those performing the coding are called *coders*. Here, we identified categories in consultation with a subject matter expert and developed a priori guidelines to measure and identify differences in content.

Our data collection process occurred in phases. First, our crawler identified 13,000 articles that met our initial list of search terms: Water Insecurity, Water Crisis & Shortages, Natural Resource Management Water, Natural Disasters Water, Clean Water, Water Rights, Water Policy Water River, Water Crisis, Water Doubt, Water Shortages, Water Desalination, Water Reclaim, and Climate Change.

We followed the method of Riffe et al (2005) for calculating the sample size, which considers a) the standard error and confidence interval of a given sample mean and b) the estimated variance of the variable in the population. Assuming a 95% confidence level and a 90% minimal level of agreement, we identified a random sample of 372 articles (Krippendorff, 1980). Several of these articles were removed from this study because they were not directly related to the issue of water insecurity. Our final sample included 316 articles, which was still sufficient for a 94.5% confidence level.

The content analysis began with a definition of our research problem: "How is the issue of water insecurity in the Southwest Region of the United States being framed in online news media over time?" Next, we developed a list of variables of interest related to the discourse around and framing of water insecurity in the Southwestern United States: 1) frame types (i.e., human interest, managerial, science, economic, conflict); 2) key actors quoted or cited, and 3) discourse and key actors over time. Based on these key variables we designed the data collection protocol. The protocol included a list of questions that researchers addressed for each item of unstructured text (e.g., online news article). For example, one question included in the protocol was "Does the article mention a science study?" Possible responses included "Yes" and "No." There were a total of 44 questions designed to identify frame types, key actors, and discourse over time. All the questions were categorical with "Yes"= 1 and "No"= 0 response options.

Figure 2. A depiction of the problem formulation

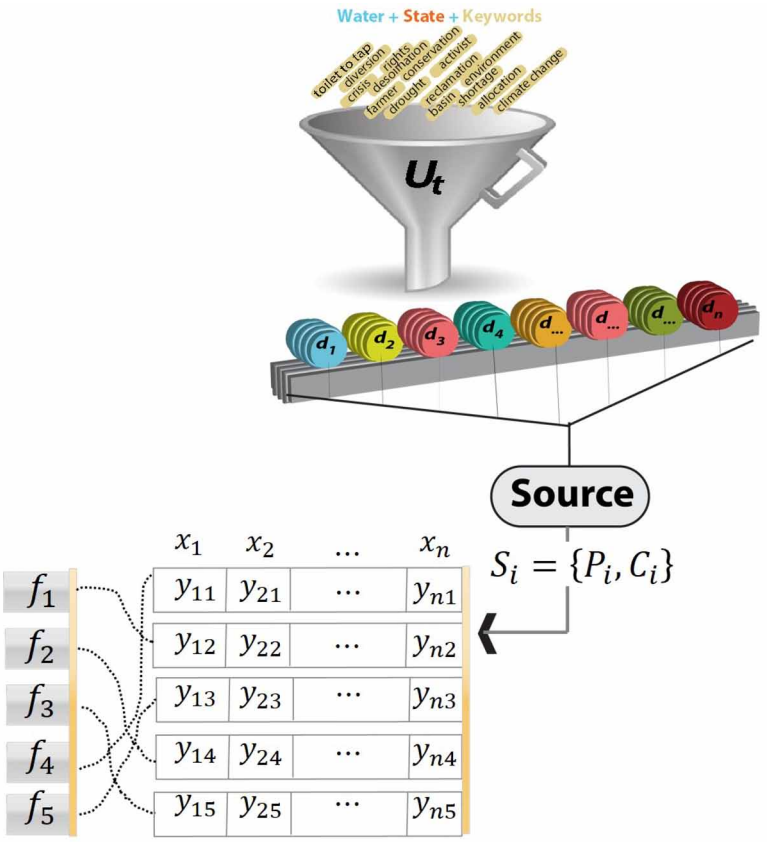
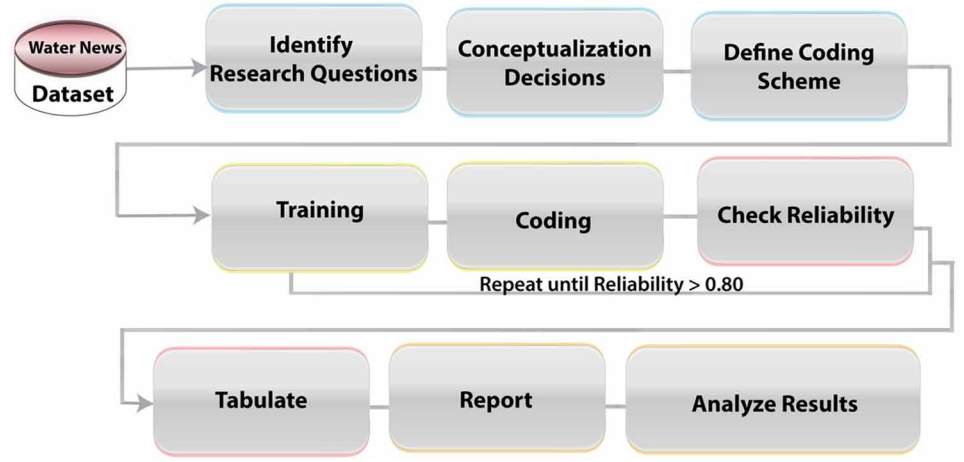


Figure 3. Content Analysis process



Content analysis typically relies on the judgments of multiple coders. To ensure coding reliability, defined as intercoder agreement or "the extent to which independent coders evaluate a characteristic of a text and reach the same conclusion" (Lombard et al., 2002, p. 589), we selected 90 articles, not included in our final sample, for coder training. Three coders, including the first author, participated in three training sessions of 30–60 minutes in length. Initial tests of intercoder agreement failed to produce reliable results; thus, the coding guide was modified and coders participated in additional training. After making these changes, we were able to achieve > 85% agreement between our coder pairs. The content analysis process was laborious and took an extensive amount of time (9 months) to achieve coder reliability. The coding guide was revised multiple times and questions that could not achieve agreement were removed from the analysis. The response to the content analysis survey questions were then put into SPSS (Version 23). Analysis identified five dominant frame types: human interest, managerial, science, economic, and conflict. These dominant frame types were established as the ground truth labels (Glaser & Strauss, 1967).

Machine Learning

In our search to discover framing within unstructured text, we look to machine learning to facilitate learning patterns. Machine learning algorithms and techniques have proven to be effective in selecting salient features for exposing emergent correlations and latent relationships. In our quest to understand how online news sources are framing the issue of water insecurity, we establish a machine learning process for discovering news frames. Our motive is to discover not only how meanings of subsequent terms, sentences, and paragraphs are related but also how the facts regarding these sentences are related. Our news frame discovery process is composed of three main steps: data transformation, frame mapping, and classification.

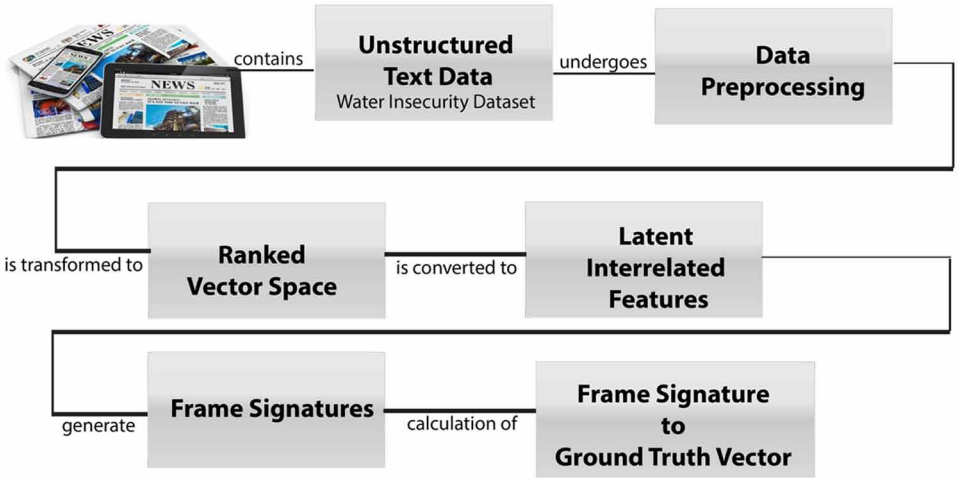
Next, we continued our data collection process, resulting in a dataset of 55,000 online news articles. Because our study involves machine learning and the expectation to learn and thereby discover frames from a small dataset, this dataset was arbitrarily split into two subsets; about 66% of the data was used for a training set and 34% for a test set. Since our approach to text mining is new, we validated our approach and tested the accuracy of our classification methods by comparing the results of our machine learning approach with a more traditional approach to frame identification: content analysis.

Data Transformation

To learn from the article text, we examine the local coherence of the text through models that allow for feature analysis. The features can capture information related to the context of the article. The semantic meaning of a term may change depending on the context of usage. News features are helpful for capturing such information. We apply two feature extraction models for learning. Our baseline model is Term Frequency-Inverse Document Frequency (TF-IDF), commonly known as bag-of-words, which is a weighting scheme used effectively to rank features based on association or co- occurrence and to build a vocabulary that will be used for deep learning (Jones, 1972; Salton & Buckley, 1998; Berger et al., 2000). Our TF-IDF baseline model is used as input to Non-negative Matrix Factorization (NMF) (Lee & Seung, 1999; Bittorf et al., 2012). NMF has properties attractive for this class of problem, namely, dimensionality reduction—under the assumption of linearity and non-negativity, clustering, and local minimum convergence (Bertsekas, 1999, Gu & Zhou, 2009). A key goal of an NMF constrained model (e.g., a model with only positive points) is to accurately identify latent relationships and to identify patterns that together explain the data as a linear combination of expression signatures.

The TF-IDF term weighting scheme, for free text documents, evaluates the specificity and arguably the importance of features contained in the article. TF-IDF transformation results in a vector space model that enables document comparisons, term similarities, and term ranking as determined by the weight, which is referred to as either the score or ranking. The primary assumption of the vector space

Figure 4. This depicts the overview process flow for discovering news frames



model is that documents that are "close together" in space are similar in meaning, thereby represents each document as a point in space (a vector in a vector space). The TF-IDF score (or ranking) of a term in a set of documents is calculated as follows:

$$w_i = tf_{dc} \times idf_c,$$

$$tf_{dc} = \frac{n_{dc}}{m_d},$$

$$idf_c = \log \frac{\left\{d_i\right\}}{1+\left|\left\{d_i \in \left\{d_i\right\} : c_i \in d_i\right\}\right|}$$

where n_{dc} is the frequency of a term c_i in a document d_i, m_d is the maximum frequency of any term in document d_i, $\left\{d_i\right\}$ is the total number of documents in the corpus, and $\left|\left\{d \in \left\{d_i\right\} : c_i \in d_i\right\}\right|$ is the number of documents in which t_i appears. If the term is not in the corpus, this will lead to a division-by-zero, therefore we adjust the denominator by adding 1. The weights w_i show the importance of the terms in each document. Documents were represented using a bi-gram of features. We were interested in retaining only those features that occur in at least 5 documents in the whole training set. The vector space model produced through the calculation of TF-IDF is a high-dimensional matrix, $\left\{d\right\} \in \mathbb{R}^{d \times c}$.

For the news article dataset, we use the TF-IDF data matrix w_i as input to the NMF matrix A that represents $d_1, \cdots d_n$ documents, such that each column corresponds to the features vector x_i from frame f_i. We use the sparse NMF extension by Hoyer (2004), using SVD initialization (i.e. NNDSVD) for addressing the convergence problem, which is common in many clustering algorithms (e.g., K-means or LDA) (Boutsidis & Gallopoulos, 2007). The non-negativity constraints in NMF result in the unsupervised selection of sparse bases that can be linearly combined to reconstruct the original

data for learning parts of the news articles (Liu & Wu, 2010). NMF of a matrix A decomposes it into two matrices W and H such that

$$A = WH.$$

If the dimension of A is $d \times c$, then W has dimensions $d \times r$ and H has dimensions $r \times c$, where r is the dimensionality (i.e., rank). We define r to be five, which corresponds to the five dominant frames (i.e., human interest, managerial, science, economic, conflict) used by online news sources. In NMF, we seek to factor A into r components for representing these frames. The H matrix denotes the five distinct signatures (patterns) where the columns are the important features for the five distinct frames.

This factorization is a constrained non-convex optimization problem with the cost function equal to

$$\left\| A - WH^T \right\|_2^2$$
$$\text{such that } W = \left[w_{d \times r} \right] \geq 0,\ H = \left[h_{r \times c} \right] \geq 0.$$

The cost function is convex only in W or H, but not convex in both together (Ding et al., 2010). The gradient descent contribution of Hoyer (2004) extended Seung & Lee's (2001) iterative update rule for overcoming the non-uniqueness starting point that results in no global solution for the algorithm while presenting an improved convergence to a local minima of the cost function.

When training our model, TF-IDF vectorization preprocessing is performed that includes a) selecting an n-gram range of two for denoting our interest in two word-phrase features, b) ignoring terms that have a document frequency strictly lower than four, c) removing from the vocabulary common words, called stopwords in text mining, which would appear to be less significant in selecting document feature, d) tokenization for breaking up given character sequences found in the document into meaningful words, and e) normalizing for different document lengths using the L_2 norm. NMF uses a) a Nonnegative Double Singular Value Decomposition (NNDSVD) seed, a method designed to enhance the initialization stage of NMF(Boutsidis & Gallopoulos, 2007), b) the rank of the decomposition equaling 5, and c) an update iteration of 50 times before timing out.

Frame Mapping

The frame mapping function is a four-step process that comprises a) sorting, b) aligning, c) measuring, and d) mapping. The NMF part-based representation matrix H, provides an efficient manner to discover hidden structure and latent relationships within data (Donoho & Stodden, 2004). Each frame signature is mapped to a document label applied during the content analysis categorization of dominant frames (i.e., human interest, conflict, economic, managerial, and science). As mentioned earlier, our content analysis is the ground truth for the low-rank approximation that results from the documents reconstructed using NMF H matrix. The feature names were preserved during the transformation.

The frame mapping process takes as input the H matrix, the best rank-r approximation of the original matrix A. In order to create a single frame signature, we start our transformation using the TF-IDF matrix D that contains the content analysis label for the dominant frame under observation, and then we calculate the NMF H. We call this H matrix our "single signature", s_1, which mean its rank-$r = 1$. We differentiate the single signature H matrix from the entire training dataset derived the H matrix where the rank-$r = 5$. We sort the s_1 H matrix according to decreasing ranked values against each frame signature $f_1, f_2, f_3, f_4, \text{and } f_5$, as shown in Figure 5.

The alignment is achieved by a) iterating each row in $f_1, f_2, f_3, f_4, \text{and } f_5$, b) calculating the intersection between the single signature and $f_1, f_2, f_3, f_4, \text{and } f_5$, c) building a feature vocabulary from the resulting intersection for establishing a master single signature, and d) using $f_1, f_2, f_3, f_4, \text{and } f_5$ to build a sub-set of the H matrix for extracting features relevant to that row. This must be done to perform a one-to-one measurement of features, see Figure 6.

Now we calculate the similarities of the s_1-sorted H matrix against the sub-set feature matrix that contain the extracted relevant aligned matrix using the L_2 norm for discovering the frame for each document. This yields a list that contain the row index and value that corresponds to $f_1, f_2, f_3, f_4, \text{ or } f_5$. The $\text{minimization of} \left(s_1 - \left\{f_1, f_2, f_3, f_4, f_5\right\}\right)$ gives the frame that is most similar to the single signature, as shown in Figure 7.

Using the row index, we map the s_1 to $f_1, f_2, f_3, f_4, \text{ or } f_5$. Thereby the document is recognized as the frame candidate, and as such receives the class label of the frame signature selected; see Figure 8.

EVALUATION AND RESULTS

To examine the effectiveness of our news frame discovery approach trained by a small dataset size in the context of the issue of water insecurity, we perform the experiment with respect to varying parameters for NMF and the use of anchoring features found to be prominent in the learned vocabulary. The basic premise for applying the ground truth is that it acts as a quality measure for comparing the content analysis labels for a subset of the data to algorithm-generated hidden structures to understand similarities and boundary differences (Lombard et al., 2002). The ground truth label is a method of validation for the machine learning prediction.

Our frame mapping function successfully maps dominant frames to $f_1, f_2, f_3, f_4, \text{ or } f_5$. To prepare for our prediction, we assign each document in our training dataset a class label. We determine the assignment by calculating the distance measure of the TF-IDF matrix that is a derivation of the training set against $f_1, f_2, f_3, f_4, \text{ or } f_5$, as shown in Figure 9. To arrive at the same matrix dimensions for performing the distance measurements, we perform a similar process to that described in the frame mapping function. This step is a form of reverse engineering the part-based representation of H matrix back to the d_i corpus.

Figure 5. Sort both H matrices

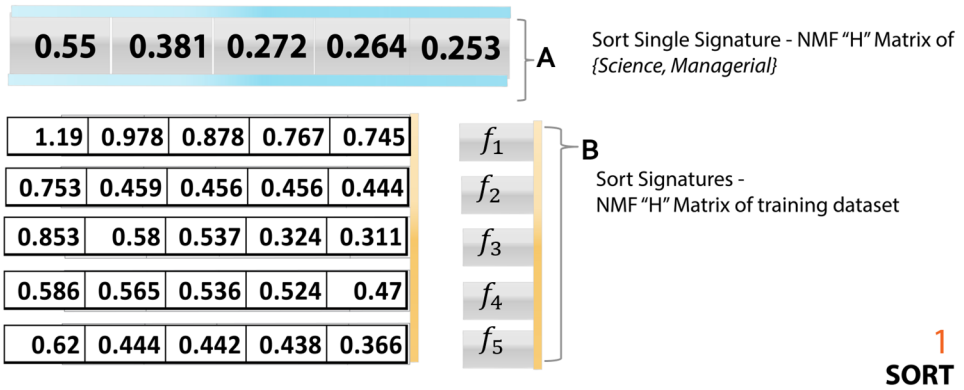


Figure 6. Align the single signature to the relevant features in preparation for distance measurement

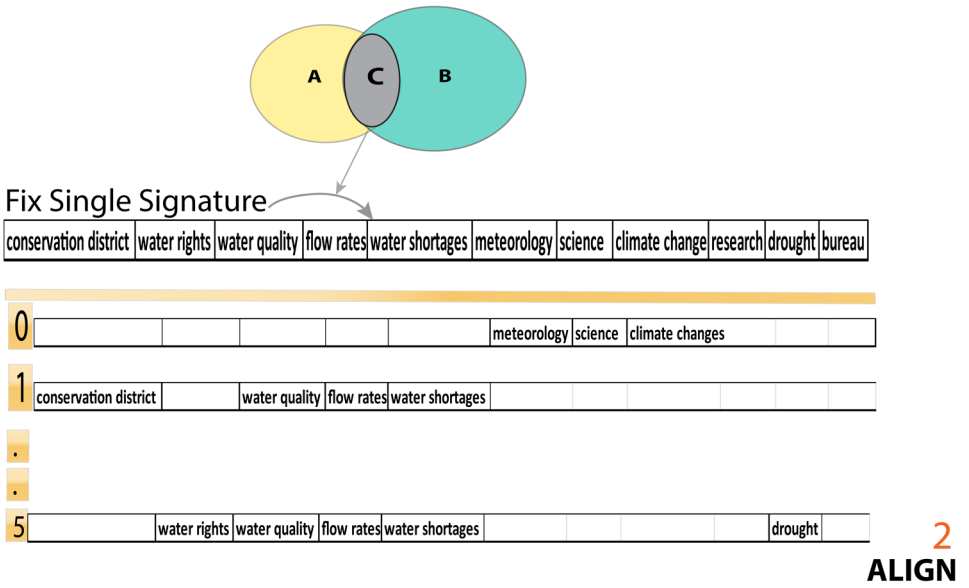
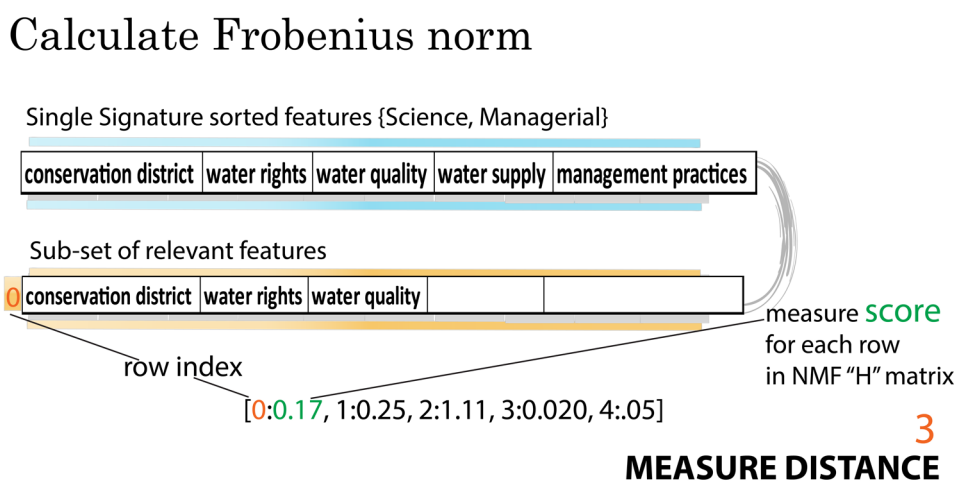


Figure 7. Calculate the distance measure



Our process is designed to handle five distinct frame signatures, our experiment is trained on the science and managerial signatures. Our frame discovery process is tested by a) observing a news article test document d_i, and b) deciding the most similar frame using the aforementioned frame mapping function. The process is trained on a subset of news articles (or documents) categorized as science frame during the content analysis, and the frame signature is tested by searching each document contained in the test dataset against the frame signature. The closest frame signature to the document is marked as machine learning frame type: science or managerial. The test data frame type is checked against the ground truth class label assigned to the document during the content analysis. Similarly, this training and testing is done for managerial frame signature. In the case when the frame signature measurement is the same—a tie—we make an arbitrary selection, thereafter we apply the selected frame signature class label.

Figure 8. Assign dominant frame to corresponding row index for discovering frame

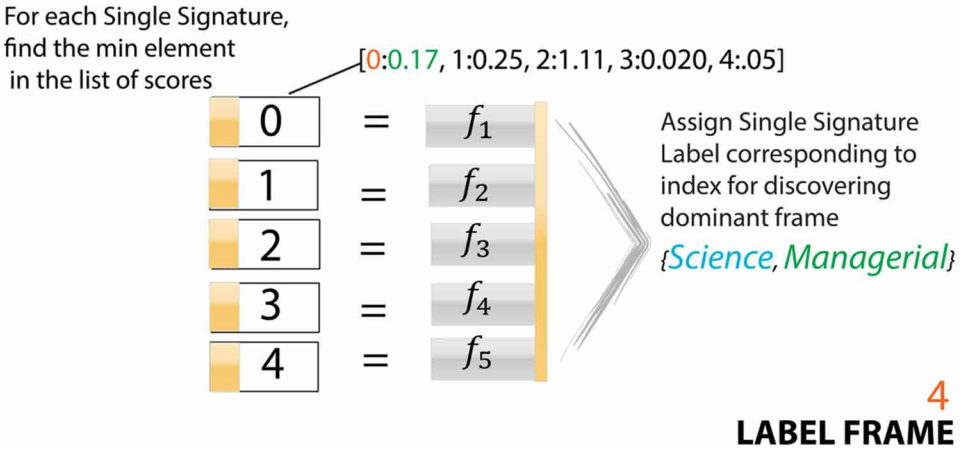
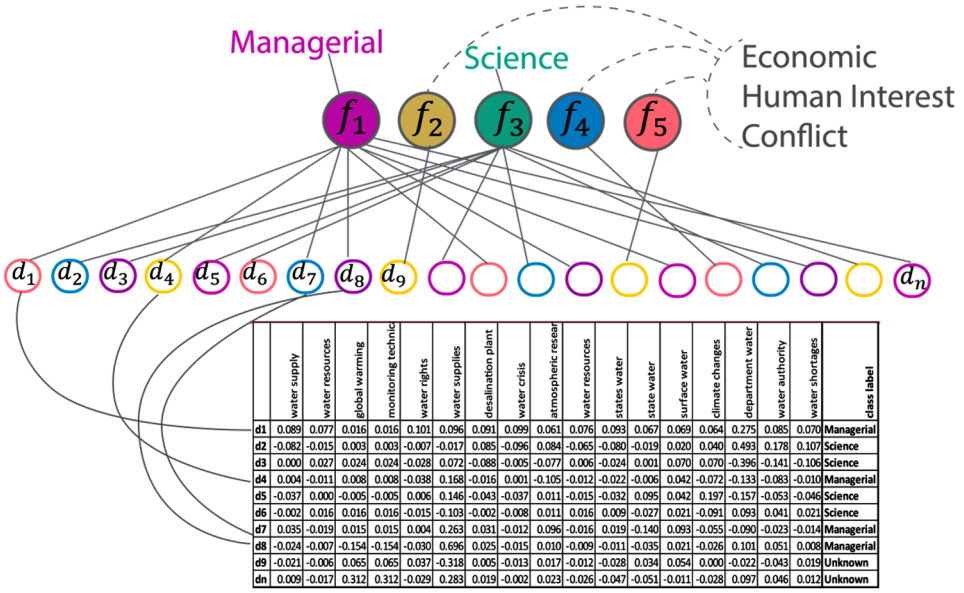


Figure 9. This depicts the assignment of dominant class label to training documents



	water supply	water resources	global warming	monitoring technic	water rights	water supplies	desalination plant	water crisis	atmospheric reser	water resources	states water	state water	surface water	climate changes	department water	water authority	water shortages	class label
d1	0.089	0.077	0.016	0.016	0.101	0.096	0.091	0.099	0.061	0.076	0.093	0.067	0.069	0.064	0.275	0.085	0.070	Managerial
d2	-0.082	-0.015	0.003	0.003	-0.007	-0.017	0.085	-0.096	0.084	-0.065	-0.080	-0.019	0.020	0.040	0.493	0.178	0.107	Science
d3	0.000	0.027	0.024	0.024	-0.028	0.072	-0.088	-0.005	-0.077	0.006	-0.024	0.001	0.070	0.070	-0.396	-0.141	-0.106	Science
d4	0.004	-0.011	0.008	0.008	-0.038	0.168	-0.016	0.001	-0.105	-0.012	-0.022	-0.006	0.042	-0.072	-0.133	-0.083	-0.010	Managerial
d5	-0.037	0.000	-0.005	-0.005	0.006	0.146	-0.043	-0.037	0.011	-0.015	-0.032	0.095	0.042	0.197	-0.157	-0.053	-0.046	Science
d6	-0.002	0.016	0.016	0.016	-0.015	-0.103	-0.002	-0.008	0.011	0.016	0.009	-0.027	0.021	-0.091	0.093	0.041	0.021	Science
d7	0.035	-0.019	0.015	0.015	0.004	0.263	0.031	-0.012	0.096	-0.016	0.019	-0.140	0.093	-0.055	-0.090	-0.023	-0.014	Managerial
d8	-0.024	-0.007	-0.154	-0.154	-0.030	0.696	0.025	-0.015	0.010	-0.009	-0.011	-0.035	0.021	-0.026	0.101	0.051	0.008	Managerial
d9	-0.021	-0.006	0.065	0.065	0.037	-0.318	0.005	-0.013	0.017	-0.012	-0.028	0.034	0.054	0.000	-0.022	-0.043	0.019	Unknown
dn	0.009	-0.017	0.312	0.312	-0.029	0.283	0.019	-0.002	0.023	-0.026	-0.047	-0.051	-0.011	-0.028	0.097	0.046	0.012	Unknown

Our initial experiment contradicted our expectations. This finding led to the re-training of our model using a subset of learned vocabulary for anchoring our frame signatures (frame types) with the original training dataset and using it as a feedback loop for improved frame classification with the original test dataset. We use the holdout method for evaluating the process for frame discovery: 1/3 data reserved for testing, 2/3 for training. The performance of our experiment is reported by using standard measures of recall, precision, recall and F1-score, as shown in Figure 10 (Buckley & Voorhees, 2000). Thus, one interpretation of our results is that our constructed model prediction of science and managerial frames has potential for extending traditional content analysis, thereby allowing for the analysis of content on large datasets. These findings expose efficiency gains and an interesting correlation between the inter-coder agreement and the machine learning of the system.

Figure 10. News frame discovery results

	PRECISION	RECALL	F1-SCORE
CLASS 0 (SCIENCE)	78%	82%	80%
CLASS 1 (MANAGERIAL)	75%	74%	75%

CONCLUSION

We present in this article the first formal computer science definition for framing, an approach for the discovery of five distinct patterns (referred to as signatures) that characterize prominent frames, and a process that advances machine learning for document classification that takes advantage of content analysis on a small number of documents for scaling up to meet the demands of large datasets. Preliminary experiments suggest the use of Non-negative Matrix Factorization (NMF) combined with Term Frequency–Inverse Document Frequency (TF-IDF) are promising approaches for discovering frames through the process of revealing latent relationships found in online news articles. Though research by computer scientists is underway in using communication frames to explore news discourse (Alashri et al., 2016; Ceran et al., 2015; Lu et al., 2016), these studies fail to answer a foundational question: how can framing be used in machine learning and in the context of computer science? Thus, our research fills an important gap in the computer science literature by providing a formal definition and process for discovering frames. Future work involves improving our model using deep learning methods. We intend to test further linked data and discriminant factors that maximize the similarity between the document representations and each frame signature. We also plan to compare our results with existing classification techniques to evaluate performance.

ACKNOWLEGMENT

This effort was partially supported by the Arizona State University Center for Policy Informatics, the Adobe Foundation, and the GEM Consortium. Special thanks to K. Hirsch for help with the preparation of this manuscript and to N.D. Taranto, H. Smith, and B. Fisher for help with content analysis coding.

REFERENCES

Alashri, S., Tsai, J.-Y., Alzahrani, S., Corman, S., & Davulcu, H. (2016). "Climate Change" Frames Detection and Categorization Based on Generalized Concepts. *Proceedings of the Tenth IEEE International Conference on Semantic Computing (ICSC-16)*, Laguna Hills, CA, USA (pp. 277-284).

An, S.-K., & Gower, K. K. (2009). How do the news media frame crises? A content analysis of crisis news coverage. *Public Relations Review*, *35*(2), 107–112. doi:10.1016/j.pubrev.2009.01.010

Baxendale, P. (1958). Machine-made index for technical literature: An experiment. *IBM Journal of Research and Development*, *2*(4), 354–361. doi:10.1147/rd.24.0354

Berendt, B., Hotho, A., & Stumme, G. (2002). Towards semantic web mining. In *The Semantic Web–ISWC 2002* (pp. 264–278). Springer Berlin Heidelberg. doi:10.1007/3-540-48005-6_21

Berger, A., Caruana, R., Cohn, D., Freitag, D., & Mittal, V. (2000). Bridging the Lexical Chasm: Statistical Approaches to Answer Finding.*Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 192-199). ACM. doi:10.1145/345508.345576

Bertsekas, D. P. (1999). *Nonlinear Programming*. Belmont, MA: Athena Scientific.

Bindela, D., Kleinberg, J., & Orenb, S. (2015). How bad is forming your own opinion? *Games and Economic Behavior*, *92*, 248–265. doi:10.1016/j.geb.2014.06.004

Bishop, G. F., Oldendick, R. W., & Tuchfarber, A. J. (1982). Effects of presenting one versus two sides of an issue in survey questions. *Public Opinion Quarterly*, *46*(1), 69–85. doi:10.1086/268700

Bittorf, V., Recht, B., Ré, E., & Troppy, J. A. (2012). Factoring nonnegative matrices with linear programs. In *Advances in Neural Information Systems* (pp. 1223–1231). NIPS.

Blei, D., Ng, A., & Jordan, M. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, *3*(Jan), 993–1022.

Boutsidis, C., & Gallopoulos, E. (2007). SVD based initialization: A head start for nonnegative matrix factorization. *Pattern Recognition*, *41*(4), 1350–1362. doi:10.1016/j.patcog.2007.09.010

Buckley, C., & Voorhees, E. M. (2000). Evaluating evaluation measure stability.*Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 33–40). ACM.

Cartwright, D. P. (1953). Analysis of qualitative material. In L. Festinger & D. Katz (Eds.), *Research Methods in the Behavioral Sciences* (pp. 421–470). New York: Dryden.

Ceran, B., Kedia, N., Corman, S. R., & Davulcu, H. (2015). Story Detection Using Generalized Concepts and Relations. *Proceedings of the2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)* (pp. 942-949). IEEE. doi:10.1145/2808797.2809312

Chung, D. S. (2008). Interactive features of online newspapers: Identifying patterns and predicting use of engaged readers. *Journal of Computer-Mediated Communication*, *13*(3), 658–679. doi:10.1111/j.1083-6101.2008.00414.x

Dave, K., Steve, L., & Pennock, D. M. (2003). Mining the peanut gallery: Opinion extraction and semantic classification of product reviews.*Proceedings of the 12th International Conference on World Wide Web* (pp. 519–528). ACM. doi:10.1145/775152.775226

Ding, C., Li, T., & Jordan, M. I. (2010). Convex and semi-nonnegative matrix factorizations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *32*(1), 45–55. doi:10.1109/TPAMI.2008.277 PMID:19926898

Donoho, D. L., & Stodden, V. (2004). *When does non-negative matrix factorization give a correct decomposition into parts?*Advances in Neural Information Processing Systems.

Druckman, J. N., & Bolsen, T. (2002). Framing, motivated reasoning, and opinions about emergent technologies. *Journal of Communication*, *61*(4), 659–688. doi:10.1111/j.1460-2466.2011.01562.x

Dumais, S., Furnas, G. W., Landauer, T. K., & Deerwester, S. (1988). Using latent semantic analysis to improve information retrieval.*Proceedings of the ACM Conference on Human Factors in Computing Systems* (pp. 281–285).

Edmundson, H. P. (1969). New methods in automatic extracting. *Journal of the ACM*, *16*(2), 264–285. doi:10.1145/321510.321519

Entman, R. M. (1993). Framing: Towards clarification of a fractured paradigm. *Journal of Communication*, *43*(4), 51–58. doi:10.1111/j.1460-2466.1993.tb01304.x

Fellbaum, C. (1998). A semantic network of English: the mother of all WordNets. In *EuroWordNet: A multilingual database with lexical semantic networks* (pp. 137–148). Springer Netherlands. doi:10.1007/978-94-017-1491-4_6

Gamson, W. A., & Modigliani, A. (1987). The changing culture of affirmative action. In R. G. Braungart & M. M. Braungart (Eds.), *Research in political sociology* (Vol. 3, pp. 137–177). Greenwich, CT: JAI Press.

Ghanem, M., Chortaras, A., Guo, Y., Rowe, A., & Ratcliffe, J. (2005). A Grid Infrastructure for Mixed Bioinformatics Data And Text Mining. *Proceedings of the 3rd ACS/IEEE International Conference on Computer Systems and Applications* (p. 41). IEEE. doi:10.1109/AICCSA.2005.1387038

Glaser, B. G., & Strauss, A. L. (1967). *The Discovery of Grounded Theory; Strategies for Qualitative Research*. Chicago: Aldine Publishing Company.

Gu, Q., & Zhou, J. (2009). *Neighborhood Preserving Nonnegative Matrix Factorization* (pp. 1–10). BMVC.

Hearst, M. A. (1997). Text data mining: Issues, techniques, and the relationship to information access. *Presentation notes for UW/MS Workshop on Data Mining*.

Holsti, O. R. (1969). *Content Analysis for the Social Sciences and Humanities*. Reading, MA: Addison-Wesley.

Hotho, A., Staab, S., & Stumme, G. (2003). Ontologies improve text document clustering.*Proceedings of the Third IEEE International Conference on Data Mining* (pp. 541–544). IEEE. doi:10.1109/ICDM.2003.1250972

Hoyer, P. O. (2004). Non-negative Matrix Factorization with Sparseness Constraints. *Journal of Machine Learning Research*, *5*(Nov), 1457–1469.

Hu, Q., Yu, D., Duan, Y., & Bao, W. (2003). A novel weighting formula and feature selection for text classification based on rough set theory.*Proceedings of the 2003 International Conference on Natural Language Processing and Knowledge Engineering* (pp. 638-645). IEEE.

Jiang, C., Coenen, F., Sanderson, R., & Zito, M. (2010). Text Classification Using Graph Mining-Based Feature Extraction. *Knowledge-Based Systems*, *23*(4), 302–308. doi:10.1016/j.knosys.2009.11.010

Karanikas, H., Tjortjis, C., & Theodoulidis, B. (2000) An approach to text mining using information extraction. *Proceedings of the Workshop Knowledge Management Theory Applications (KMTA 00)*.

Karlberg, A. (1997). News and conflict, How adversarial news frames limit public understanding of environmental issues. *Alternative Journals*, *23*(1), 22.

Kempe, D., Kleinberg, J., & Tardos, E. (2003). Maximizing the spread of influence through a social network. *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 137–146). ACM. doi:10.1145/956750.956769

Krippendorff, K. (1980). *Content Analysis: An Introduction to its Methodology*. Beverly Hills, CA: Sage.

Krosnick, J. A. (1999). Survey research. *Annual Review of Psychology*, *50*(1), 537–567. doi:10.1146/annurev.psych.50.1.537 PMID:15012463

Kushal, D., Lawrence, S., & Pennock, D. (2003). Mining the peanut gallery: Opinion extraction and semantic classification of product reviews.*Proceedings of the 12th international conference on World Wide Web*, Budapest, Hungary (pp. 519−528). ACM.

Lee, D. D., & Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, *401*(6755), 788–791. doi:10.1038/44565 PMID:10548103

Liu, H., & Singh, P. (2004). Commonsense reasoning in and over natural language. *Proceedings of theInternational Conference on Knowledge-Based and Intelligent Information and Engineering Systems* (pp. 293–306). Springer Berlin Heidelberg. doi:10.1007/978-3-540-30134-9_40

Liu, H., & Wu, Z. (2010). Non-negative matrix factorization with constraints.*Proceedings of the 24th AAAI Conference on Artificial Intelligence* (pp. 506–511).

Lombard, M., Snyder-Duch, J., & Bracken, C. C. (2002). Content analysis in mass communication research: An assessment and reporting of intercoder reliability. *Human Communication Research*, *28*(4), 587–604. doi:10.1111/j.1468-2958.2002.tb00826.x

Lu, Y., Steptoe, M., Burke, S., Wang, H., Tsai, J., Davulcu, H., & Maciejewski, R. et al. (2016). Exploring evolving media discourse through event cueing. *IEEE Transactions on Visualization and Computer Graphics*, *22*(1), 220–229. doi:10.1109/TVCG.2015.2467991 PMID:26529702

Luhn, H. P. (1958). The automatic creation of literature abstracts. *IBM Journal of Research and Development*, *2*(2), 159–165. doi:10.1147/rd.22.0159

Luo, C., Li, Y., & Chung, S. M. (2009). Text document clustering based on neighbors. *Data & Knowledge Engineering*, *68*(11), 1271–1288. doi:10.1016/j.datak.2009.06.007

Maron, M. E., & Kuhns, J. L. (1960). On Relevance, Probabilistic Indexing and Information Retrieval. *Journal of the ACM*, *7*(3), 216–244. doi:10.1145/321033.321035

Murphy, P. (2001). Affiliation bias and expert disagreement in framing the nicotine addiction debate. *Science, Technology & Human Values*, *26*(3), 278–299. doi:10.1177/016224390102600302

Raghavan, P., Amer-Yahia, S., & Gravano, L. (2004). Structure in Text: Extraction and Exploitation. *Proceedings of the 7th International Workshop on the Web and Databases (WebDB)*. ACM SIGMOD/PODS.

Riffe, D., Lacy, S., & Fico, F. G. (2005). *Analyzing Media Messages: Using Quantitative Content Analysis in Research*. New York, NY: Routledge.

Salton, G., & Buckley, C. (1988). Term-weighing approaches in automatic text retrieval. *Information Processing & Management*, *24*(5), 513–523. doi:10.1016/0306-4573(88)90021-0

Salton, G., Wong, A., & Yang, C. S. (1975). A vector space model for automatic indexing. *Communications of the ACM*, *18*(11), 613–620. doi:10.1145/361219.361220

Scheufele, D. A. (1999). Framing as a theory of media effects. *Journal of Communication*, *49*(1), 103–122. doi:10.1111/j.1460-2466.1999.tb02784.x

Scheufele, D. A., & Tewksbury, D. (2007). Framing, agenda setting, and priming: The evolution of three media effects models. *Journal of Communication*, *57*(1), 9–20.

Seung, D., & Lee, L. (2001). *Algorithms for non-negative matrix factorization* (pp. 556–562). Advances in Neural Information Processing Systems.

Severin, W. J., & Tankard, J. W. (1997). *Communication Theories: Origins, Methods, and Uses in the Mass Media* (4th ed.). New York, NY: Longman Publishing Group.

Shah, D., Watts, M. D., Domke, D., & Fan, D. P. (2002). News framing and cueing of issue regimes: Explaining Clintons public approval in spite of scandal. *Public Opinion Quarterly*, *66*(3), 339–370. doi:10.1086/341396

Spärck Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. *The Journal of Documentation*, *28*(1), 11–21. doi:10.1108/eb026526

Stavrianou, A., Andritsos, P., & Nicoloyannis, N. (2007). Overview and semantic issues of text mining. *SIGMOD Record*, *36*(3), 23–34. doi:10.1145/1324185.1324190

Steinheiser, R., & Clifton, C. (1998). Data Mining on Text.*Proceedings of the 22nd Annual IEEE International Computer Software and Applications Conference*.

Tan, A., Zhong, N., & Zhou, L. (2000). Text Mining: The state of the art and the challenges.*Proceedings of the PAKDD 1999 Workshop on Knowledge Disocovery from Advanced Databases* (Vol. 8, pp. 65–70).

Tan, S., Cheng, X., Wang, B., Xu, H., Ghanem, M. M., & Guo, Y. (2005). Using dragpushing to refine centroid text classifiers.*Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 653–654). ACM. doi:10.1145/1076034.1076174

Tankard, J. (2001). The empirical approach to the study of media framing. In S. D. Reese, O. H. Gandy Jr, & A. E. Grant (Eds.), *Framing Public Life: Perspectives on Media and Our Understanding of the Social World* (pp. 95–106). Mahwah, NJ: Erlbaum.

Tewksbury, D., & Scheufele, D. A. (2009). News framing theory and research. In J. Bryant & M. B. Oliver (Eds.), *Media Effects: Advances in Theory and Research* (pp. 17–33). Hillsdale, NJ: Erlbaum.

Loretta H. Cheeks is pursuing a Computer Science PhD in Fulton School of Engineering/CIDSE at Arizona State University. Prior to that, she developed large scale systems & led development teams within the communications, radio, avionics, instrumentation & control and chemical industries. She received BS and MS in Computer Science from Southern University, Baton Rouge, LA. Her research and collaborations aim to discover methods for machine learning and influence dynamics that will enable communities of interest to effectively make decisions about critical socio-environmental resources.

Tracy L. Stepien is a Postdoctoral Research Associate in the Department of Mathematics at the University of Arizona. Prior to that, she was a Visiting Assistant Professor in the School of Mathematical and Statistical Sciences at Arizona State University. She received her BS in Mathematics from the University at Buffalo, The State University of New York in 2008 and her MA in 2010 and PhD in 2013 in Mathematics from the University of Pittsburgh. Her research interests include mathematically modeling physical systems with an emphasis on biological systems, dynamical systems, ordinary and partial differential equations, and related numerical methods.

Dara M. Wald is an assistant professor in environmental communication and sustainability in the Greenlee School of Journalism and Communication, Iowa State University. Prior to this, she was a postdoctoral fellow in the Center for Policy Informatics and the Decision Center for a Desert City, Arizona State University. She received her PhD from the Department of Wildlife Ecology and Conservation, University of Florida, where she was selected as an NSF-GK12 Fellow and awarded an Alumni Graduate Fellowship, the highest graduate student award available at UF. Her research focuses on the causes and consequences of environmental conflict and collaboration, with an emphasis on how emotions, risk perceptions, and communication drive socio-political conflict and policy resistance.

Ashraf Gaffar is an Assistant Professor in the Fulton School of Engineering/CIDSE at Arizona State University. He Received his PhD in Computer Science/Software Engineering from Concordia University, Montreal, Canada, where he designed and built a framework to model and analyze extremely large sets of unstructured data. He received several academic awards, including "Superior Academic Achievement". Gaffar comes with rich industrial experience in software design, data visualization and business intelligence. After receiving his PhD, he worked at SAP as a Senior Design Expert, where he helped design and develop complex software products used by most Fortune 500 companies. He was also instrumental in developing new design guidelines for SAP mobile business applications and received several prestigious industrial rewards for his work. He authored or co-authored over 40 conference and journal papers as well as book chapters. His research focuses on human cognitive abilities to analyze big data, human centered design, and usability engineering. He is interested in developing new analytical and visualization methods to increase comprehension of big data.