Discovering News Frames: Exploring Text, Content, and Concepts in Online News Sources to Address Water Insecurity in the Southwest Region

Loretta H. Cheeks[1], Tracy L. Stepien[2], Dara M. Wald[3]

[1]Department of Computing, Informatics and Decision Systems Engineering,
Arizona State University, Tempe, Arizona, U.S.A., loretta.cheeks@asu.edu

[2]School of Mathematical and Statistical Sciences,
Arizona State University, Tempe, Arizona, U.S.A., tstepien@asu.edu

[3] Greenlee School of Journalism & Communication,
Iowa State University, Ames, Iowa, U.S.A., dwald@iastate.edu

Abstract

The Internet is a major source of online news content. Current efforts to evaluate online news content, including text, story line 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. Our experimental evaluation shows that the proposed process is an effective and efficient semi-supervised machine learning method to inform data mining for inferring classification.

Keywords – Text Mining; Non-Negative Matrix Factorization; News Frames; Content Analysis; Clustering; Data Mining

1. INTRODUCTION

Approximately 80% of all data today exists in unstructured formats (e.g., news, e-mails, social media feeds, contracts, memos, clinical notes, and legal briefs) [28]. Text mining, also known as text data mining [14], is a multidisciplinary field, involving information retrieval, text analysis, information extraction, clustering, categorization, visualization, database technology, machine learning, and data mining. The ability to unlock 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. Text mining algorithms, techniques and methodologies offer a value to the data integration tasks: they inform the similarities between heterogeneous sources and text features, which reduce uncertainty and risk exposure when performing the integration tasks.

Text mining, coupled with an interest in understanding social influence and the information diffusion for document summarization (i.e., topic modeling, sentiment analysis, and opinion mining), is an active area of research [37], [38]. "The formation and transmission of group standards, values, attitudes, and skills are accomplished largely by means of verbal communication" [39]. Thus, efforts to understand social interaction, cooperation and influence require the study of text. Given the widespread use of unstructured text, such as email, online news and social media feeds, for individual and mass communication, it is particularly important to understand how social influence and information about complex socio-environmental issues is transmitted through online content.

The effect of media communication and specifically, the influence of the words and frames the media uses on public perceptions of social and environmental issues has been studied extensively. Framing refers to the strategic use of keywords, concepts and sentences for the purpose of conveying latent meanings about an issue [40]. Framing involves the selection, organization or presentation of key ideas or concepts within a story [41]. The framing of news stories can shape public interpretation of social and environmental news [42], [44].

One traditional method of exploring how text and media framing can influence social perceptions is content analysis. Content analysis, which draws on the concept of grounded theory [45], was developed specifically to aid in

978-1-5090-3207-5/16 $31.00 © 2016 IEEE
DOI 10.1109/IRI.2016.67

the interpretation of social discourse or text [46] for communication research [47]. Content analysis involves methodical evaluation and categorization of text [48]. Current approaches to content analysis, however, require scholars and researchers to comb through documents in search of patterns in the text. This approach to text analysis is dependent on humans and limits the applicability of this approach for the analysis of large-scale unstructured text.

The widespread availability of large data repositories, like those found in online news articles, creates an opportunity to develop new methods of text mining using 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 classification and clustering. This research aligns well with concurrent efforts in machine learning and data mining that seek to discover novel patterns and latent relationships in unstructured text for deep learning [5]. This method has several improvements over traditional methods of text mining: First, unlike the traditional approach to processing unstructured texts by focusing on the characterization of form (syntactic) and its meaning (semantic meaning) [32], this method views news texts as organized symbolic devices (frames) that act as carriers or signed vehicles (or cues) influencing perceptions about an issue that will interact with individuals' existing beliefs (i.e., other senders or receivers of news articles). Second, by placing framing within computer science, this approach allows for 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 [26], [31] 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, 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.

2. 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 Hans Peter Luhn [49] that describes work being done at IBM in the 1950s. In his work, Luhn proposed that the frequency of a particular word provides a useful measure of its significance. Luhn's contribution is the identification of the concept *term frequency* (TF), which says it is possible to identify the significant terms just based on the term frequency calculated within that document and relates to average information, or entropy, of a term or group of terms ranking in relationship to each other.

In the same year of the Luhn article publication, Baxendale [50] 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 and Kuhns [51] published "On Relevance Probalitic Indexing and Information Retrieval". This paper was the first on ranked retrieval and ranking of documents by their computed values of probability of relevance, which is important because two-valued indexing of documents could be replaced by weighted indexing, where the weights were to be interpreted as probabilities.

Edmundson [52] was the first to describe a system that produces document extracts. His primary contribution was the development of a typical structure for an extractive summarization experiment that integrates features of word frequency and positional importance, borrowed from the works of Luhn [49] and Baxendale [50].

Critical to text summarization is the contribution of Gerard Salton [29], credited with developing the vector space model (VSM). Extending the work of Luhn, Salton's contribution for describing similarities (and dissimilarities) computed using both extracted words and cited data remains a principle concept upon which many scholarly research leverages. According to Salton, a retrieval model represents documents, description features (such as index terms), queries, and the relationships within and across those sets.

Meanwhile, in 1972, Karen Spärck Jones published in the Journal of Documentation a paper called "A statistical interpretation of term specificity and its application in retrieval" [18]. Her contribution proved to be a giant leap in text summarization and information retrieval. She is credited with developing the measure of term specificity that later became known as *inverse document frequency*, or IDF; 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, and the measure was a heuristic implementation of this intuition.

Of Susan Dumais' [53] research contributions, the one that aligns with this research study is her improvement to 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 [29] and models construction [34], [33], [35]; data dimensions reduction research in feature extraction [20], [17]; research on mining algorithm of text classification [35], [19] and clustering [15], [25]; and deep semantic mining based on natural language process [3], [54].

Frame discovery is motivated by framing theory, as known in communications research, which focuses on understanding the latent meanings of observable messages in their contexts (e.g., [55]), 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; [53]), Social Network Analysis (SNA), which focuses on the importance of relationships among interacting units [57], [56].

3. 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 use, water policy and water decisions are framed affects water rights allocations, policy decisions, human consumptions, emerging technologies, farming techniques, and agricultural outcomes [61]. The issue of water insecurity in the Southwest Region is particularly important in that the seven states that make up the Southwest Region (Colorado, New Mexico, Utah, Wyoming, Nevada, Arizona and California) 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.

When online news sources use strategic devices for presenting prominent aspects and perspectives about an issue, using certain keywords as well as stereotyped images and sentences for the purpose of conveying latent meanings about an issue, it is called framing [40]. Public opinion, attitudes, beliefs, and behaviors can be influenced by how an issue is framed, particularly when framing comes from elites who have the power to go beyond persuasion to manipulation [42]. 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 [64] that are published online. Moreover, 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 [62] or respondents with reduced exposure to or interest in an issue [63]; 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 will describe the methods, an experimental evaluation, and a discussion of our results.

4. METHODS

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

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. Namely a) the news articles being published by online news sources, b) the news source that generate the articles, c) the frequency of news publications by sources overtime, and d) the salient language contained in the article.

4.1 Dataset Description

The data is comprised of a set of news articles published on the Internet by online news sources within a defined time period. 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 gathering online news articles that contain pertinent features needed for this research study. The URL seeds are used as input to the crawler.

Our researchers wrote a Python program for crawling news articles using Google News online search engine for dates ranging from January 2006 to March 2015. 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, the articles 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 is 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.

4.2 Frame Discovery Methods

News frames are a central organizing idea for news that supplies a context and highlights key issues and ideas by selecting, emphasizing and elaborating on particular perspectives and/or excluding others [59]. Using this definition, we experimented with existing techniques and algorithms for text mining.

Scholarship exists for text mining for novel knowledge discovery of patterns as mentioned above. However, this is the first known research to apply text mining for discovering communication frames found in online news articles. Moreover, a discovery of frames may expose clues of influence that shed light on the intent of creating discourse about the issue central to the online news article and how adversarial news frames limit public understanding of environmental issues, such as climate change effects, water insecurity, agricultural insecurity, health disparities, civil unrest, and death [60].

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. 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) 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 [58]. For the issue of water insecurity, we use standard frame types [2], in particular, human interest, conflict, economic, managerial, and science. Let $f_1, f_2, f_3, f_4,$ and f_5 correspond to each of these frames, where for 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 will depend on the quality of the vector space of documents $\{d_i\}$. Our study applied three methods for showing the quality of the vector space and evidencing machine-discovery accuracy. The techniques applied are 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. This is 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.

4.2.1 Content Analysis.

In order to explore the relationship between online news source framing, our team of researchers 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 [48]. 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 [48]. Content analysis is used to measure dimensions (features) of content of groups of text, so a study must identify a sample of texts. The term *coding* is the processing 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.

We followed the Lacy, Robinson, and Riff [48] method 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. Our data collection process occurred in phases. In early 2015, when our content analysis began the data crawling resulted in 13,000 articles. According to [47], a selection of approximately 700 of the collected articles would allow reliable inferences about the content of the online news articles. A total of 1000 randomly selected news articles were extracted from our news dataset as our sample size, which is over the suggested number. In the summer 2015, we were able to continue our data collection process, resulting in a dataset of 55,000 online news articles. Because our study involves machine learning (described in subsequent section) and the expectation to learn and thereby discover frames from a small dataset, we selected 350 news articles from our dataset for training and testing. Since our approach to text mining is new, we compared the tests of our machine learning approach with a more traditional approach to frame identification: content analysis. To validate our approach, we explored the sample articles with both content analysis and machine learning. We then compared the results for classification accuracy.

The content analysis involved the following steps: A definition of our overarching research problem was defined: "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, 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 needed to address 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 as No=1; Yes=2. There were a total of # questions designed to identify frame types, key actors and discourse over time. All the questions addressed here were categorical with No=1 and Yes=2 response options.

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), we selected 90 articles 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. We used SPSS statistical software for all the steps in the content analysis. Analysis showed five dominant frame types—human interest, managerial, science, economic, and conflict.

4.2.2 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 has 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 interest is to discover not only how meanings of subsequent terms, sentences, and paragraphs are related but also how the facts that these sentences refer to are related. Our news frame discovery process is composed of three main

steps: data transformation, frame mapping, and classification.

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, and 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. Our TF-IDF baseline model is used as input to the Non-negative Matrix Factorization (NMF) [22]. NMF has properties attractive for this class of problem, namely, dimensionality reduction—under the assumption of linearity and non-negativty, clustering, and local minimum convergence [65]. 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 model is that documents that are "close together" in space are similar in meaning. The TF-IDF score (or ranking) of a term in a set of documents is calculated as follows:

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

$$tf_{dc} = \frac{n_{dc}}{m_d}, \quad idf_c = \log \frac{\{d_i\}}{|\{d_i \in \{d_i\}: c_i \in d_i\}|} + 1$$

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. $\{d_i\}$ is the total number of documents in the corpus, and $|\{d \in \{d_i\} : c_i \in d_i\}|$ 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. 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 vectorization produced through the calculation of TF-IDF is a high-dimensional vector, $\{d\} \in \mathbb{R}^{d \times c}$.

For the news article dataset, we use the TF-IDF data matrix w_i as input to the NMF matrix A that represents $d_i, \cdots d_n$ documents, such that each column corresponds to the features vector x_i from frame f_i. We use the sparse NMF extension by [16], using SVD initialization (i.e. NNDSVD) [6] for addressing the convergence problem, which is common in many clustering algorithms (e.g., K-means or LDA). 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. NMF of a matrix A decomposes it into two matrices W and H such that

$$A = WH$$

If A has dimension of $d \times c$, 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:

$$F = \| A - WH^T \|_F^2$$
$$\text{such that } W = [w_{d \times r}] \geq 0$$
$$H = [h_{r \times c}] \geq 0$$

The cost function is convex only in W or H, but not convex in both together. The gradient descent contribution of Hoyer [16] extended Lee and Seung's [30] 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) we normalize for different document lengths using L_2norm. NMF uses a) a Nonnegative Double Singular Value Decomposition (NNDSVD) [6] seed, a method designed to enhance the initialization stage of NMF, b) the rank of the decomposition is 5, and c) the NMF updates iterates 50 times before timing out.

Frame Mapping. The frame mapping function is a two-step process that comprises a) aligning the features represented as columns in the TF-IDF matrix D and calculating the L_2 norm. The feature names associated with w_i for document d_i in the TF-IDF matrix $\{d_i\}$ are sorted against each frame $f_1, f_2, f_3, f_4,$ and f_5 signature, which must remain stationary. Each frame signature can be mapped to a document label applied during the content analysis categorization of dominate 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. Now we calculate the similarities of $\{d_i\}$ and H frame signature matrix using the Frobenius norm, commonly referred to as the L_2norm, for discovering the frame for each document. The minimization of (d_i- f_1, f_2, f_3, f_4, and f_5) gives the frame that is most similar to the document. Thereby, the document is recognized as the frame candidate, and as such receives the class label of the frame signature selected.

5 EVALUATION AND RESULTS

Though 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 as follows: a) it observes a news article test document, d_i and b) decides 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. 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 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 1. 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 exposes efficiency gains and an interesting correlation between the inter-coder agreement and the machine learning of the system.

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

Fig. 1. News frame discovery results.

6 CONCLUSION

We present in this article a) the first 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. Preliminary experiments suggest the use of Non-negative matrix factorization (NMF) combined with Term Frequency–Inverse Document Frequency (TF-IDF) are promising for discovering frames through the process of revealing latent relationships found in online news articles. Though research by computer scientist is underway in using communication frames to explore news discourse [9], [1], [24], those studies fail to answer a foundational question, which is how framing can be used in machine learning and in the context of computer science. Thus, this research fills an important gap in the computer science literature by providing a formal definition and process for discovering frames. Over the course of the next month, we will improve our model using deep learning methods. We intend to test further linked data and discriminant factors that maximizes the similarity between the document representations and each frame signature. In future research, we will compare our results with existing classification techniques to evaluate performance.

7 REFERENCES

[1] S. Alashri, J-Y Tsai, S. Alzahrani, S. Corman, H. Davulcu, "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, 2016.
[2] S-K An, K. K. Gower, “How do the news media frame crises? A content analysis of crisis news coverage”, Public Relations Review 35, 107–112, 2009.
[3] B. Berendt, A. Hotho, G. Stumme, "Towards semantic web mining", The Semantic Web—ISWC 2002, Springer Berlin Heidelberg, 2002. 264-278.
[4] V. Bittorf , B. Recht , E. Re′, J.A., Troppy (2013), “Factoring nonnegative matrices with linear programs”, in Advances in Neural Information Systems (NIPS ′12), 1223-1231, 2012.
[5] D. Blei, A. Ng, and M. Jordan, “Latent Dirichlet allocation,” Journal of Machine Learning Research, vol. 3, pp. 993–1022, 2003.
[6] C. Boutsidis, E. Gallopoulos, “SVD based initialization: A head start for nonnegative matrix factorization”, Pattern Recognition, 41:1350-1362, 2007.
[7] C. Buckley and E. M. Voorhees. Evaluating evaluation measure stability. In ACM SIGIR’00, pages 33 – 40, Athens, Greece, ACM Press, 2000.
[8] D. Cai, X. He, J. Han, and T. S. Huang, “Graph regularized nonnegative matrix factorization for data representation,” Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 33, no. 8, pp. 1548–1560, 2011.
[9] B. Ceran, N. Kedia, S.R. Corman, H. Davulcu, “Story Detection Using Generalized Concepts and Relations”, Proceedings of International Symposium on Foundation of Open Source Intelligence and Security Informatics (FOSINT-SI), in conj. with IEEE ASONAM 2015, Paris, France, 2015.
[10] C. Ding, T. Li, M. I. Jordan, “Convex and semi-nonnegative matrix factorizations”, IEEE Transactions on Pattern Analysis and Machine Intelligence, 32(1), 45-55. [4685898]. 10.1109/TPAMI.2008.277, 2010
[11] D. L. Donoho, V. Stodden, “When does non-negative matrix factorization give a correct decomposition into parts?”, in Proceedings of the NIPS, MIT Press, 2004.
[12] M. Ghanem, A. Chortaras, Y. Guo, A. Rowe, J. Ratcliffe, "A Grid Infrastructure For Mixed Bioinformatics Data And Text Mining", Computer Systems and Applications, 2005. The 3rd ACS/IEEE International Conference on (Vol.29, pp.41-I), 2005.
[13] Q. Gu and J. Zhou, “Neighborhood Preserving Nonnegative Matrix Factorization,” in BMVC, pp. 1–10, 2009.
[14] M. A. Hearst, “Text data mining: Issues, techniques, and the relationship to information access”, Presentation notes for UW/MS workshop on data mining, July 1997.
[15] Hotho, Andreas, Steffen Staab, and Gerd Stumme. "Ontologies improve text document clustering." Data Mining, 2003. ICDM 2003. Third IEEE International Conference on. IEEE, 2003.
[16] P. O. Hoyer, “Non-negative Matrix Factorization with Sparseness Constraints”, Journal of Machine Learning Research 5, 1457-1469, 2004.
[17] Hu, Qinghua, et al. "A novel weighting formula and feature selection for text classification based on rough set theory." Natural Language Processing and Knowledge Engineering, 2003. Proceedings. 2003 International Conference on IEEE, 2003:638-645.
[18] K. S. Jones, “A statistical interpretation of term specificity and its application in retrieval”, Journal of Documentation, Vol. 28, pp. 11–21, 1972.
[19] C. Jiang, F. Coenen, R. Sanderson, M. Zito, “Text Classification Using Graph Mining-Based Feature Extraction”, Knowledge-Based Systems, vol. 23, no. 4, pp. 302-308, 2010.
[20] H. Karanikas, C. Tjortjis, and B. Theodoulidis. "An Approach to Text Mining using Information Extraction", Proc. Workshop Knowledge Management Theory Applications (KMTA 00), 2000.
[21] D. Kushal, S. Lawrence, and D. Pennock, "Mining the peanut gallery: Opinion extraction and semantic classification of product reviews", Proceedings of the 12th international conference on World Wide Web, ACM, Budapest, Hungary, 2003, 519−528.
[22] D. D. Lee, H. S. Seung, “Learning the parts of objects by non-negative matrix factorization,” Nature, vol. 401, no. 6755, pp. 788–791, 1999.
[23] H. Liu, Z. Wu, “Non-negative matrix factorization with constraints,” in Twenty-Fourth AAAI Conference on Artificial Intelligence, 2010.
[24] Y. Lu, M. Steptoe, S. Burke, H. Wang, J. Tsai, H. Davulcu, D. Montgomery, S.R. Corman, R. Maciejewski, “Exploring Evolving Media Discourse Through Event Cueing”, IEEE Transactions on Visualization and Computer Graphics, 22(1): 220-229, 2016.
[25] C. Luo, Y. Li, S. M. Chung. "Text document clustering based on neighbors" Data & Knowledge Engineering, 68.11:1271-1288, 2009.
[26] P. Murphy, “Affiliation bias and expert disagreement in framing the nicotine addiction debate”, Science, Technology, & Human Values, 26, 278–299, 2001.
[27] Oracle Corporation, WWW, www.oracle.com, 2008.
[28] P. Raghavan, S. Amer-Yahia and L. Gravano eds., “Structure in Text: Extraction and Exploitation”, In. Proceeding of the 7th international Workshop on the Web and Databases(WebDB), ACM SIGMOD/PODS 2004, ACM Press, Vol 67, 2004.
[29] Salton, G., A. Wong, and C. S. Yang. "A vector space model for automatic indexing." In Communications of the ACM 18 (11), 1975: 613-620.
[30] D. Seung and L. Lee, “Algorithms for non-negative matrix factorization,” Advances in neural information processing systems, vol. 13, pp. 556–562, 2001.
[31] D. Shah, M. D. Watts, D. Domke, D. P. Fan, “News framing and cueing of issue regimes: Explaining Clinton’s public approval in spite of scandal”. Public Opinion Quarterly, 66(3), 339–370, 2002.
[32] A. Stavrianou, P. Andritsos, N. Nicoloyannis, “Overview and Semantic Issues of Text Mining”, Special Interest Group Management of Data (SIGMOD) Record, (Vol. 36, No. 3), September 2007.
[33] Steinheiser, R., and C. Clifton. "Data Mining on Text." 2012 IEEE 36th Annual Computer Software and Applications Conference IEEE Computer Society, 1998:630.
[34] A. Tan, N. Zhong and L. Zhou, "Text Mining: The state of the art and the challenges." Proceedings of the PAKDD Workshop on Knowledge Discovery and Data Mining, 65—70, Beijing, China 2000.
[35] S. Tan, X. Cheng, B. Wang, H. Xu, M. M. Ghanem, Y. Guo, “Using dragpushing to refine centroid text classifiers”, In Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '05). ACM, New York, NY, USA, 653-654. 2005.

[36] D. Tewksbury, D. A. Scheufele, News framing theory and research. In: Bryant J, Oliver MB Media effects: Advances in theory and research. 3rd ed. Hillsdale, NJ: Erlbaum ; pp. 17-33, 2009.
[37] D. Bindela, , J. Kleinberga, , S. Orenb, "How bad is forming your own opinion?", Games and Economic Behavior, Volume 92, Pages 248–265, July 2015.
[38] D. Kempe, J. Kleinberg, and ´ E. Tardos, "Maximizing the spread of influence through a social network," in *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, pp. 137–146, 2003.
[39] D. P. Cartwright, "Analysis of qualitative material", In *L. Festinger & D. Katz (Eds.), Research Methods in the Behavioral Sciences* (pp. 421-470). New York: Dryden, 1953.
[40] R. M. Entman, "Framing: Towards clarification of a fractured paradigm", *Journal of Communication*, *43*(4), 51-58, 1993.
[41] W. A. Gamson, A. Modigliani, The Changing Culture Of Affirmative Action. *Research in Political Sociology*, 3, 137–177, 1987.
[42] J. N. Druckman, T. Bolsen. Framing, motivated reasoning, and opinions about emergent technologies. Journal of Communication, 2002.
[43] D. A. Scheufele, D. Tewksbury, Framing, agenda setting, and priming: The evolution of three media effects models. *Journal of Communication 57 (1)*:9-20, 2007.
[44] Tewksbury, D., and D. A. Scheufele, News framing theory and research. In *Media effects: Advances in theory and research*, edited by J. Bryant and M. B. Oliver. Hillsdale, NJ: Erlbaum, 2009.
[45] B. G. Glaser, A. L. Strauss, The Discovery of Grounded Theory; Strategies for Qualitative Research, *Chicago: Aldine Pub. Co.*, 1967.
[46] O. R. Holsti, Content Analysis for the Social Sciences and Humanities*, Reading, MA: Addison-Wesley*, 1969.
[47] K. Krippendorff, Content analysis: An introduction to its methodology. Beverly Hills, CA: *Sage*, 1980.
[48] D. Riffe, S. Lacy, F. G. Fico, Analyzing Media Messages: Using Quantitative Content Analysis in Research, New York, NY: *Routledge*, 2005.
[49] H. P. Luhn. "The Automatic Creation of Literature Abstracts," IBM Jor~rrzul of Research and Development, 2, No. 2, 159 (April 1958).
[50] P. Baxendale, Machine-made index for technical literature - an experiment, *IBM Journal of Research Development*, 2(4):354–361, 1958.
[51] M. E. Maron, J. L. Kuhns. "On Relevance, Probabilistic Indexing and Information Retrieval", *Journal of the ACM*, 216-244, 1960.
[52] H.P. Edmundson, New methods in automatic extracting, *Journal of the ACM*, 16(2), 264–285, 1969.
[53] S. Dumais, G.W. Furnas, T.K. Landauer, S. Deerwester, Using latent semantic analysis to improve information retrieval. In *Proceedings of CHI'88: Conference on Human Factors in Computing*, New York: ACM, 281-285, 1988.
[54] K. Dave, L. Steve, D. M. Pennock. "Mining the peanut gallery: Opinion extraction and semantic classification of product reviews", *Proceedings of the 12th international conference on World Wide Web*. ACM, 2003.
[55] D. A. Scheufele, "Framing as a theory of media effects", *The Journal of Communication*, 49(1), 103-122, 1999.
[56] H. Liu, P. Singh, Focusing on conceptnet's natural language knowledge representation. In *Proc. Of the 8th Intl Conf. on Knowledge-Based Intelligent Information and Engineering Syst*, 2004.
[57] C. Fellbaum, "A Semantic Network of English: The Mother of all WordNets," in: *Computers and the Humanities* 32: 209-220. 1998.
[58] Tankard, J. (2001). The empirical approach to the study of media framing. In *S. Reese, O. Gandy, & A. Grant (Eds.),* Framing public life (pp. 95–106). Mahwah, NJ: Erlbaum.
[59] W.J. Severin, J.W. Tankard, Communication theories: origins, methods, and uses in the mass media, 4th Edition, *Longman Publishing Group*, New York, 1997.
[60] A. Karlberg, News and Conflict, How adversarial News Frames Limit Public understanding of Environmental Issues, *Alternative Journals*, 1997.
[61] M. Lombard, J. Snyder-Duch, C. C. Bracken, Content analysis in mass communication:
Assessment and reporting of intercoder reliability. *Human Communication Research*, 28, 587–604, 2002.
[61] Bates, B.C., Z.W. Kundzewicz, S. Wu and J.P. Palutikof, Eds., Climate Change and Water. *Technical Paper of the Intergovernmental Panel on Climate Change*, IPCC Secretariat, Geneva, 210 pp, 2008.
[62] G.F. Bishop, R.W. Oldendick, A.J.Tuchfarber,. Effects of presenting one versus two sides of an issue in survey questions. *The Public Opinion Quarterly* 46, 69–85, 1982.
[63] J.A., Krosnick, Survey research. *Annu. Rev. Psychol*. 50, 537–567, 1999
[64] Resource Media 2009. Water and Climate Change in the West: Polling and Media Analysis; March 2009. Carpe Diem - Western Water and Climate Change Project. San Francisco: Resource Media, 2009.
[65] D. P. Bertsekas, Nonlinear Programming, Athena Scientific, Belmont, MA 02178-9998, second edition, 1999.