Environmental Modelling & Software 62 (2014) 128–138

Contents lists available at ScienceDirect

Environmental Modelling & Software

journal homepage: www.elsevier.com/locate/envsoft

A text mining framework for advancing sustainability indicators

Samuel J. Rivera [a, *], Barbara S. Minsker [a], Daniel B. Work [a], Dan Roth [b]

[a] Department of Civil and Environmental Engineering, 205 N. Mathews Ave., University of Illinois at Urbana-Champaign, Urbana IL 61801, USA
[b] Department of Computer Science, 201 N. Goodwin Ave., University of Illinois at Urbana-Champaign, Urbana IL 61801, USA

ARTICLE INFO

Article history:
Received 16 October 2013
Received in revised form
18 July 2014
Accepted 15 August 2014
Available online 16 September 2014

Keywords:
Sustainability indicators
Text mining
Informatics
Knowledge discovery

ABSTRACT

Assessing and tracking *sustainability indicators* (SI) is challenging because studies are often expensive and time consuming, the resulting indicators are difficult to track, and they usually have limited social input and acceptance, a critical element of sustainability. The central premise of this work is to explore the feasibility of identifying, tracking and reporting SI by analyzing unstructured digital news articles with text mining methods. Using San Mateo County, California, as a case study, a non-mutually exclusive supervised classification algorithm with natural language processing techniques is applied to analyze sustainability content in news articles and compare the results with SI reports created by Sustainable San Mateo County (SSMC) using traditional methods. Results showed that the text mining approach could identify all of the indicators highlighted as important in the reports and that the method has potential for identifying region-specific SI, as well as providing insights on the underlying causes of sustainability problems.

Software availability

Name of software: SI News Classifier
Developers: Samuel Rivera (srivera2@illinois.edu)
Contact address: 4129 Newmark Lab, MC-250,205 N. Mathews Ave. Urbana, IL 61801 USA
Availability and Online Documentation: Free download with installation manual and supporting material at GitHub account of the Environmental Informatics and Systems Analysis group at the University of Illinois at Urbana–Champaign (https://github.com/EISALab).
License: The University of Illinois/National Supercomputer Application Center (NCSA) Open Source License (http://opensource.org/licenses/NCSA).
Year first available: 2014
Software required: Matlab and Dataless Classification (developed by the Cognitive Computation Group led by Professor Daniel Roth at the University of Illinois at Urbana–Champaign and available at: http://cogcomp.cs.illinois.edu/page/software_view/Descartes.)
Programming language: Matlab
Program size: 72 KB

1. Introduction

1.1. Motivation

The planet's environmental challenges, economic instability, and finite resources have raised interest in assessing sustainability and measuring progress towards sustainable development (Farr, 2008; Rockström et al., 2009; Solomon et al., 2009; UNDESA, 2010). Over the last 20 years, *sustainability indicators* have emerged as the preferred method to track this progress, and to aid decision making towards sustainable development (Dahl, 2012).

Sustainability indicators are metrics that track the current state and evolution of these complex systems (Hammond et al., 1995; IISD, 2000), such as the number of people living in poverty or the health of endangered species. To be comprehensive, these indicators must address the political, economic, social, and environmental components of communities, and must be understood by all members of society (Innes and Booher, 2000; Dahl, 2012). Indicators are most effective when they are aligned with the values and concerns of the target audience (Dahl, 2012). Lastly, these indicators must be created at multiple scales of governance, including global, national, regional, and city scales in order to be effectively utilized within local cultural, social, political and economic characteristics of each community (Bossel, 1999; Innes and Booher, 2000; Gahin et al., 2003; Dahl, 2012).

To date, 895 initiatives exist worldwide to develop sustainability indicators ranging in scales from cities to global projects (IISD,

* Corresponding author. Tel.: +1 787 240 0044.
E-mail addresses: srivera2@illinois.edu, sammy.rivera14@gmail.com (S.J. Rivera), minsker@illinois.edu (B.S. Minsker), dbwork@illinois.edu (D.B. Work), danr@illinois.edu (D. Roth).

http://dx.doi.org/10.1016/j.envsoft.2014.08.016
1364-8152/

2013). However, not all of the methodologies and guidelines for developing and implementing sustainability indicators have been effective (Gahin et al., 2003; Krank and Wallbaum, 2011). Most indicator projects seek input from representative populations to decide which problems should be addressed (Innes and Booher, 2000; Yli-Viikari, 2009; Dahl, 2012), for example through surveys, public hearings, professional meetings, and other means. Unfortunately, these approaches often fail to achieve large-scale participation, and consequently may not be representative of the community's true values and concerns (Innes and Booher, 2000; Gahin et al., 2003; Adinyira et al., 2007; Yli-Viikari, 2009; Scerri and James, 2010; Krank and Wallbaum, 2011; Dahl, 2012; Moldan et al., 2012). Furthermore, collecting this input is often extremely time consuming and resource intensive, which ultimately leads to large latencies in the reported data (Innes and Booher, 2000; Gahin et al., 2003; Moldan et al., 2012).

These limitations contribute to the challenges associated with using information from sustainability indicators in environmental modeling and management. Approaches such as integrated environmental modeling (Laniak et al., 2013; Jakeman et al., 2008) and decision making, participatory modeling and socio-environmental modeling (Krueger et al., 2012; Voinov et al., 2014) require input on the state of sustainability indicators in order to inform the interactive decision making process. Additionally, it has been argued that the presentation of scientific evidence without an understanding of the preferences and values affecting the decision and actions of stakeholders will usually fall short of being effective (Laniak et al., 2013). It is necessary to have an understanding of the perceptions and values of stakeholders on sustainability issues to generate adequate (re)actions in the form of policies and management strategies (Krueger et al., 2012; Voinov et al., 2014).

The objective of this work is to begin addressing these limitations through the development of a new method to design and track sustainability measures using digital news media and recent progress in text mining. More specifically, this paper addresses the question of whether the news media contains relevant information that can enable fast identification, tracking, and reporting of sustainability indicators for a region. The hypothesis to be tested is that the unstructured data of news media can provide insight into sustainability problems within the cultural and contextual characteristics of a community, thereby also addressing the social component that has been underdeveloped in previous approaches.

1.2. Related work

Recent statistics show that half of the news across the world have been digitized, and are supplanting print and broadcast news (Leetaru, 2011). News media represents a “mediated public sphere” (Holt and Barkemeyer, 2012) with the potential to influence people's mindsets and create a feeling of worldwide connectedness by changing the public's level of awareness and attention to a specific issue (Szerszyski et al., 2000; Holt and Barkemeyer, 2012). Also, it has been argued that newspapers are the most important source of information used to spread scientific knowledge (Nelkin, 1995; Morse, 2008), and are very effective in placing topics in the public mind (Holt and Barkemeyer, 2012). Moreover, it has been shown that there is a causal relationship between thematic priorities of the media and the relevance of social problems in the population (Rogers et al., 1993). Furthermore, news articles contain far more than just factual details; they provide insights into the cultural context upon which they are written, a spatial and temporal component to the facts, and a window for forecasting many social behaviors using text mining techniques (Thøgersen, 2006; Tang et al., 2009; Leetaru, 2011; Michel et al., 2011).

To identify actionable information from the large volumes of unstructured digital news, efficient text mining tools are needed to process the data and extract trends. Recent studies suggest that analysis of text archives can generate new knowledge about the functioning of society (Leetaru, 2011). Moreover, text mining tools can detect the tone of news articles, which enables applications such as forecasting social behaviors ranging from ticket movie sales to stock market trends (Mishne and Glance, 2006; Tang et al., 2009; Leetaru, 2011; Michel et al., 2011). In one recent study, an analysis performed on the global news tone of a 30-year worldwide news archive demonstrated that this type of analysis could have forecasted the revolutions in Tunisia, Egypt, and Libya, including the removal of Egyptian President Mubarak (Leetaru, 2011). This suggests the use of text mining techniques to quantify and assess the social components of sustainability could hold promise.

Text mining techniques have previously been applied to measure some elements of sustainability. “Trends in Sustainability” is a Web application that searches for predetermined keywords related to different sustainability topics in 115 newspaper sources from 41 different countries (Barkemeyer et al., 2009). The result of the application is a display of the trend of the volume of news containing the keyword across time. The “Carbon Capture Report”, is a similar application that searches the social media (e.g. news, blogs, Twitter, and Youtube) for predetermined keywords to identify relevant articles. However, in the “Carbon Capture Report” the data are further processed by implementing *natural language processing* (NLP) techniques and a sentiment analysis that provide further information. The application displays a time series analysis of the volume of data with an overall tone (positive, neutral, negative) and activity of the topics and color coded world regions that indicate the magnitude of their contribution to the data, as related to that topic. A similar application is the Media Watch on Climate Change, a public Web portal that aggregates large archives of digital news and social media coverage on climate change and related issues (Scharl et al., 2013). Using an interactive dashboard the location, the frequency, and the sentiment of the information is displayed to stakeholders with the idea of increasing awareness and availability of environmental information.

1.3. Contributions

Previous sustainability studies using text mining have focused on tracking general trends in different sustainability and climate change topics at the national or global scale. This study focuses on demonstrating a faster method for identifying, tracking, and reporting of sustainability indicators specific to a region using news articles that can better incorporate society's values. Furthermore, the approaches taken in this study provide links between observed indicator trends and their underlying causes; previous indicator methods focus primarily on tracking the issues.

Additionally, this study develops a new methodology that combines a suite of text mining methods to more accurately classify sustainability articles given the limited training set of regional news articles and their non-mutually exclusive topic areas (e.g., an article on the effects of pesticides on water quality would be equally relevant to the water quality and pesticide indicators). The methodology is applied in San Mateo County, CA, to demonstrate feasibility of the approach. Future work is then recommended to extend the methodology to other types of data available on the Web (e.g., social media data and blogs).

2. Methodology

The tracking and extracting of information from sustainability related news articles is accomplished by integrating different classification approaches and NLP

techniques. The methodology has three main components (Fig. 1): pre-processing of the unstructured textual news data, classification of the documents under pre-determined labels, and NLP information retrieval. In the pre-processing step the textual data from a set of documents is represented by a word-document matrix. The rows of the matrix represent the documents and the columns represent all words in the documents, also known as n-dimensional features. The matrix is constructed by assigning an importance metric to each word in the article (e.g. frequency of occurrence of the word in the document). These metrics are then stored in the document's row of the matrix. This representation allows the calculation of similarity between the new document and a pre-labeled set, which is used for classification of the document under the most related sustainability indicators. Finally, NLP techniques are used to assign a geographical location and identify co-occurring topics in the set of documents. This information can provide additional insights that help explain the current state of the sustainability indicators. Fig. 1 presents a conceptual overview of the steps involved in the approach, which are defined in more detail in the sub-sections below.

2.1. Pre-processing: from textual data to word-document matrix

Typical pre-processing methods used for the transformation of unstructured textual data to a word-document matrix (Fig. 1 block 1) include the tasks of tokenization, elimination of stop words, and the calculation of an importance metric. Tokenization is the process of splitting the text into individual words or tokens. Tokenization within the English language is often done by using blank spaces and punctuation marks as token delimiters (Miner et al., 2012). During this process, "stop words" are also eliminated. Stop words are 'functional' words that don't carry any meaning, such as "the", "for", "that", "to", etc., and do not provide insightful representation of the document's context. Eliminating these words decreases computational cost and often increases the accuracy of classification tasks (Feldman and Sanger, 2007; Weiss et al., 2010).

The final pre-processing step involves converting the occurrence of words within a document to a metric that represents its relative importance. In this study the binary representation of the occurrence of a word and the *term frequency-inverse document frequency* (TF-IDF) were used as metrics of importance. The binary representation is computed by assigning a 1 if a word is present in the document and 0 otherwise. This metric is used to filter out words that only appear in one document in the set, a necessary step for the correct implementation of the generalized discriminant analysis discussed in Section 2.2.1. The TF-IDF computes the frequency of a word in a document and all the other documents in the set as a metric of the importance of a word in a document relative to all of the other words and documents (see details in Appendix A). The TF-IDF is one of the most commonly used importance metrics for unstructured text and has been demonstrated to be among the best approaches for classification tasks (Weiss et al., 2010).

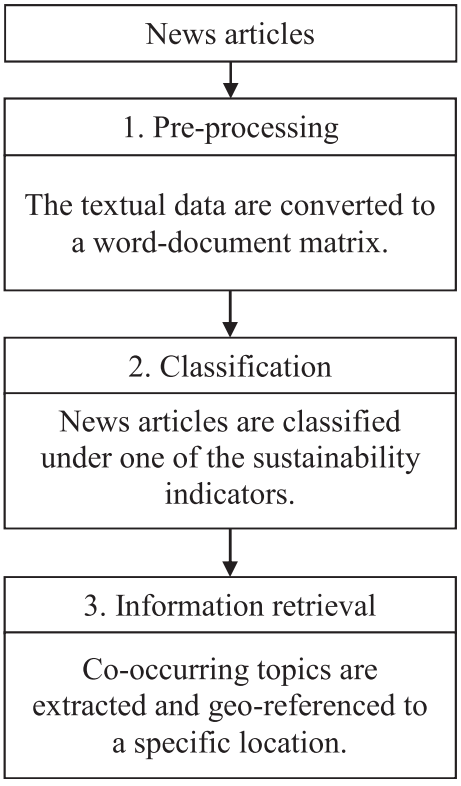

Fig. 1. Methodology overview.

These pre-processing steps are executed using Rapidminer, an open source system for data and text mining. Rapidminer has been used in other text mining studies successfully (Miner et al., 2012), and it has the benefit of providing a user-friendly interface. The documents are pre-processed in batches of 5000 documents due to the significant computational time required when more documents are added simultaneously. For this application, processing the data in batches had no effect on the ranking of the words within the documents or the magnitude of the TF-IDF.

2.2. Assignment of sustainability indicator labels to news articles

Text classification is the task of assigning unlabeled documents to a set of pre-determined classes based on their similarity to a pre-labeled training set (Fig. 1 block 2). It is often performed using supervised classification algorithms (Feldman and Sanger, 2007; Weiss et al., 2010). These type of algorithms perform best with a representative and uniform distribution of examples for each class in the training set (Feldman and Sanger, 2007; Weiss et al., 2010) and their performance is usually tested on mutually exclusive datasets (e.g. music v.s. sports). Furthermore, the training sets are usually created by manual human labor to classify randomly selected documents. Application of such methods to sustainability news articles posed new challenges due to the non-mutually exclusive characteristics of the news articles, which often span multiple sustainability indicators, and the labor associated with the task of pre-labeling a training set. Given the large number of news articles not related to sustainability indicators, manual labeling of the training set was a time consuming and resource intensive task that produced an unusually small training set for each sustainability indicator. Thus, the performance of typical classification approaches was initially poor because of these issues.

To improve performance, the classification of news articles into different sustainability indicators was done using two supervised classification algorithms and sustainability indicators categorization taxonomy, often refer to in the literature as a hierarchical category tree. A hierarchical category tree is a labeling process where news articles are pre-labeled as part of internal (parent) nodes and leaf (child) nodes. Within the tree, child nodes can only be associated with their immediate parent nodes (see the nodes used for sustainability topics in Table 1). Hierarchical category tree structures have been used to eliminate the possibility of misclassification due to similarity between words in documents that are not in the same category (e.g. a news article talking about wind power versus one taking about a storm) (Sun and Lim, 2001, Silla and Freitas, 2011). In this study, the hierarchical category tree aids the classification process by reducing the non-mutuality of sustainability-related news articles. Parent nodes of the tree were selected to be mutually exclusive, thereby minimizing use of the same words in different classes. Once a parent node is selected, the number of possible classes is reduced to the child nodes of that parent, concentrating the classification task around that set of indicators. The hierarchical tree used in this study has a root node with only two labels: sustainability related (SR) or non-sustainability related (NSR). Additional levels in the tree are structured as shown in the columns in Table 1.

Once the hierarchical tree is established, the task to classify each document within the hierarchy is composed of three main steps (Fig. 2): (1) transformation of the data using a generalized discriminant analysis (GDA) to reduce dimensionality of the classification problem, (2) classification of documents as related or non-related to the sustainability indicators using K-Nearest Neighbors (KNN) and (3) classification of the documents under the parent and child nodes of the category tree using KNN, Dataless classification, and a majority voting rule. Given the large set of sustainability indicators, the approach uses KNN to suggest candidate class labels to the Dataless classification algorithm, which takes only two labels as input and uses Wikipedia as its training set to classify the news articles. The majority voting rule then combines the KNN and Dataless results to determine a final classification of each article. This novel combination of these techniques aims at reducing the time required for the task of pre-labeling news articles and allowing the use of a small training set in the classification model. These steps are described in more detail in the sub-sections below.

2.2.1. Generalized discriminant analysis (GDA)

Most document classification algorithms start with dimensionality reduction to reduce computational cost and increase classification accuracy (Feldman and Sanger, 2007). Generalized linear discriminant analysis reduces dimensionality by simultaneously minimizing within-class distance and maximizing between-class distance using a linear transformation of the n-dimensional feature space (in this application, the words in the news documents) of the training set. The new dimensions of the transformed feature space are reduced to at most the total number of different classes in the training set. This study adopts the generalized discriminant analysis developed by Li et al. (2008), which uses a generalized eigenvalue decomposition to create a transformation of the feature space that reflects the inherent similarity of the data and allows for efficient implementation of the document classification. For this application, GDA helps reduce the non-mutually-exclusive characteristic of the words and improves the classification accuracy by creating more distinct classes at the parent node level of the hierarchical sustainability indicator tree. For more details on the implemented GDA please refer to Li et al. (2008).

Table 1
Structure of hierarchical category tree for sustainability indicators. (* indicators not included in the SSMC reports).

Root node	Parent nodes	Child nodes	
Sustainability related	Education	Academic performance index	
		High school graduation	
		School drop out	
		Vocational education	
	Poverty	Affordable housing	
		Food stamps*	
		Homeless	
	Agriculture	Agricultural production value	
		Crops	
		Farm	
		Farmers market*	
		Organic agriculture	
		Community garden*	
	Children health	Child care	
		Children obesity	
	Community health	Chronic illness	Diabetes
			Cancer
			Heart disease
			Asthma
			Arthritis
			Alzheimer's disease*
			Hepatitis*
			Osteoporosis*
		Average life expectancy	
	Health care		
	Pollution	Air quality	Carbon emissions
			smog
		Beach closure	
		Water pollution	
	Crime	Gang violence	
		Child abuse	
	Drought		
	Flooding		
	Disaster preparedness		
	Energy	Energy consumption	
		Renewable energy	Solar power
			Wind power
	Solid waste		
	Unemployment		
	Public transit		
	Habitat protection	Invasive species	
		Open spaces	
		Endangered species	
	Green buildings		
	Pesticide		
	Deforestation*		
	Overfishing*		
	Fecal coliform*		
	Coastal erosion*		
Not sustainability related			

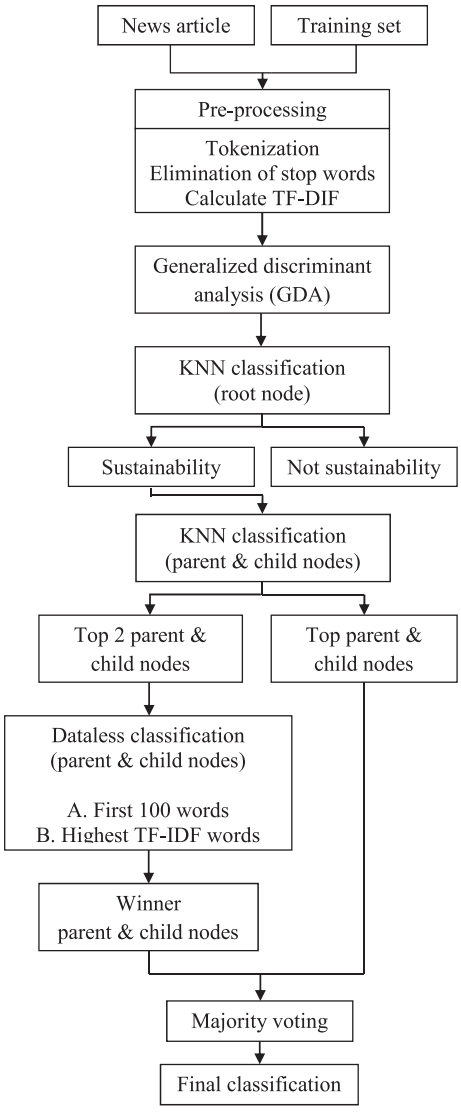

Fig. 2. Flow diagram of classification approach.

2.2.2. K-nearest neighbors

The second step of the classification process identifies sustainability-related news articles and the most probable sustainability indicators for each article using KNN. The KNN algorithm is a supervised machine learning technique, typically used for classification tasks, that compares vectors and measures their similarity. KNN is widely used for text mining and information retrieval because of simplicity, its ease of implementation, and scalability (Feldman and Sanger, 2007). Depending on the application, KNN has many different implementations and can use different measures of similarity to compare documents. The implementation adopted in this study assigned news articles to the class with the highest number of occurrences within the set of the k most similar documents. The similarity of the documents is measured using Euclidean distance, a classic approach that calculates the length of the line segment connecting two vectors (in this case, each row of the word-document matrix). The method was selected because it followed the GDA implementation described by Li et al. (2008), and was shown to give the best predictive values when TF-IDF is used as the importance metric.

KNN is applied twice in the classification algorithm shown in Fig. 2. The first application is for the binary classification at the root node, determining whether each news article is related to sustainability or not. The second application is a multi-class classification where news articles are classified in the lower levels of the hierarchical tree. The results of the second application are used to suggest the two most probable parent nodes, and their respective child nodes, of the sustainability indicators relevant to the article, which are then input to the Dataless classification algorithm in the next step. KNN was selected for these classification tasks for its robustness when classes are not linearly separable (Hastie et al., 2009). This characteristic was observed at the boundary of many of the classes because of the similarity of words used in the news articles. Furthermore, the simplicity of implementing KNN allowed for rapid classification of a large number of articles and outperformed other supervised classification approaches (Fig. 3 and Fig. 4).

2.2.3. Dataless classification

The third step in the classification process uses Dataless classification to classify articles at the parent and child node levels using the previously suggested labels from KNN. Unlike KNN, Dataless classification does not need annotated training data, as it interprets a string of words as a set of semantic concepts (i.e., concepts with an intuitive meaning, idea, or thought associated with them), using Wikipedia as a source of semantic knowledge (Chang et al., 2008). Developed by members of the Cognitive Computation Group led by Professor Daniel Roth at the University of Illinois at Urbana–Champaign, the classification algorithm takes as input two labels, a single word or phrase (in this case, the two most probable sustainability indicators

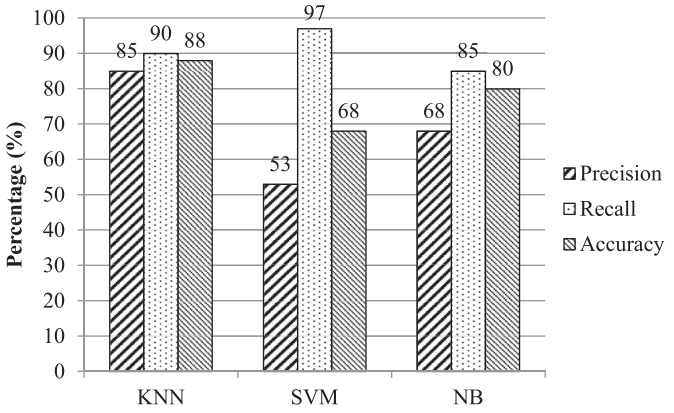

Fig. 3. Performance metrics for sustainability-related news articles (KNN– K Nearest Neighbors, SVM – Support vector machine, NB – Naive Bayes).

suggested by KNN), and a string of text (from the original news article), and outputs a number of associated *concepts* related to each label.

Dataless classification assumes each article in Wikipedia corresponds to a *concept* and an article can be associated with many concepts. The algorithm identifies which of these concepts are relevant to the document being classified by associating the words in the document with the text of the Wikipedia entries by using Explicit Semantic Analysis (ESA). The concepts are then combined in a weighted vector and a final list of concepts associated with the text is ordered by their weight. Finally, each concept in the list is assigned to one of the two input labels based on the likelihood of appearance of the label in that concept (e.g., the concept *pesticide* is more likely to be associated with the label *agriculture* than with the label *child abuse)*. The label assigned to each news article is the one with the highest number of associated concepts.

The success of the Dataless algorithm lies in using Wikipedia as the "world's knowledge" to infer the semantic meaning of labels to be analyzed. Given the extensive list of varied and interconnected topic related to sustainability and the number of news articles not related to sustainability, creating a training set with sufficient examples for each sustainability indicator would require an extremely time-intensive human task. Using two sustainability indicators as input, Dataless classification aids the classification process by allowing the use of a smaller number of pre-labeled examples for each indicator in the training data, while maintaining good classification accuracy. In this study the Dataless classification algorithm was applied in two different ways concurrently, each time using a different portion of the article. The first implementation of the method used only the first 100 words of the news articles, given that the main topic is often stated in the first paragraph. The second implementation involved using the 50 words with the highest TF-IDF values; this approach is used for articles that introduce the topic in the body of the text. Each implementation provides a candidate label that could be assigned to the news articles.

Once the Dataless algorithm completes its classification of each article, at the parent node and child node levels, the final classification is made by reconciling the results of the KNN and Dataless classifications using a majority voting rule. Each of the two different applications of the Dataless classification and the results of the KNN receives a vote. All votes receive equal weight and the child node label with the majority of votes is chosen as the final label of each article. If the suggested child nodes are all different, then the parent nodes of the child nodes are used to label the news article. If all documents in the *k* set from KNN correspond to the same class, then that class is automatically assigned to the news article and the use of the Dataless classification is omitted. This unique combination of the two algorithms was observed to significantly outperform individual implementation of the KNN and Dataless classification methods in terms of computational cost and classification accuracy.

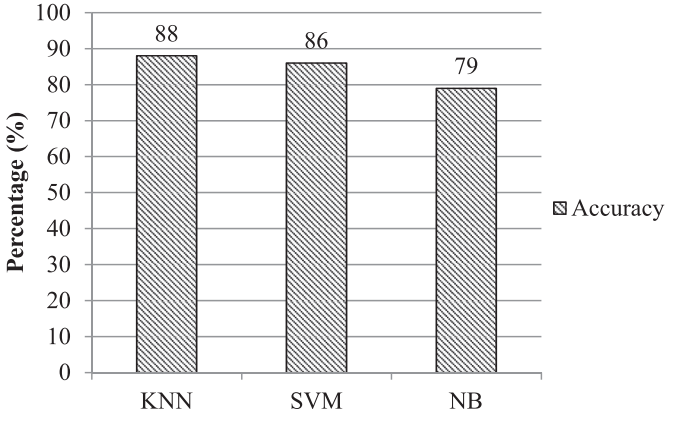

Fig. 4. Classification accuracy for parent nodes of sustainability-related news articles (KNN– K Nearest Neighbors, SVM – Support vector machine, NB– Naive Bayes).

The superior performance achieved by the combination of the two algorithms is likely due to the synergetic nature of the approach. KNN has a low computational cost but requires a large number of training examples to achieve good classification accuracy. Dataless classification uses Wikipedia to construct a large and diverse training set that allows it to achieve good classification accuracy, however the method suffers from high computational cost due to the limitation that the method takes only two labels at a time. By combining the two methods, KNN is used to provide an *intelligent* guess of the potential sustainability labels to the Dataless classification, thus reducing the high computation cost without sacrificing classification accuracy.

2.3. Information retrieval tasks

The final step of the method is to automatically extract and analyze content from the already classified documents that help explain the current state of the sustainability indicators. Two main techniques are used to complete this task: (1) the creation of frequent concept sets and association rules that identify and summarize recurrent topics across all news articles classified under the same sustainability indicator, and (2) Part-of-Speech (POS) tagging and a gazetteer for geo-referencing of news articles, identifying areas with the highest interest in a particular sustainability indicator. These two steps are defined in more detail in the sections below.

2.3.1. Frequent concept sets and association rules

Frequent concept sets represent the co-occurrence of different concepts represented in a document collection, given a co-occurrence frequency threshold known as the minimal support level. They have been widely used in data mining applications ranging from market research to text mining to discover unknown patterns in the data (Feldman and Sanger, 2007). The Frequent-Pattern Growth (FP-growth) algorithm is implemented in this study, which is a widely used algorithm that discovers frequent concepts without generating a candidate concept set, reducing computational cost (Feldman and Sanger, 2007).

Next, association rules are derived that identify the direct relationship between concepts or a set of concepts taking the form of $A \Rightarrow B$, indicating that every time A is present B is also present (Feldman and Sanger, 2007). The creation of association rules is subject to a minimum level of confidence defined as the percentage of time the rule is met (e.g. the percentage of documents that include all the concepts in *B* within the subset of those documents that include all of the concepts in *A*) (Feldman and Sanger, 2007).

The sequential use of both frequent concept sets and association rules allows extraction of keywords and their relationships to provide information about the underlying causes of events related to the sustainability indicators.

2.3.2. Part-of-speech tagging and geo-referencing of news articles

Part-of-speech (POS) tagging is the process of labeling different words in the text of a document as a part of a grammatical class or part of speech given the context and the definition of the word (Feldman and Sanger, 2007). In natural language processing, POS is often used to recognize names of people, places, and organizations. In this study, the Meandre OpenNLP POS tagger (NCSA, 2013; OpenNLP, 2013), developed by the National Center of Supercomputing Applications at the University of Illinois at Urbana–Champaign, is utilized to identify places mentioned in the news articles. By applying the POS to the news articles, proper singular nouns can be identified and searched within a gazetteer, a database of geographic locations with the latitude and longitude of each record. News articles can then be associated with the location that is most frequently mentioned in the text and is within a specified region of interest. This type of analysis allows the geo-referencing of news articles related to each indicator to be correlated with a specific location.

3. Case study application and discussion

To illustrate the feasibility of using news articles to identify and track sustainability indicators, a case study was examined in the region of San Mateo County, California. This location was selected because of its long-standing sustainability indicator program and the availability of a digitized local newspaper archive. Sustainable San Mateo County (SSMC), a non-profit public benefit corporation, initiated its sustainability initiatives in San Mateo County in 1992. The SSMC has generated sustainability indicator reports annually since 1998. In these reports, the current states of 49 sustainability

indicators are presented to the community. The current states of these indicators are defined by a series of individually tracked sub-indicators that have specific data associated with each of them. For example, the sustainability indicator for *air quality* is defined by sub-indicators such as suspended particulate matter, ozone level, and total emissions of greenhouse gases.

In this case study, 22 sustainability indicators and 36 sub-indicators were selected for the feasibility analysis (Table 1). The majority of the selected indicators were extracted from the SSMC reports based on the expected likelihood of their appearance in the news articles. The rest were selected from other sustainability reports of international agencies, such as the United Nations (UN) and the European Union, and regional reports such as those presented by Sustainable Seattle and Central Texas Sustainability Indicator Project. These additional indicators represented sustainability problems often seen in similar coastal regions, and were included in the analysis as a validation to the proposed methodology. Lastly, the analysis was conducted using news articles from the San Mateo County Times newspaper. The San Mateo County Times is among the newspapers with the highest circulation in the region, and the only one with an accessible digitized archive. The total number of analyzed news articles was 28,619 and 39,166 for years 2007 and 2009, respectively (2008 data was unavailable). The training set used by the classification algorithm included a total of 1830 sustainability-related news articles, an average of 35 news articles per indicator. The results of the case study include a performance evaluation of the classification approach and a set of illustrative results that demonstrate the feasibility of using news articles to identify and track sustainability indicators. To validate these illustrative results the SSMC reports were used as a representation of the experts' opinions. SSMC reports were developed by a group of experts and thus the replication of the information presented in the report provides sufficient validation to indicate potential usefulness of the tool.

3.1. Performance of classification algorithm

The typical metrics used to assess the performance of text classification algorithms are accuracy, precision, and recall. Accuracy refers to the percentage of correctly classified news articles out of the total number of classified articles (i.e. correct and incorrect classified articles). Within a label, precision refers to the proportion of correctly classified articles to the total number of articles classified under that label. Finally, recall is the percentage of all news articles corresponding to a label that were classified with that label. As an illustrative example, consider a hypothetical set of 10 articles for which the correct labeling is 5 articles with label A and 5 articles with label B. If the classification algorithm correctly classified 7 of the 10 articles, then the accuracy is 70% (7/10). Furthermore, if 6 of the 10 articles were classified as label A, but only 4 of the 6 were correctly classified as label A, then the precision of label A is 67% (4/6). Finally, if only 4 articles were correctly classified as label A, then the recall of label A is 80% (4/5).

These metrics were used to assess the performance of three different supervised classification methods at the root (binary classification scheme) and parent nodes (multi-class classification scheme). The performance at these levels is critical because the algorithm is classifying which news articles are sustainability-related and what two parents should be suggested to the Dataless classification. The GDA-transformed test data described in the next paragraph was used to conduct the assessment. The results in Figs. 3 and 4 show that KNN outperformed the other classification methods; therefore KNN was selected as the preferred classification algorithm for these levels of the hierarchical tree.

Before conducting the performance assessment of the classification algorithm, the optimal number of nearest neighbors k for KNN was selected by performing a leave-one-out cross validation using only the training set. The optimal number was $k = 10$ at the root and parent node levels with 96% and 97% accuracy, respectively. The performance of the classification approach was assessed using a test set of 500 news articles labeled at the child node level by three independent annotators with inter-annotator agreement of 86%. At the root node (classifying sustainability vs non-sustainability articles), the classification approach achieved 88% accuracy with 90% recall and 85% precision for sustainability-related articles. As expected, even with a poorly defined boundary, the KNN achieved good classification accuracy at the root node given the binary nature of the problem. However, at lower levels of the multi-label tree, the overlap of words becomes higher, producing less distinct classes and a harder classification problem. Thus, at the parent and child node level (see Table 1 for the terms at each level) the method achieved 88% and 71% accuracy, respectively. Nevertheless, the performance is satisfactory given that the percentage agreement between humans performing the same task was only 86%. Moreover, as will be seen in the analysis in the next section, the classification accuracy at lower levels of the hierarchical tree was still sufficient to allow extraction of meaningful information from news articles.

3.2. Analysis of sustainability news-related articles

Having tested the accuracy of the classification algorithm, the feasibility of identifying and tracking the indicators using news articles was explored. Each of the 2007 and 2009 articles from the San Mateo County Times were classified under one of the sustainability indicators; the final number of news articles related to sustainability were 5842 (20%) and 9781 (25%) in 2007 and 2009, respectively. To analyze the results, the classified news articles were grouped into their respective parent nodes (Fig. 5) and the volume of news articles was used as a measure of the community's perception of importance of each indicator. This measure was based on the use of volume of media coverage as a traditional measure of the relative importance of an issue over time in previous studies (Mueller, 1973; Benton and Frazier, 1976; Naisbitt, 1976; Beniger, 1978; Dearing and Rogers, 1996; Carvalho, 2005; Holt and Barkemeyer, 2012).

Fig. 5 shows that most of the indicators perceived as important by the community were also included in the SSMC's reports, while the indicators that were extracted from other reports received a low level of attention. However, some of the indicators that were included in the SSCM's reports resulted in a low volume of news articles as well (e.g., pesticides). This can be explained by the nature of the data; although newspaper articles are assumed to be a good objective representation of society's concerns, not all topics are newsworthy all the time.

To further demonstrate the capabilities of the method, the results obtained for the *chronic illness* child node (under the *community health* parent node in Fig. 5) were analyzed. The *chronic illness* child node is composed of 8 illnesses, 5 which were extracted from the SSCM's reports. As seen in Fig. 6, it is clear that the method was able to identify the illnesses most relevant to the community of San Mateo County. Moreover, the two leading causes of death in the county, *cancer* and *heart diseases*, obtained the highest volume of news articles for both years. *Diabetes* and *asthma* followed in magnitude of article volume and were reported among the most common chronic illnesses in the county (SSCM Sustainability indicators report, 2010). These results provide additional insights into the community's perception of chronic illnesses, suggesting that the community has greater concerns for death-related

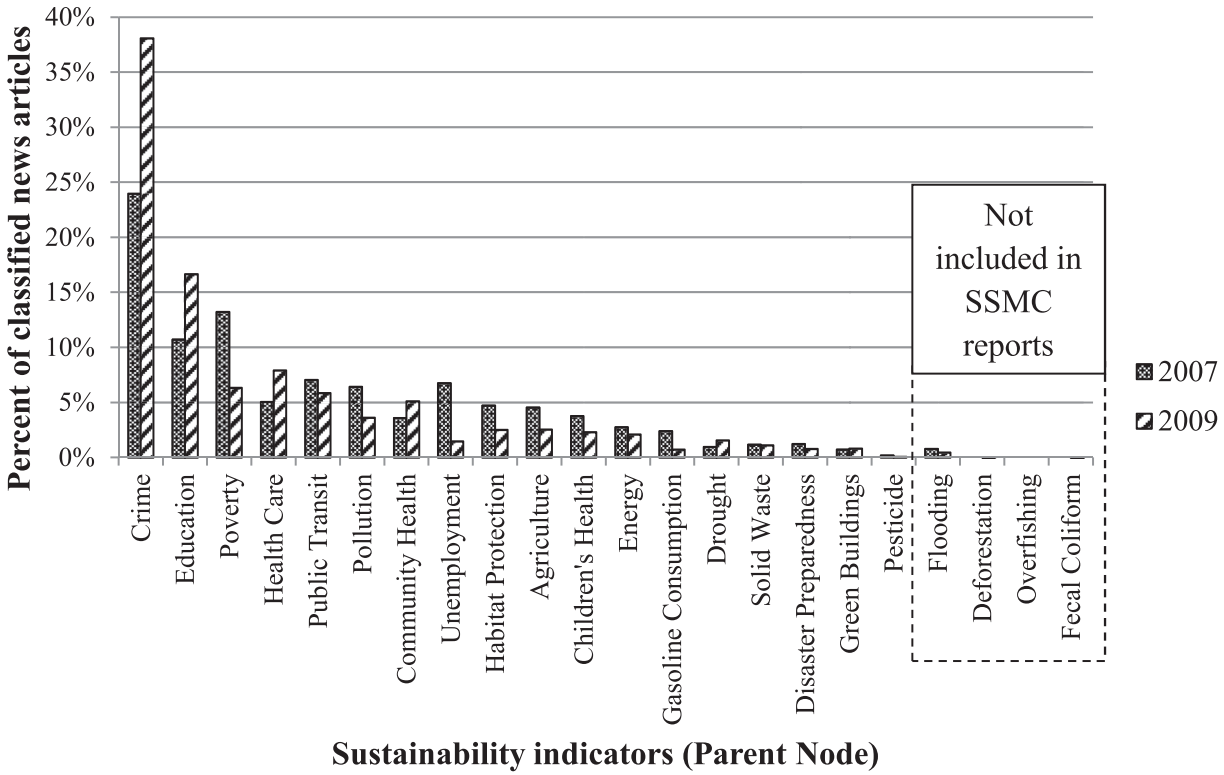

Fig. 5. Distribution of sustainability topics in the sustainability related articles.

illnesses. This finding demonstrates the capabilities of the method for identifying sustainability issues specific to a region by providing insights on the relative importance given to different sustainability indicators by the community, an issue that had been overlooked in the past and which is necessary for the creation of value-based indicators.

The indicators were also tracked by calculating the monthly volume of news articles using their date of publication. Additionally, information retrieval techniques were used to analyze months with unusually high numbers of articles to identify events that led to such high volumes. As a feasibility demonstration, the high-volume time series for *health care, flooding* and *pollution* parent node indicators for the years of 2007 and 2009 were analyzed (Fig. 7 and Fig. 8, respectively).

Fig. 7 shows that the parent node *pollution* indicator peaked in November of 2007. The analysis of the volume of news articles at the child node level revealed that the *water pollution* and *beach closure* indicators were the main contributors to the volume of this peak, and suggested a relationship between the two indicators. Application of the FP-growth algorithm and the creation of association rules using the news corresponding to these peaks allowed the extraction of keywords to gain insights into the general topic of these articles. The keywords with the highest frequency for both indicators were: *Cosco, Busan, San Francisco, oil, water, spill, fuel,* and *clean up,* suggesting that an oil spill had occurred in the San Francisco area during this time. Further validation confirmed that indeed on November 7, 2007 the freighter Cosco Busan caused an oil spill of 58,000 gallons in the San Francisco Bay area (SSMC Sustainability Indicators report, 2008; NFWF, 2013). In addition, the 2008 SSCM report indicated that in 2007, six beaches were closed for four days due to the Cosco Busan oil spill, for a total of 24 total closed beach days (SSMC Indicator report, 2008).

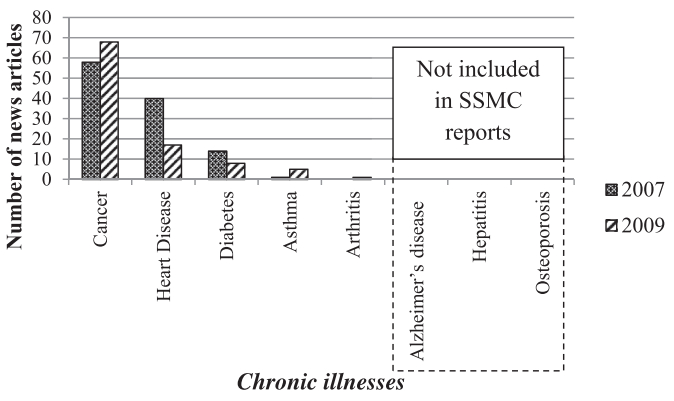

Fig. 6. Distribution of *Chronic illnesses* in the sustainability related articles.

In the analysis of the 2009 *pollution* time series (Fig. 8), the volume of articles was divided into the child nodes for pollution. The *water pollution* indicator peaked in the month of February, for which the keywords *water, control, sewage, storm* and *city* were extracted. These keywords suggested the pollution of the water due to an event related to sewage and a storm, thus these other related indicators were analyzed. The results show that the *flooding* indicator also had distinct peaks in February and October. The application of information retrieval techniques to these two peaks suggested the occurrence of two storms given that the keywords with highest frequency included: *mph, wind, storm, rain* and *strong.* The correlation of the results in their peaks and the corresponding keywords suggest that storms in the area caused sewage overflows that contaminated the water in the Bay Area. This hypothesis was validated by the 2010 SSCM Sustainability Indicators report where it stated that storm flooding often caused sewage overflows, decreasing the water quality. These results, and Fig. 5, suggest that

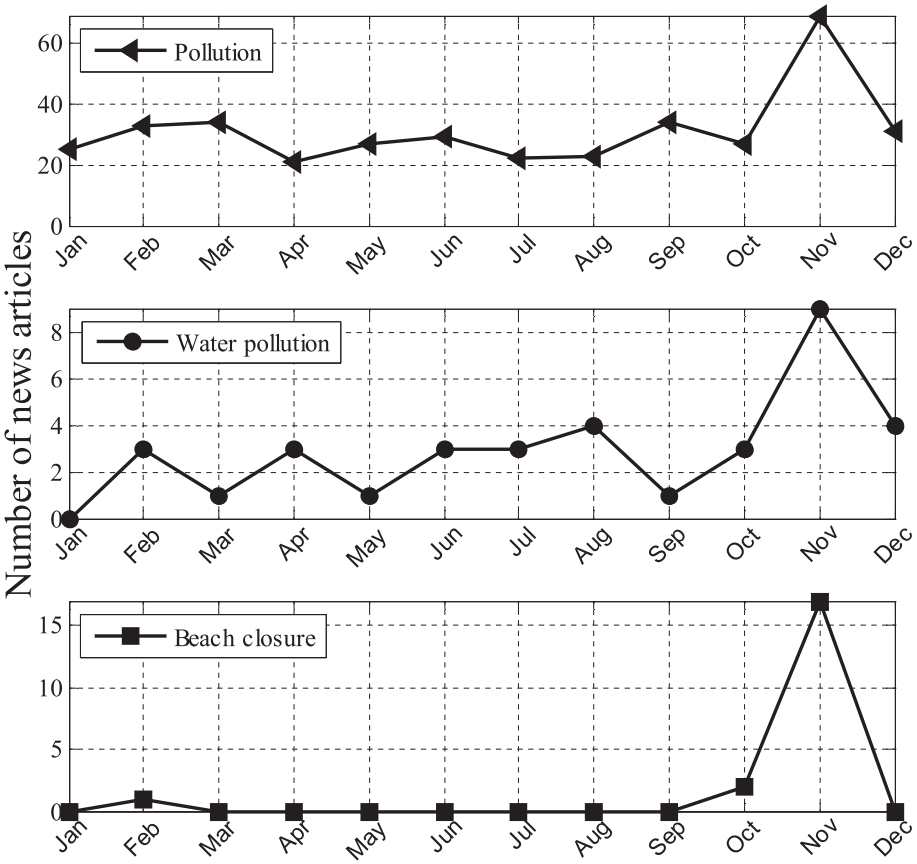

Fig. 7. Volume of news articles for selected indicators in the 2007 time series.

flooding is a concern in the region, although it was not included as one of the indicators in the SSMC reports.

Another distinct peak in the 2009 pollution time series occurred in January and December, due to the volume of news obtained by the child node of air quality. For these two peaks the keywords obtained included: *winter, wood, burning, fireplace, air* and *pollution*. These keywords suggest a decrease in the air quality due to particulate matter due to wood burning in the winter months, just as reported in the 2010 SSCM Sustainability Indicators report. Furthermore, the time series of volume of news articles for air quality were found to have good correlation with 2.5 particulate matter measurements during the same time period ($\rho \approx 0.7$).

Lastly, the 2009 *health care* time series demonstrated a distinct peak in the month of September. Further analysis of the peak suggested the keywords: *Obama, health care, plan,* and *Congress*. Further validation corroborated that in September 2009 there was a discussion in Congress about the changes to be made to the health care system proposed by President Obama.

Finally, the ability to locate the sustainability problems within the region of San Mateo County was evaluated by using the POS techniques and the GeoNames database (www.geonames.org/) to georeference news articles to a specific area within the county. The method was able to identify correctly the specific location of the Cosco Busan oil spill by suggesting *Oakland* and *San Francisco* as the two places most frequently mentioned within news articles related to the observed peak from Fig. 7. Furthermore, the method was able to correctly identify 4 of the 6 beaches closed in 2007 due to the oil spill.

These events are fairly concentrated in a specific area. To analyze the capability of the method in identifying less concentrated events, the indicator of *crime* was analyzed, which had a large volume of articles. The results, shown in Fig. 9, demonstrated that the method was able to correctly identify high crime areas using only the volume of news articles as a metric of crime frequency. The results were validated using the aggregated police reports for the region (www.citydata.com), also shown in Fig. 9. Nevertheless, the problems related to some other indicators with a high frequency of news articles, such as *education* and *poverty,* could not be georeferenced to a specific place within the county. These indicators are often discussed at the county levels, and specific locations are not included in the articles.

In summary, these results indicate significant promise for text mining to address the timely and resource-intensive nature of identifying, tracking, and reporting the indicators by providing regular updates (e.g., daily, weekly, or monthly) on: (1) the relative importance of different sustainability indicators to a region, (2) the current state of these indicators and (3) underlying events affecting them. The analysis of data to provide an explanation to the observed changes is likely one of the most human intensive and time consuming tasks associated with indicator reporting. The method was able to provide the same explanations for the behavior of the sustainability indicators in near real time by suggesting associated keywords and correlation between many of the indicators. This additional information allowed the identification of the events causing shifts in discussions of these sustainability issues. All of these results correlated well with the information provided in the SSMC sustainability reports, thus validating the capabilities of the method for tracking and reporting the state of the sustainability indicators.

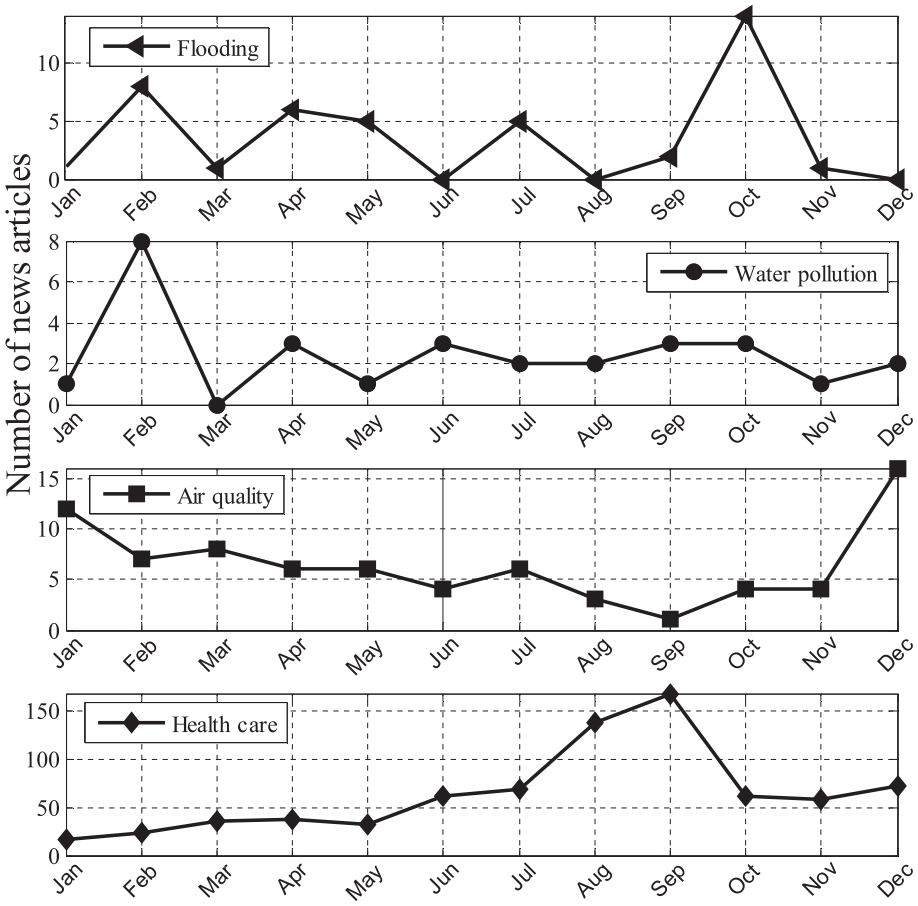

Fig. 8. Volume of news articles for selected indicators in the 2009 time series.

However, the results of the study were limited by sparse data that prevented the identification of some of the more chronic problems in the region (e.g. loss of endangered species, crop loss due to invasive species, and carbon emission due to vehicles). These types of sustainability issues are less newsworthy, limiting the approach to acute events that are more likely to be frequently reported. Other online news sources (e.g., blogs associated with citizens groups that are concerned about a particular chronic topic) may be more fruitful for tracking chronic problems. Nonetheless, the low frequency of news articles related to chronic sustainability issues suggests that society may place a lower value on chronic problems.

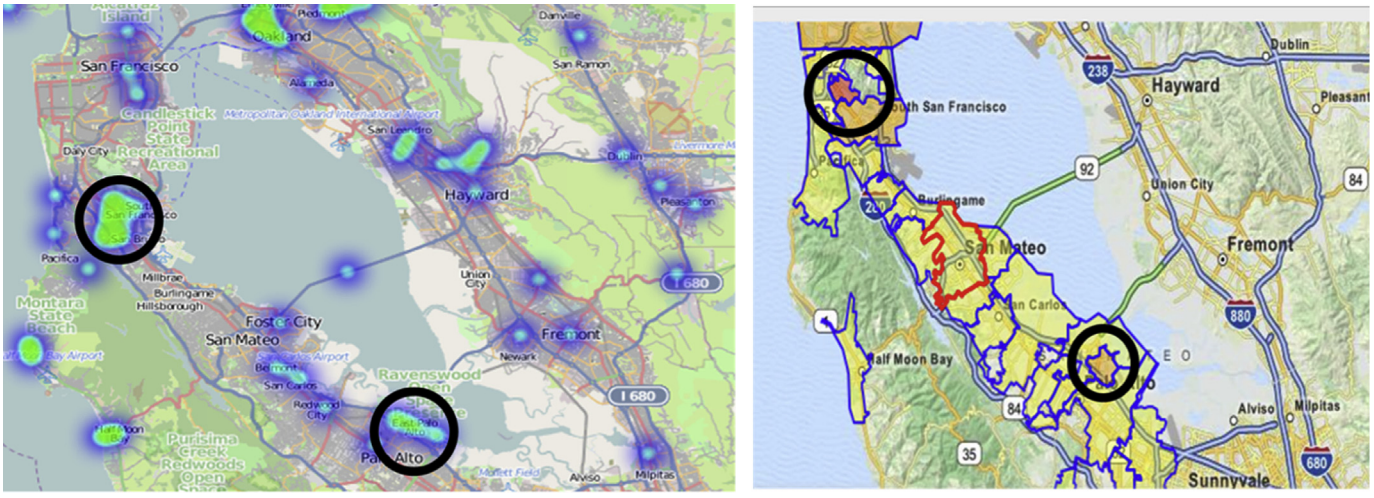

Fig. 9. Mapping of high crime areas (left: heat map derived from *crime*-related news articles, right: map derived from aggregated San Mateo County police reports [Map: City-Data.com]).

4. Conclusions

This study is a first attempt at utilizing news media to more easily create effective sustainability indicators. Using a non-mutually-exclusive document classification algorithm and the incorporation of different information retrieval techniques, the analysis demonstrated that mining the growing amount of digitized news media can provide useful information for identifying, tracking, and reporting sustainability indicators. Using San Mateo County as a case study, the results suggest that the approach has potential for reducing time and resources dedicated to the identification and tracking of sustainability problems specific to a region. Furthermore, the use of news media allowed incorporation of society's values in the process of selecting the indicators, a component that had been under developed in previous approaches. This element could be expanded in future work by incorporating more online news and social media sources. The method was also able to detect significant changes in the sustainability indicators and provide an explanation for such behavior in real time. All of these capabilities provide decision-makers with faster access to reliable information on the sustainability indicators at lower cost, enabling better planning aligned with the values and concerns of the community.

The limitation of the method proved to be the identification and geo-referencing of some of the most chronic problems in the region. This limitation can be attributed to the nature of the data given that not all topics are newsworthy all the time. Further research is needed to explore whether online news and social media sources, such as environmental blogs or Facebook groups, could provide more information on these chronic problems, or whether input from professionals and data analysis are still needed to address these problems. Additionally, the performance of the methods should be evaluated to assess its value when including these additional media sources, as their inclusion may require different text mining techniques or the combinations of different methods. Furthermore, the method should be evaluated in different regions across the nation and using a longer time series to more thoroughly assess its utility. Lastly, to further validate the approach and its usefulness, the opinions of experts and potential end users of the system should be evaluated. These interviews would help gather the necessary knowledge needed to further developed the text mining approach (e.g., by introducing new data sources) and determine appropriate ways to present the results for best use in different applications.

Data ownership and licensing of news media articles were a major challenge in this study that prevented a further analysis at this time. Nonetheless, current trends point towards an easier dissemination of news data in the future, and thus a larger potential to further investigate the approach.

Acknowledgments

We acknowledge the financial support of the College of Engineering Support for Under-Represented Groups in Engineering (SURGE) Fellowship program and the Campus Research Board Program at the University of Illinois at Urbana -Champaign.

Appendix A. Mathematical formulation of the term frequency – inverse document frequency (TF-IDF)

The TF-IDF takes into account the frequency of a term (word) in a document and all the other documents in the set to calculate a metric of importance of a term in a document relative to all the other words and documents. Mathematically, the TF-IDF can be expressed as the following function:

$$\mathrm{TFIDF}(t,d,D) = \mathrm{TF}(t,d)*\mathrm{IDF}(t,D)$$

where t is the term index, d is the document index, D is the set of documents, $\mathrm{TF}(\cdot,\cdot)$ is the normalized term frequency of term t in document d and $\mathrm{IDF}(\cdot,\cdot)$ is the inverse document frequency. The normalized term frequency, TF, represents the importance of a word within a document. As a function it can be denoted as $\mathrm{TF}(\cdot,\cdot)$:

$$\mathrm{TF}(t,d) = \frac{f(t,d)}{\max\{f(t,d) : t \in d\}}$$

where $f(t,d)$ is the number of occurrences of a term t in document d. The inverse document frequency, IDF, is a factor that accounts for the appearance of the same term t in other documents in the set D. For a document d, the IDF reduces the importance (weights) of terms that occur very frequently in the set of documents and increases the weights of those that occur rarely. Mathematically it is represented as $\mathrm{IDF}(\cdot,\cdot)$:

$$\mathrm{IDF}(t,D) = \log\frac{|D|}{|\{d \in D : t \in d\}|}$$

References

Adinyira, E., Oteng-Seifah, S., Adjei-Kumi, T., 2007. A review of urban sustainability assessment methodologies. In: International Conference on Whole Life Urban Sustainability and its Assessment.

Barkemeyer, R., Figge, F., Holt, D., Hahn, T., 2009. What the papers say: trends in sustainability: a comparative analysis of 115 leading national newspapers worldwide. J. Corp. Citizsh. 33, 69–86.

Beniger, J., 1978. Media content as social indicators: the greenfield index of agenda-setting. Commun. Res. 5 (4), 437–453.

Benton, M., Frazier, P., 1976. The agenda setting function of the mass media at three levels of 'information holding'. Commun. Res. 3 (3), 261–274.

Bossel, H., 1999. Indicators for Sustainable Development: Theory, Method, Applications. International Institute for Sustainable Development (IISD), Manitoba, Canada. Retrieved from. http://iisd.ca/about/prodcat/ordering.htm.

Carvalho, A., 2005. Representing the politics of the greenhouse effect: discursive strategies in the British media. Crit. Discourse Stud. 2 (1), 1–29.

Chang, M.-W., Ratinov, L., Roth, D., Srikumar, V., 2008. Importance of semantic representation: dataless classification. In: Proceedings of the National Conference on Artificial Intelligence, vol. 2, pp. 830–835.

Dahl, A., 2012. Achievements and gaps in indicators for sustainability. Ecol. Indic. 17, 14–19.

Dearing, J., Rogers, E., 1996. Agenda-setting. SAGE Publications, Thousand Oaks, CA.

Farr, D., 2008. Sustainable Urbanism : Urban Design with Nature. Wiley, Hoboken, N.J.

Feldman, R., Sanger, J., 2007. The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data. Cambridge University Press.

Gahin, T., Veleva, V., Hart, M., 2003. Do indicators help create sustainable communities? Local Environ. 8 (6), 661–666.

Hammond, A., Adriaanse, A., Rodenburg, E., Bryant, D., Woodward, R., 1995. Environmental Indicators: a Systematic Approach to Measuring and Reporting on Environmental Policy Performance in the Context of Sustainable Development. World Resources Institute, Washington, D.C.

Hastie, T., Tibshirani, R., Friedman, J., 2009. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer, New York.

Holt, D., Barkemeyer, R., 2012. Media coverage of sustainable development issues - attention cycles or punctuated equilibrium? Sustain. Dev. 20 (1), 1–17.

International Institute for Sustainable Development (IISD), 2000. International Institute for Sustainable Development Annual Report, 2000.

International Institute for Sustainable Development (IISD), 2013. IISD Compendium of Sustainable Development Indicator Initiatives. Retrieved, 2013, from. http://www.org/publications/pub.aspx?id=607.

Innes, J., Booher, D., 2000. Indicators for sustainable communities: a strategy building on complexity theory and distributed intelligence. Plan. Theory Pract. 1 (2), 173–186.

Jakeman, A.J., Chen, S.H., Rizzoli, A.E., Voinov, A.A., 2008. Modelling and software as instruments for advancing sustainability. In: Environmental Modelling, Software and Decision Support: State of the Art and New Perspectives, vol. 3, pp. 1–13.

Krank, S., Wallbaum, H., 2011. Lessons from seven sustainability indicator programs in developing countries of Asia. Ecol. Indic. 11 (5), 1385–1395.

Krueger, T., Page, T., Hubacek, K., Smith, L., Hiscock, K., 2012. The role of expert opinion in environmental modelling. Environ. Model. Softw. 36, 4–18.

Laniak, G.F., Olchin, G., Goodall, J., Voinov, A., Hill, M., Glynn, P., et al., 2013. Integrated environmental modeling: a vision and roadmap for the future. Environ. Model. Softw. 39, 3–23.
Leetaru, K., 2011. Culturomics 2.0: forecasting large-scale human behavior using global news media tone in time and space. First Monday 16 (9), 2–2.
Li, T., Zhu, S., Ogihara, M., 2008. Text categorization via generalized discriminant analysis. Inf. Process. Manag. 44 (5), 1684–1697.
Michel, J., Yuan, K., Aiden, A., Veres, A., Gray, M., Pickett, J., Hoiberg, D., Clancy, D., Norving, P., Orwant, J., Pinker, S., Nowak, M., Aiden, E., 2011. Quantitative analysis of culture using millions of digitized books. Science 331 (6014), 176–182.
Miner, G., Elder, J., Hill, T., Delen, D., Fast, A., 2012. Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications. Academic Press.
Mishne, G., Glance, N., 2006. Predicting movie sales from blogger sentiment. Technical report, SS-06-03. In: AAAI Spring Symposium, pp. 155–158.
Moldan, B., Janoušková, S., Hák, T., 2012. How to understand and measure environmental sustainability: indicators and targets. Ecol. Indic. 17, 4–13.
Morse, S., 2008. Post-sustainable development. Sustain. Dev. 16 (5), 341–352.
Mueller, J., 1973. War, Presidents and Public Opinion. Wiley, New York.
Naisbitt, J., 1976. The Trend Report: a Quarterly Forecast and Evaluation of Business and Social Development. Center for Policy Process, Washington, DC.
National Fish and Wildlife Foundation, 2013. Cosco Busan Oil Spill Settlement – Recreational Use Grant Program. Retrieved, 2013, from. http://www.nfwf.org/Pages/coscobusanrec/home.aspx#.UeP3NMrjL5U.
National Center of Supercomputing Applications, 2013. Meandre - Documentation. University of Illinois at Urbana-Champaign. Retrieved, 2013, from. http://www.seasr.org/meandre/documentation/.
Nelkin, D., 1995. Selling Science: How the Press Covers Science and Technology. Freeman, New York.
Rockström, J., Steffen, W., Noone, K., Persson, Å., Chapin, F., Lambin, E., Lenton, T., Scheffer, M., Folke, C., Schellnhuber, H., Nykvist, B., De Wit, C., Hughes, T., Van Der Leeuw, S., Rodhe, H., Sörlin, S., Snyder, P., Costanza, R., Svedin, U., Falkenmark, M., Karlberg, L., Corell, R., Fabry, V., Hansen, J., Walker, B., Liverman, D., Richardson, K., Solomon, S., Plattner, G., Knutti, R., Friedlingstein, P., 2009. A safe operating space for humanity. Nature 461 (7263), 472–475.
Rogers, E., Dearing, J., Bregman, D., 1993. The anatomy of agenda-setting research. J. Commun. 43, 68–84.
Scerri, A., James, P., 2010. Accounting for sustainability: combining qualitative and quantitative research in developing 'indicators' of sustainability. Int. J. Soc. Res. Methodol. 13 (1), 41–53.
Scharl, A., Hubmann-Haidvogel, A., Weichselbraun, A., Lang, H., Sabou, M., 2013. Media watch on climate change - visual analytics for aggregating and managing environmental knowledge from online sources. In: Proceedings of the Annual Hawaii International Conference on System Sciences, pp. 955–964.
Silla Jr., C., Freitas, A., 2011. A survey of hierarchical classification across different application domains. Data Min. Knowl. Discov. 22 (1–2), 31–72.
Solomon, S., Plattner, G., Knutti, R., Friedlingstein, P., 2009. Irreversible climate change due to carbon dioxide emissions. Proc. Natl. Acad. Sci. U. S. A. 106 (6), 1704–1709.
Sun, A., Lim, E., 2001. Hierarchical text classification and evaluation. In: Proceedings – IEEE International Conference on Data Mining. ICDM, pp. 521–528.
Sustainable San Mateo County, 2008. Retrieved, 2012, from. http://www.sustainablesanmateo.org/indicators-report/.
Sustainable San Mateo County, 2010. Retrieved, 2012, from. http://www.sustainablesanmateo.org/indicators-report/.
Szerszyski, B., Urry, J., Myers, G., 2000. Mediating global citizenship. In: Smith, J. (Ed.), The Daily Globe Environmental Change, the Public and the Media. Earthscan, London, pp. 97–114.
Tang, X., Yang, C., Zhou, J., 2009. Stock price forecasting by combining news mining and time series analysis. In: Proceedings – 2009 IEEE/WIC/ACM International Conference on Web Intelligence, WI 2009, vol. 1, pp. 279–282.
The Apache Software Foundation, 2013. Apache OpenNLP Developer Documentation. Retrieved, 2013, from. http://opennlp.apache.org/documentation/1.5.3/manual/opennlp.html.
Thøgersen, J., 2006. Media attention and the market for 'green' consumer products. Bus. Strategy Environ. (John Wiley Sons, Inc) 15 (3), 145–156.
United Nations Department of Economic and Social Affairs (UNDESA), 2010. 2009 Revision of World Urbanization Prospects.
Voinov, A., Seppelt, R., Reis, S., Nabel, J.E.M.S., Shokravi, S., 2014. Values in socio-environmental modelling: persuasion for action or excuse for inaction. Environ. Model. Softw. 53, 207–212.
Weiss, S., Indurkhya, N., Zhang, T., 2010. Fundamentals of Predictive Text Mining, first ed. Springer Publishing Company, Incorporated.
Yli-Viikari, A., 2009. Confusing messages of sustainability indicators. Local Environ. 14 (10), 891–903.