A Visual Analytics Framework for Identifying Topic Drivers in Media Events

Yafeng Lu, *Member, IEEE,* Hong Wang, Steven Landis and Ross Maciejewski, *Senior Member, IEEE*

Abstract—Media data has been the subject of large scale analysis with applications of text mining being used to provide overviews of media themes and information flows. Such information extracted from media articles has also shown its contextual value of being integrated with other data, such as criminal records and stock market pricing. In this work, we explore linking textual media data with curated secondary textual data sources through user-guided semantic lexical matching for identifying relationships and data links. In this manner, critical information can be identified and used to annotate media timelines in order to provide a more detailed overview of events that may be driving media topics and frames. These linked events are further analyzed through an application of causality modeling to model temporal drivers between the data series. Such causal links are then annotated through automatic entity extraction which enables the analyst to explore persons, locations, and organizations that may be pertinent to the media topic of interest. To demonstrate the proposed framework, two media datasets and an armed conflict event dataset are explored.

Index Terms—Semantic Similarity, Media Annotation, Visual Analytics, Causality Modeling, Social Media.

1 Introduction

As citizen news reports, micro-blogs and other media outlets have increased, a variety of tools have emerged for analyzing media data collections. Such tools tend to focus on topic extraction [1], [2], event detection [3], and information flows [4], [5] as a means of quickly assessing the development of ongoing stories. However, recent work often focuses on exploring evolving media discourse in isolation. What is needed are tools and methods that can enable analysts to link together multiple data sources of interest in order to identify patterns and drivers that exist between datasets that are not fully captured or represented in any single dataset alone. To that end, new technologies have been developed for fusing media data and secondary data sources to provide contextual information. For example, recent work from Wanner et al. [6] and Hullman et al. [7] explored methods for time series analysis to identify text features of interest in conjunction with quantitative phenomena observed in stock prices, Liu et al. [8] proposed *TextPioneer* with a combination of hierarchical tree visualization and a twisted-ladder-like visualization to present and analyze the lead-lag patterns in a topic between different corpus, and work by Lu et al. [9] explored methods for identifying intervention points in news media data to cue analysts into the exploration of secondary datasets of interest.

However, fusing datasets and providing means of identifying and annotating potential temporal drivers is still fraught with challenges. For example, imagine collecting a corpus of text from Twitter discussing a sale product (e.g., tennis shoes) as well as a collection of product reviews on tennis shoes from Amazon. In this case, an analyst may want to see if discussion on Twitter is driving ratings and comments in the product reviews. Challenges here could be that the language used on Twitter and the language used on the product review site do not have a one to one matching (e.g., "This shoe is sick" could be counted as a positive review, but the language on the product review site may use less slang). As such, keyword searches to filter the document collections to only positive reviews may not be able to rely on traditional topic modeling tools and often need domain expert intervention. Once a dataset is curated, then further automated analysis to explore drivers between the datasets must be performed (for example, do positive tweets about a product proceed positive product reviews?). Annotations of key events in the timeline and key actors in the text corpora are also relevant and need to be annotated in the hypothesis exploration and analysis phase to help the analyst navigate large document collections and identify key components of the event drivers.

As such, this paper proposes a visual analytics framework (Figure 1) that focuses on the exploration, linkage, and annotation of multiple media sources to explore drivers of discourse. First, we apply semantic matching to identify keywords and concepts that an analyst considers to be related between two datasets. A novel widget enables domain experts to quickly cluster, split, and merge keywords from a semantic dictionary to ensure that meaningful similarities are captured through a visual to parametric interface while allowing for analyst-guided language disambiguation. While there are known limitations of keyword searches, by enriching an analyst's choice of keywords with semantic meaning, we enable a broader matching that better aligns with the user's mental model. In this way, we move away from searching on one (or several) keyword(s) and instead search on semantic meanings that embeds the analyst's domain knowledge into the search.

Once users are satisfied with the semantic grouping of keywords, filtering is performed and a raw count of semantically related articles and events per time step can be extracted from each of the time-oriented textual datasets. Using these time series, a secondary annotation step is performed where causality measures are applied to the derived time series to extract possible drivers. These causality measures could indicate that past events contained in time series A contain information that can help predict time

- *Yafeng Lu, Hong Wang and Ross Maciejewski, are with Arizona State University.*
  E-mail: {lyafeng, hxwang, rmacieje}@asu.edu.
- *Steven Landis is with University of Nevada, Las Vegas*
  E-mail: steven.landis@unlv.edu

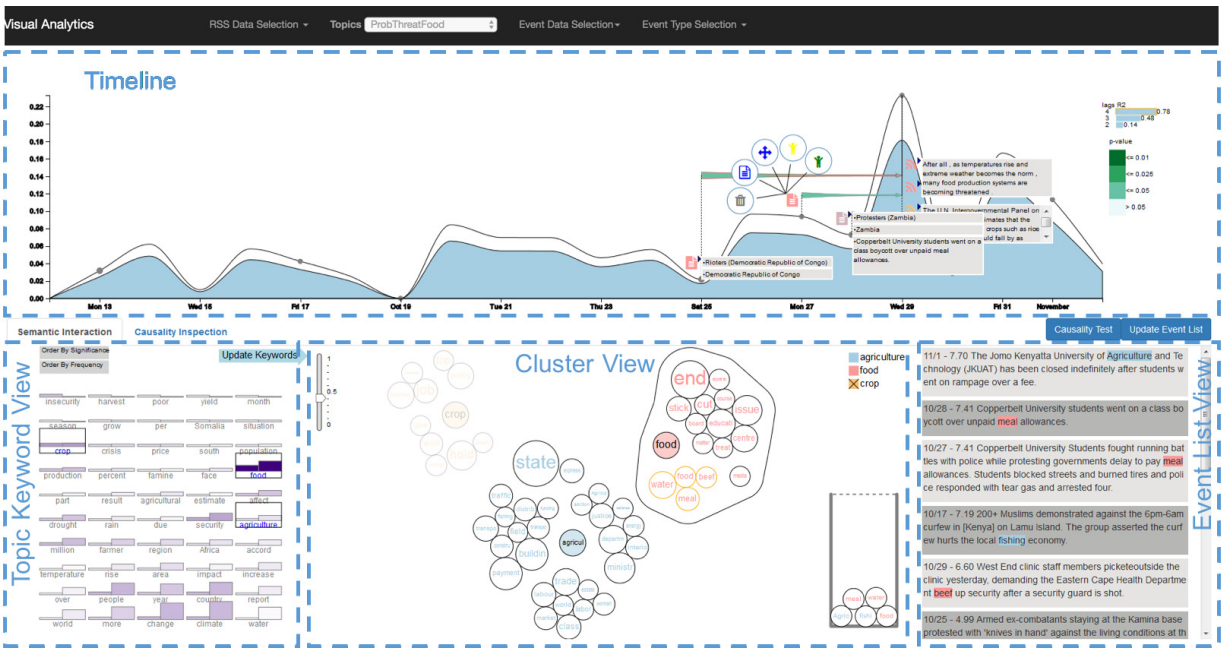


Fig. 1. Semantic annotation framework exploring the period between Oct.12 and Nov.2, 2014 of the climate change media dataset and the ACLED dataset. The Timeline shows annotations of possible drivers of media reports on climate change framed around food insecurity. Keywords "agriculture" and "food" are used for the semantic mapping between the media topic and the text content in the ACLED dataset. The causality model indicates a goodness of fit $R^2 \approx .8$ with $p\text{-}value \leq .05$. Actors, locations and descriptive text are annotated on the Timeline.

series B above and beyond the information contained only in time series B. If a causal link is established, the framework then indicates the temporal lag under which causality was identified and provides interactions to further filter and annotate the time series based on relationships between locations, actors and other derived information. Contributions include:

1) A user-guided semantic lexical matching scheme.
2) Applied causality metrics for identifying media drivers.
3) A causality-driven annotation scheme for exploring potential media drivers.

Thus, our framework enables the development of narratives that cross beyond the boundaries of a single corpus and enable exploration into external events that may serve as drivers to discourse. In order to demonstrate the effectiveness of our work, we have partnered with experts from political science to explore media drivers with respect to armed conflicts. While media streams are the primary source for annotating conflict events, when extracting topics, for example climate change or election cycles, the text corpus related to a topic will likely be disjoint from topics of armed conflicts. While these topics can be explored in parallel, there is a need to extract, explore and annotate specific events between separate topics in order to formulate and explore hypotheses about narrative drivers. In this work, we frame our discussion on two case studies: exploring armed conflict events and their relationship to how the media is framing climate change stories, and; exploring the plausibility of climate-induced civilian abuse during the 2014 Greater Horn of Africa drought. While our case studies are specific to our domain expert's area, our framework is flexible to allow ingestion and semantic annotation between multiple sources of data, topics, and events. To apply this framework to general textual datasets, a preprocessing step that includes topic modeling and word pair similarities is required.

2 Related Work

In this section, we review related work on visual text analytics and annotation to position our contributions.

2.1 Visual Text Analytics

In fields ranging from law to journalism to science, the need to organize documents and synthesize knowledge has been an underlying driver for many text analytics algorithms and systems. In the SPIRE system [10], documents are clustered by keywords and projected to a 2D space through dimension reduction techniques in order to extract themes. This idea of finding themes, trends, topics and narratives across collections of documents has been an ongoing subject of research. ThemeRiver [5] explored the evolution of keyword-based topics, visualizing document collections as thematically labeled stacked areas over time to reveal trends. More recent work has focused on topic modeling techniques (e.g. Latent Dirichlet Allocation [11]) for document summarization [2], document clustering [1], topic evolution [4], and competition analysis [12], [13]. In addition to analyzing themes, work such as LeadLine [3] has focused on associating topical themes with events that may be driving changes in discourse. LeadLine applies event detection methods to detect "bursts" from topic streams and further associates such bursts with people and locations to establish meaningful connections between events and topics. Similar to the work of LeadLine [3], our proposed framework also focuses on identifying discourse drivers in time-varying media collections. However, instead of detecting events in the same text collection, as is done in LeadLine, our work focuses on fusing multiple data sources to provide external context that may not be captured in the primary data stream. We focus on utilizing secondary datasets to link contextual information and investigate possible causality relationships between data sources.

While many visual text analytics systems, including those discussed above, focus on analyzing documents from a single data source (e.g., scientific literature collections [14], [15], Wikipedia articles [16], and social media messages [12], [13], [17]), more recent approaches have begun exploring the integration of multiple data sources in visual text analytics. For example, Narratives [18] combines keywords from news articles with reactions from social media and visualizes them on a line graph associated with those keywords. Work from Wongsuphasawat et al. [19] monitors events from large-scale logs on Twitter and links them to commercial products. Vox Civitas [20] is a visual analytics tool designed to help journalists and media professionals extract news from large-scale aggregations of social media content around broadcast events. webLyzard [21] is a visual analytics system which aggregates news and social media content for interactive exploration and knowledge extraction. Scharl et al. [22] proposed a visual analytics tool, Westeros Sentinel, which explores the text content of news media together with four social media platforms in order to analyze public opinion towards television shows. Their work integrates topic analysis, sentiment analysis, and content-aware semantic processing to analyze factual and affective information from the text. Our work builds upon previous work and further introduces causality analysis as a mechanism for framing the relationship between multiple textual data sources to assist the annotation of media topics and ties to potential driving-effects.

Other than combining media and social media data, work such as Contexifier [7] has focused on linking news stories to stock prices to detect and highlight events which might cause price changes. Similar to Contexifier, Wanner et al. [6] develop an integrated visual analysis system for stock market data and media content to help economists identify text properties in news articles that might affect stock prices. Their focus is on time series pattern identification as a means of identifying potential drivers of market fluctuations. Our work is similar in the sense that we are exploring methods of identifying causal drivers of the media stories from linked secondary events; however, our focus is on linking media and events based on semantic similarities and then assessing if drivers between two semantically similar streams exist (as opposed to identifying temporal patterns of interest). This is similar to Diakopoulos et al.'s work [20], which uses message similarity to filter for related responses to an event; however, we provide new methods for defining semantic relevance as well as advanced filtering, analysis, and annotation utilizing causality metrics to identify potential drivers.

2.2 Annotation

Along with linking secondary datasets and identifying drivers, we also focus on methods for annotating relationships between media streams. Document annotation has been used in many visual analytics systems and has been shown to be an effective way to organize information and transform it into knowledge. For example, Zheng et al. [23] proposed a structured annotation approach that uses a unified annotation model to record and organize co-authors' insights in document revision tasks. Click2Tag [24] and Fingerprint [24] are systems developed to help users generate and browse annotations for online documents, and Click2Annotate [25] allows insight externalization and semi-automatic annotation of features such as clusters and outliers. Contextifier [7] provides customized annotations for stock timeline graphs with references to the content in a news article, and NewsViews [26] is a geo-spatial visualization system that generates interactive, annotated thematic maps for news articles. NewsViews supports trend identification and data comparison relevant to a given news article. Similar to our work, the annotation text in NewsViews [26] is derived not only from media articles but also from other sources that are geospatially related to the article. Other annotation work includes TimeLineCurator [27], which is an interactive authoring tool that extracts event data and generates annotated timelines based on temporal information in an article. Our work expands on these with additions of semantic interactions for event linkage, providing measures of causality to help directly identify statistically significant drivers, and enabling annotation between datasets through entity extraction and analysis.

3 Framework Design

This framework has been designed for researchers to explore and annotate the underlying driving effects between media datasets (text-based). Methods and techniques utilized in this work were developed based on feedback from our previous work [9] and discussions with collaborators from political science and communication. The communication experts consist of a professor and a researcher in the university communication school who were analyzing the relationship between climate change and social unrest. Our collaborator from political science is one of the authors of this paper. These domain scientists have text data collected from news outlets, social media platforms, and curated event databases. This section introduces the analytic tasks our collaborators have and the design requirements derived from their tasks.

3.1 Analytic Tasks

As an example analysis, we consider a scenario in which a political scientist wants to explore the relationship between local conflicts and the 2015 Nigerian election. Prior to the Nigerian election, domain experts had hypothesized that widespread riots and social unrest would occur as part of the election cycle. The domain expert wanted to analyze social unrest news articles and local conflict events in Africa during the time leading up to the election to inspect if the election campaign media was a precursor to violence (or vice-versa). First, the data for analysis is collected from two sources (news media and a curated event dataset documenting violent conflicts in Africa). These datasets are a superset of the data needed for analysis. Thus, the first task is filtering the data for relevant text. Once a subset of the media posts and the conflict event records are curated, the documents need to be checked for relevancy, and a temporal aggregation must be performed to enable cause-effect relationship analysis. In this case, the analyst is interested in relationships between the subset of media data related to the election in Nigeria and the subset of conflict events in Africa occurring in Nigeria. As key events, actors and locations are identified in the data, the analyst also needs to annotate their findings in order to explain the events and the news contents and support the hypothesis generation and explanation phase.

Given this example analysis task, we generalize the analyst's workflow into several key steps for the identification of topic drivers in media data:

1) Identify topics within the datasets in order to explore their evolution over time;
2) Link and filter datasets so that the extracted media items are relevant to the analysis;

3) Identify critical entities within the datasets, and;
4) Provide methods for cause-effect identification and identify potential leading or lagging indicators.

3.2 Requirements

From the proposed task workflow, we discussed possible visual analytics solutions with our domain collaborators and from these discussions, we focused on deriving their key analysis needs. The functional requirements from our domain collaborators can be summarized as:

- When exploring multiple media (text) datasets, visualizing more than co-occurrence is desired. Metrics that detail cause-effect relationships should be prioritized.
- The relationship between event drivers should be temporal.
- Raw text should be easy to access for detailed analysis.
- Important entities should be highlighted for a quick summary of events.

Along with requirements from our collaborators, we also have incorporated principles from the visualization literature into our framework design.

- Interactions should maintain (when possible) visual coherency so that smooth updates occur in the visualization.
- Views should be automatically synchronized for linked exploration and analysis.
- The perception of the meanings embedded in the visual encoding should be intuitive to the domain experts.

We incorporate the functional requirements generated from our collaborations with domain experts and visualization design principles abstracted from literature. The key designs of the proposed framework can be summarized into three main components.
Content + Temporal: The textual analysis of multiple sources of data requires both content analysis to find shared information and temporal analysis to identify correlations and causality with lead and lag. The proposed framework links the datasets based on textual content and posted time and performs causality testing on multiple different lags.
Summary + Detail: A quick summary of the data can facilitate text browsing before locating a subset of data for detailed investigation. In addition, the details of the data are also needed to validate an analyst's hypothesis and reconstruct the story. This framework provides word summaries as well as raw text after an initial search and filter. In the annotation step, short term entities, icons, and editable texts are available to represent discovered events.
Interactive: Interactions should be enabled directly on the visualization and no drastic change of the views should occur when the data updates. To enable interactive steering, the proposed framework automatically filters the data and updates the linked views after each interaction to minimize the number of required user actions while still providing direct and intuitive data access to the users [28]. The visual encodings are designed to be intuitive to users, and the design has focused on the color, the icons, and the linking relationships provided during the analysis results.

3.3 Framework Overview

The goal of our visual analytics framework is to facilitate the annotation of time-oriented media collections and semantically linked events through querying, filtering, and causality testing. Our framework consists of three main views:

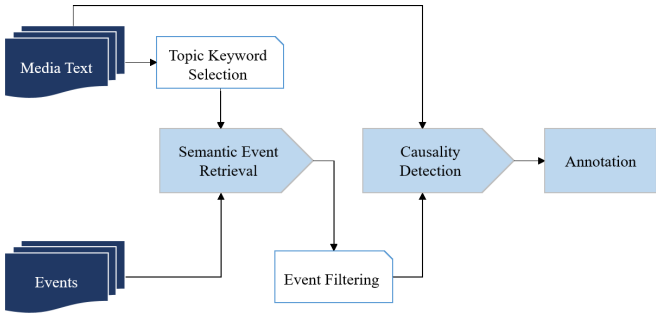
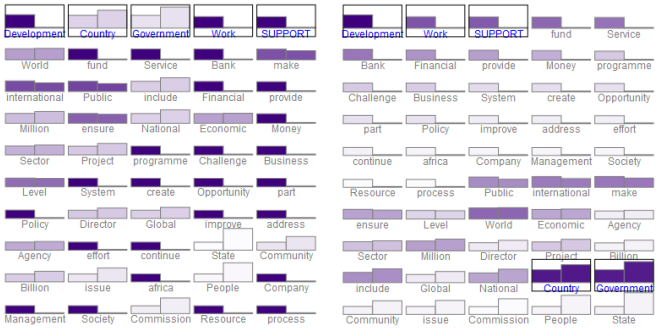


Fig. 2. The framework contains three key analytics steps: semantic event retrieval, causality detection, and annotation.

1) The Cluster View, which displays the semantic clusters of related words organized by concepts between the primary and secondary data sources and enables semantic interactions to improve data fusion and annotation;
2) The Bipartite View, which connects the media articles and events indicated by the causality models to analyze document and keyword correlations, and;
3) The Timeline view, which visualizes the media topic volume, the semantically linked events volume, the statistical results from causality models, and annotations added by the analyst.

In addition to these three main views, our framework also contains a Topic Keyword View for selecting representative keywords from media topics, an Event List view to show relevant events sorted by a semantic similarity measure displayed to the right of the Cluster View on the Semantic Interaction tab, and two detail views displaying the complete text of the media articles and linked events to the sides of the Bipartite View on the Causality Inspection tab.

The analytical cycle of our semantic annotation framework is illustrated in Figure 2. The analyst begins by selecting a topic (or other categorization) from the desired media dataset and choosing a secondary data source that the analyst feels is relevant to the chosen topic. The selected media topic is then summarized in a Topic Keyword View, where the analyst can select important keywords to begin building a semantic dictionary for linking the media topic to the secondary data source. Next, a semantic lexical match method is developed to link the secondary dataset and identify documents of interest. Documents in the secondary dataset are extracted based on the applied semantic filtering and organized and presented to the analyst based on a semantic similarity score. The analyst can refine the keyword selection as well as interactively identify and group relevant synonyms in the semantic dictionary. Once the analyst is satisfied with the applied semantic filtering, the relevant events in the secondary dataset are aggregated as a time series and compared to the media stream. Causality modeling is applied between the two time series to suggest relevant events as well as identify statistically significant lags and leads. The analyst can refine keyword selection and document extraction and re-run the causality tests on these different inputs. The analyst can annotate the documents on the Timeline to externalize useful knowledge discovered during their hypothesis generation.



(a) Ordered by frequency (b) Ordered by significance

Fig. 3. The Topic Keyword View shows the 50 most frequent topic keywords for *Economy* in the social unrest media collection. Keywords are ordered by their frequency in (a) and by topic significance in (b). Color refers to the other measure not ordered by. The two bars of each keyword represents its frequency inside and outside the topic respectively.

4 Semantic Event Retrieval

4.1 Semantic Lexical Match

In many application areas, the key to successful data analysis and reasoning involves integrating data from different sources, for example, linking financial data with news reports may help analysts develop models for predicting stock market responses [6]. One critical task in linking multiple datasets is performing text query matches based on document similarities. This task usually leverages information retrieval methods [29]; however, many media posts contain short text messages or other limited information which may not contain the specific query word, restricting the effectiveness of simple keyword matching. In order to solve this problem, word semantic similarity measures have been studied [30], [31]. The general idea is to match the word sets from text segments by pairing every word in one data collection with its most similar word from the other data collection and then calculate a weighted sum of all pairs. Though this method can be used to measure the semantic similarity between text segments, there are two major challenges in querying for relevant events in media:

1) One dataset under exploration may have a wide word coverage which would tend to increase the maximum similarity measurements between the two datasets.
2) The knowledge-based word similarity measure returns the highest similarity score among all possible word senses [32]; however, not all word senses are relevant to the analyst's semantic definition. For example, the word "crop" can relate to agriculture or a type of haircut, but an analyst studying agriculture would likely be uninterested in articles about hair care.

Topic Keyword View: In order to reduce the issues with word coverage, our framework first extracts a list of keywords from the selected media topic. Next, the analyst can explore the uniqueness of these keywords through the Topic Keyword View, Figure 3. In this view, a small multiples bar graph is used to show the 50 most frequent keywords. The height of the left bar represents the frequency of the word with respect to documents that are classified into the selected topic. The height of the right bar represents the frequency of the word in all other documents in the dataset. This view has two options for ordering these keywords, either by the word frequency in this topic or by the significance of this word with respect to all other topics in the dataset. The significance measure of a keyword w to a topic d is defined by:

$$significance(w,d) = \frac{f(w,d) - \sum_{t \in T, t \neq d} f(w,t)}{\sum_{t \in T} f(w,t)}, \qquad (1)$$

where T denotes the set of all topics extracted from the media collection, $f(w,t)$ is the frequency of word w in topic t, and d is the topic under analysis. The range of this metric is $[-1,1]$ where values closer to 1 means the word is more significant in the chosen topic. This measure is also perceptually visible based on the height of the two bars. If the left bar is taller than the right bar, the significance value is positive; if the left bar is shorter than the right bar, the significance value is negative. Analysts can select keywords by clicking on the bar graph and they will be highlighted by a rectangular box. In Figure 3, the five most frequent words in the topic, *Economy*, have been selected and are shown at the first line by the frequency order (Figure 3(a)). When the view is reordered based on the significance metric, three of the five most frequent keywords are also listed as the top five most significant keywords while the other two fall into the last ten words of the list (Figure 3(b)). Using this view, our framework provides a keyword selection reference so that the analysts can choose representative words for the media topic, thereby injecting domain knowledge into the semantic annotation pipeline while reducing the word set chosen for semantic matching.

Similarity Measure: Once keywords are chosen, semantic matching to identify relevant links between the datasets is performed. We calculate a semantic similarity score between the selected keywords and the documents in the secondary dataset which is then filtered by this score and the documents that have a high semantic similarity score are returned for evaluation. The detailed calculation is introduced as follow.

We use a knowledge-based word semantic similarity metric, Wu and Palmer [33], to first calculate the word-word similarity between selected media topic keywords and all words in the secondary dataset. This metric measures the depth of two given senses in the WordNet taxonomy [34], along with the depth of the least common subsumer (LCS). The sense-to-sense similarity is calculated as follow:

$$SenseSim = \frac{2 \times depth(LCS)}{depth(sense1) + depth(sense2)}$$

This is a sense-to-sense similarity measure, but it can be used as a word-to-word similarity measure by selecting the highest similarity score among all the similarities between the senses of these two words. Thus, word similarity can be defined as $WordSim = Max(SenseSim)$, which is a score between 0 and 1.

Given a keyword set representing the media topic and the word-to-word similarity measure, a semantic similarity metric can be developed to measure the relatedness of a document to this media topic. The media topic can be described as a set of keywords $K = \{k_1, k_2, \ldots, k_m\}$ where k_i is one of the m keywords. Similarly, the document in the secondary dataset can also be represented by a set of words $E = \{w_1, w_2, \ldots, w_n\}$ where w_i is the word occurring in the document from the secondary dataset. Note that both the media keywords in K and the words in the secondary dataset E are preprocessed to remove stop words and are lemmatized using CoreNLP [35] for consistency. Using the above notations, our

similarity score between a topic in dataset one and a document in dataset two is calculated as follows:

$$EventSim(E,K) = \sum_{\substack{w,\exists k_i \in K,\\ WordSim(k_i,w)>\theta}} \delta_w \text{tfidf}(w,E), \qquad (2)$$

where θ is a threshold to filter for semantically similar words (by inspection, $\theta = 0.8$ was a reasonable choice and is used as the threshold values for all examples in the paper), the tf-idf is used to weight the word's importance, and $0 \leq \delta \leq 1$ is a weight for each semantically matched word. The value of δ is initially set to be 1 for all words, and δ can be changed during the visual to parametric interaction methods that are described in section 4.2. We use the augmented frequency for $\text{tf}(w,E)$ to prevent a bias towards longer documents, where

$$\text{tfidf}(w,E) = \left(0.5 + \frac{0.5 \times f_{w,E}}{max\{f_{w',E} : w' \in E\}}\right) \times idf(w,D)$$
$$idf(w,D) = log\frac{N}{|\{E \in D : w \in E\}|},$$

and w is one word in a document of the secondary dataset E, and the whole secondary data collection is D, the size of which is N. We use the logarithmically scaled inverse document frequency for $\text{idf}(w,D)$. This approach returns a list of documents ordered by their similarity scores together with the words that are semantically similar to at least one of the media topic keywords. The semantically matched documents from the secondary dataset are shown in the Event List view (Figure 1) since in our analysis each document in the secondary dataset describes one event.

4.2 Semantic Similarity Update

As previously mentioned, a direct calculation of semantic similarity from a knowledge base has issues with one keyword semantically belonging to multiple related concepts, for example, if you look at relationships for "food" in the knowledge base, both "bread" and "education" appear; however, these represent two very different concepts which may not be the intention of the analyst. Thus, we have developed a visual to parametric interface, building on the conceptual work of Leman et al. [36], in which the analyst can cluster the concepts returned from the knowledge base to better refine the semantic similarity matching. In addition to clustering concept words, the analysts can mark entries returned from the secondary dataset as relevant or irrelevant which will update the word similarity weight (δ in Equation 2) thereby modifying the semantic scores and reorganizing the event list.

4.2.1 Semantic Interaction for Concept Word Clustering

Again, even though a chosen keyword may semantically match a word in the secondary dataset, this matching may not align based on the contextual concept in which an analyst is working. For example, the word "food" has the following three noun senses:

- food#n#1: any substance that can be metabolized by an animal to give energy and build tissue;
- food#n#2: any solid substance (as opposed to liquid) that is used as a source of nourishment, and;
- food#n#3: anything that provides mental stimulus.

If an analyst wants to relate concepts of food and agriculture, sense 1 and 2 are likely related to the semantic search; however, sense 3 is unrelated.

Cluster Force Layout: Based upon the interface design of IN-SPIRE [10] that uses word clusters to represent document themes, we have developed a cluster force layout to group words in the semantic dictionary based on their word-to-word similarities. Our contributions include methods that enable the analyst to steer and update this clustering in order to develop an appropriate concept map for semantic annotation. To separate words by their meaning, we use complete-link agglomerative hierarchical clustering [37], in which the similarity between two clusters is decided by the smallest similarity of all word pairs between the two clusters. To develop a concept map, an analyst can interact with the cluster force layout (Figure 1 bottom middle). In the layout, each node represents a word. The nodes with a solid background represent the selected keywords from the media data, and the nodes without a background color represent the words extracted from the secondary dataset that are semantically related to the selected keywords. This view uses a categorical color scheme to separate different keyword bubble groups. Words in the secondary dataset are attracted to their corresponding keyword in the media data, and words belonging to the same cluster are further attracted together. Collision detection is applied to ensure that the nodes do not overlap each other, and nodes in different clusters are separated by a larger margin. As a result, each keyword and its semantically related words naturally form a bubble group, along with internal bubble clusters formed by the clusters of the semantically related words. We call the former the keyword bubble group and the after the bubble cluster. The nodes in each keyword bubble group are colored to match the keyword legend on the top right corner. The size of each node is proportional to the frequency at which the word appears in the event records. In the case where a word becomes too small to see, the analyst can mouse over the nodes to show the words in a tooltip. Sometimes the analyst may select many keywords and the keywords may contain many semantically related words, then we will not have enough space to display all the words in the view. In practice, a 900×450 space for this view can support 4 keyword bubble groups with around 100 semantically similar word bubbles. To enable the capability of analyzing more words, we also allow zooming and panning on the cluster view. The analyst can also freely drag the keyword nodes and the cluster nodes to adjust their relative position. As the clusters move close to each other, an attractive/repulsive force will be activated based on the similarity between the two clusters. This similarity score is calculated by taking the average of all word pair similarities. If the similarity between the two clusters is greater than 0.5, the clusters will attract each other, otherwise, they will repel each other.

Implementation of Cluster Force Layout: We used the force directed layout from the d3 library. The clusterings were formed by adding a gravitational force to each of the selected keywords such that the keywords would only attract the words that belong to their clusterings. Then, the clusters were formed by adding a force between words that belong to the same clusters. We also added a collision detection force to prevent the nodes from overlapping each other, and the collision detection force will separate nodes in different clusters by a larger space than nodes in the same clusters. To stabilize the force layout and prevent jittering during interaction, we also added a repulsion force between each node to neutralize the other attracting forces when the layout has already formed its shape to prevent the nodes from constantly colliding.

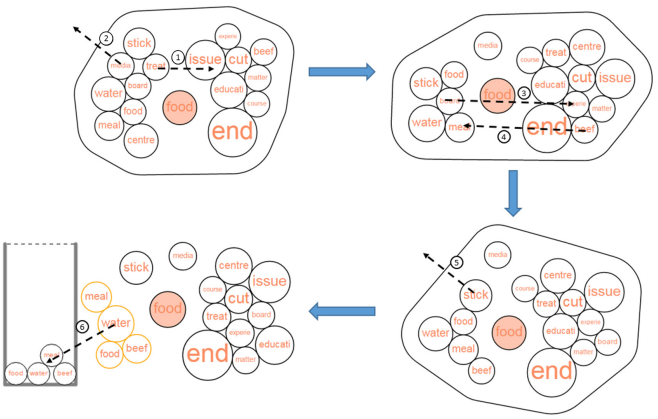


Fig. 4. This figure shows an interaction process of steering the cluster force layout to update the word concept clusters. Each step is marked numerically on the figure. User can drag a word away from all clusters or towards another cluster. User can drag a group of words to select them.

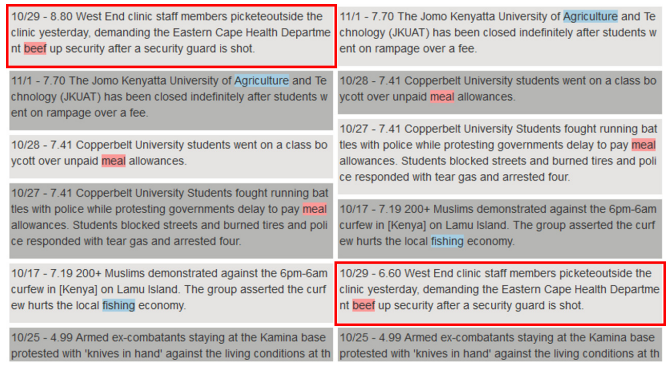


Fig. 5. Interaction on the Event List view to update word weight and the event semantic similarity.

However, when a cluster moves close to another cluster that belongs to a different clustering, the repulsive force would turn into an attractive force toward the nodes between the two clusters. Also, whenever the user drags any clusters on the layout, the strength of the attractive forces for the cluster will be slightly increased to ensure the cluster shape is preserved.

Semantic Interaction: In addition to visualizing word clusters, the cluster layout allows analysts to select sub-clusters and filter out words through semantic interactions [38], [39], [40]. The underlying concept is that by allowing users to directly manipulate data in the visualization space, updates to the positions of data elements on the screen can be tied back to weights in the analytic modules on the backend, which can then be translated to the model updates. Our cluster force layout supports semantic interactions for creating concept clusters. Here, a user can change the number of clusters by dragging the slider on the top left corner to set the similarity threshold in the hierarchical clustering and change the similarities between words through drag and drop interactions on the bubbles. Let k denote the number of clusters shown in one keyword bubble group and its word set can then be represented by clusters $C_i = \{w_{ij}\}$, where $i = 1,2,\ldots,k$, $j = 1,2,\ldots,n_i$, such that cluster C_i contains n_i words. An analyst can drag a word, w, from its current cluster C_i to another cluster $C_{i'}$ such that the similarities between w to all other words in C_i decrease and the similarities between w to all words in $C_{i'}$ increase while the similarities between w to the words not in C_i nor $C_{i'}$ do not change. The new similarities of word w to other words updates as follows:

$$sim'(w,w') = \begin{cases} sim(w,w') + (1 - sim(w,w')) \times 0.1, & w' \in C_{i'} \\ sim(w,w') \times 0.9, & w' \in C_i \\ sim(w,w'), & otherwise \end{cases}$$

Here $sim(w,w')$ represents the similarity between the word w and w' before the interaction, and $sim'(w,w')$ represents the new similarity after the interaction. If the analyst finds that there are no clusters that the word w can join, he can try to reduce the similarity between this word and all the other words in the keyword bubble group. To do this, he can click on any nodes in this keyword bubble group to activate the convex boundary and then drag the word w outside the boundary. Doing so will reduce the similarity between the word and other words in this keyword's group by 50%. The change of these similarities will trigger an update of the hierarchical clustering and the cluster force layout. To hide the boundary, the analyst can click any of the nodes in the bubble.

Once an analyst is satisfied with a concept cluster, he can choose the words in the cluster to be used for semantic similarity matching for the event records. The analyst can select the cluster by holding the mouse on any of the nodes in the cluster, and the selected cluster will be highlighted by an orange border. The analyst can then drag this cluster into the container in the bottom right corner to select the words, and those selected words will remain in the container and be used for semantic similarity matching. Alternatively, the analyst can also drag any individual words to the container. Selected words can be removed by dragging them out of the container. An example of these interactions is illustrated in Figure 4 which shows how we can use the cluster force layout to eventually select a subset of words related to "food". In Figure 4, the analyst is creating a concept map for the keyword "food". The bubbles contain semantically related words as captured using the WordNet similarity. First, the analyst inspects the different clusters. In step 1, the analyst wants to refine the cluster containing "meal", "food", and "water". The word "treat" is moved into the other cluster within the keyword bubble group and the clustering updates. Due to the semantic similarity score between "treat" and "centre", "centre" is also moved with "treat". In step 2, the analyst wants to remove "media" entirely from the analysis and drags it outside the convex boundary of the "food" clustering. Then, in step 3, "board" is moved away from "food" but positioned next to "cut" as the analyst feels those may be conceptually similar. After having a cluster with words "food, water, meal", the analyst notices that the word "beef" might also relate to the concept of food, so in step 4 "beef" is dragged to the cluster of food. This turns into a state where the word "stick" is also clustered together with food and the analyst drags it away as shown in step 5. Finally, step 6, the analyst chooses to use the words in the highlighted cluster for the semantic matching.

4.2.2 *Word Weights Update*

In addition to semantically interacting with the cluster force layout to refine the concept words, users can also interact with the Event List showing the text details from elements in the secondary dataset (Figure 1 bottom right). Each record in the Event List view contains the date, similarity score, and the text for an event. The selected semantically similar words are highlighted using the same color as the related media keywords. For example, the Event List in Figure 1 shows events queried by topic keywords "food"

and "agriculture", which match to qualitative colors "pink" and "blue". The semantically similar words (e.g. "meal" is related to "food" and "fishing" is related to "agriculture") are highlighted by the corresponding color in the Event List. When browsing related events, the analyst can mark an event as relevant or irrelevant by clicking directly on the text. When marked as irrelevant, the word weight δ in Equation 2 will decrease by 0.25 (until reaching 0), and δ will increase by 0.25 (until reaching 1.0) if marked as relevant. Through this interaction, the scoring measure of the events will update while the word similarity cluster does not change. For example, in Figure 5, we filter a list of events based on words similar to food, and we notice that the word "beef" in the first event does not mean the meat for eating but means "to strengthen", and we mark this event as irrelevant. The weight of the word "beef" then decreased and the rank of this event drops, as shown in the right side list.

4.3 Limitations

Given the computational demands in the framework, several of the features are limited based on the data size. The similarity calculation time is proportional to the data size. On average, the calculation takes about 20 seconds to process 10,000 pairs of words on a computer with an Intel i7 2.67GHz 8 core CPU and 20GB of RAM. This step is done as preprocessing to enable interactive exploration. In our use cases, we have three datasets with 5421 (Climate Change Media), 1820 (Social Unrest Media), and 113297 (ACLED) documents. We calculate the similarity of 102,201,615 pairs of words. Pre-processing takes approximately 40 hours for all the documents. The other major performance bottleneck is querying for similar words. Similarity searches took approximately 28 seconds to return an uncached query.

Along with the computational complexities, the visualization also has several limiting factors. Specifically, the force directed layout can only display a limited number of words. In our experiments, a screen size of 900×450 was found to be sufficient for displaying about 115 words at the initial zoom level. This represents approximately 4 to 5 clusterings of words.

5 Causality Detection

While the semantic similarity analysis and interactions enable the analyst to filter and link events between two datasets, these methods do not provide any indication of whether or not the events identified in the secondary dataset seem to be driving the media topics under analysis. As such, our framework leverages statistical causality models to provide a quantitative indicator of significance under the hypothesis that the current event series is driving the media discourse. We formulate the input to the causality model as two time series. Time series $Y(T) = \{y_1, y_2, \ldots, y_t, \ldots, y_T\}$ represents the volume per time step of documents in our media data classified into the topic of interest. Time series $X(T) = \{x_1, x_2, \ldots, x_t, \ldots, x_T\}$ represents the volume per time step of related events identified during the semantic matching procedure. Then, we can test for causality between these two time series, where Y is the effect and X is the cause. Here, it is important to note that there may be other relevant factors in a larger universe to X and Y which cannot be modeled practically, so spurious causalities may also be identified; however, these measures are still able to provide insight and help in the hypothesis generation, testing, and exploration process. While no statistical technique can provide a definitive test for causality, a causality test is able to provide explanations of effects as the results of potential causes and suggest whether a change in the media stream might be correlated to some local events [41]. Another issue in this test is that the causal effect can be bidirectional, which means X can cause Y and Y may also cause X. In such a situation, a feedback mechanism should appear. In this application, we focus only on one directional causality, exploring the question of X causes Y. The following assumptions on our dataset are also made:

1) The cause shall appear before of the effect;
2) The information in a larger universe not coded in $U = \{X, Y\}$ will be irrelevant, and;
3) Both X and Y are stationary series, which means their means neither change over time nor follow any trends. This should be reasonable for natural events, otherwise they shall first be transformed to stationary processes.

We apply the **Granger causality test** [42]. In a simple causal model (no instantaneous causality and no feedback mechanism), causality is tested by fitting the following two linear regression models and testing if the prediction variance is statistically significantly improved in the second model.

$$y_t = a_0 + \sum_{i=1}^{m} a_i y_{t-i} + \varepsilon_t$$

$$y_t = a_0 + \sum_{i=1}^{m} a_i y_{t-i} + \sum_{i=1}^{m} a_i x_{t-i} + \varepsilon_t$$

Here, ε_t is an uncorrelated noise series, i.e, $E[\varepsilon_t \varepsilon_s] = 0, s \neq t$, and m can be any integer in $[1, T]$.

Let $\sigma^2(A|B)$ be the variance of $\varepsilon_t(A|B)$ which is the error series in the prediction model that series A is the response and series B is the predictor. In our test, given a value of m, we say that X Granger-causes Y at lag m is statistically significant if

$$\sigma^2(Y|Y - Y(t-m), X - X(t-m)) < \sigma^2(Y|Y - Y(t-m)). \quad (3)$$

The F-test is used here to test the significance of the increment of the explanatory power of adding X by comparing the overall fit of the model using only Y and then by using both Y and X. A corresponding *p-value* is also used to show the significance level, where the null hypothesis is that the variance has not decreased by adding X. An R^2 value is used to indicate the performance of the second model by showing how much variance can be fit using both X and Y. The causality test treats the two timeseries data as two arrays of data, the length of time gaps between each data point depends on the granularity of the timeseries. Currently in our framework, the granularity of our data is in days, but the causality test will also work for hourly or monthly data. Since the question of "true causality" requires field testing and controlled experiments, the applied statistical method should only be considered as "predictive" causality which tests whether one time series is useful in forecasting another [43]. However, this is useful as an indicator for trends and drivers and can help an analyst in exploring hypotheses. This serves as a basis for choosing which points in the Timeline may need further annotation.

6 Annotation

Annotations have been used in visualizations to highlight interesting data points, provide context, and display detected events [7], [26], [27], [44]. Our framework allows analysts to annotate media articles and related events, and provides causality modeling indicators in the Timeline for correlation discovery and externalization.

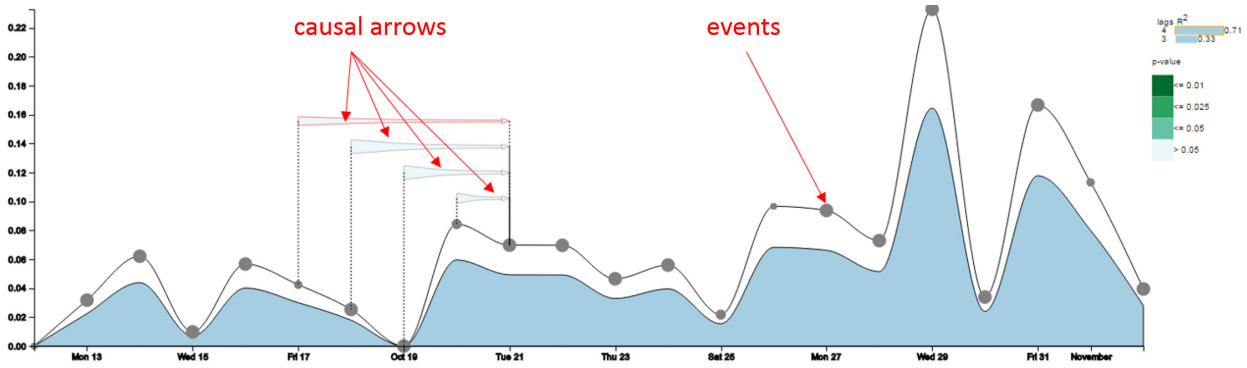


Fig. 6. Initial Causality results without filtering events. The causal arrows are colored based on their significance, and the current result shows no significant causality between the current media stream and events, and the best fit model is displayed with $R^2 \approx 0.6$ and $0.1 < p\text{-}value \leq 0.5$ at a lag of 4. The height of the area curve represents the amount of variance explained by the model.

6.1 Timeline with Events and Causalities

One of the main views in our framework is the Timeline, Figure 6, which starts from a single line denoting the volume of the media stream to be augmented gradually by adding relevant events, causalities, and descriptive text annotations. The solid black line indicates the trend of the percentage volume of the selected media topic. The dots in gray on this line represents related events that happened on a given day, and the size of a gray dot represents the number of events that happened on that day. The size and the amount of the dots will update after interactions with the Topic Keyword view, the Cluster Force Layout, and the Event List view.

When the analyst is satisfied that the relevant events have been semantically linked, they can click on the "Causality Test" button (see Figure 1) to run the causality test on the retrieved events and the media topic. This causality test returns the statistics for all models with possible lag smaller than 10. For each model, the *p-value* and R^2 are displayed for evaluation. These causality models with different lags are indicated at the top right corner of the Timeline by a legend consisting of several bars, one for each model. To the left of each bar, the lag of the model is shown, and the length of the bar indicates the R^2 value. When one model is selected, a stream area, referred to as the explanatory area, is shown below the topic volume line to represent the R^2 which denotes how much variance is explained by the model. For example, when a model's $R^2 = 0.6$ the explanatory area will cover 60% of the area under the media stream line, Figure 6. The analyst can move the mouse along the timeline to browse the model for each date, and the identified lags and events will be shown on top of the timeline illustrated with arrows (causal arrows in Figure 6) connecting the possible driving events to the effect date of media articles. The color gradient, in a sequential color scheme, shows the respective significance levels referring to the *p-value* color legend. For example, in the Figure 6 the color of the causal arrows are light green which means the causality is only significant at a level lower than 90%.

For a model with lag m, the media stream is fitted with a linear regression model $y_t = a_0 + \sum_{i=1}^{m} a_i y_{t-i} + \sum_{j=1}^{m} b_j x_{t-j}$ and only the past m time steps have been used as predictors. The start point of each arrow matches the time of the event and the end point (the point with an arrow) matches the time t where the volume of news articles are being predicted. The width of an arrow's starting point corresponds to the amount of events happening that day. For the purpose of perception and aesthetics, the arrows are ordered by their length either bottom up or top down, based on the position of the gray dots which represent the potential causal events. Our timeline view currently supports only the results from one causality test on one pair of series, and future work will explore methods to overcome this limitation.

6.2 Annotating with Text Information

When the causality test result is positive, it shows the possibility that event series is the cause of the media series, however, to make such hypothesis it is necessary to look into the content of the media articles and the detail information of those events. We used a bipartite view to allow the analysts to navigate through the events and media articles, and then the analysts can pick interesting events and media articles to be annotated on the timeline.

6.2.1 *Bipartite View*

The Bipartite View (Figure 7) displays the connections between media articles and event records linked by the causal arrow. The bipartite view updates while the analysts anchor on a date to investigate (by moving the mouse on the timeline while holding the left key). Initially the bipartite view displays the events and media articles on the same day. Once the analyst selects a causal arrow by clicking it, the bipartite view displays the events and media articles linked by the arrow. The right side of the Bipartite View lists all media articles and the left side lists relevant events. Both media articles and events are colored according to the semantically matched keywords. For media articles, the semantically matched keywords are the selected media keywords that the articles contain. For events, the matched keywords are the selected media keywords that have some of their semantically related words contained in the event notes. When the article or event record has one matched keyword, it will be represented as a rectangle in the color corresponding to the keyword's legend. If more than one keyword has been matched, this rectangle will use a blended color from all the matched keywords' colors. Both media articles and events are grouped according to their color, i.e. the semantically matched keywords. The edges connect the groups by their keyword co-occurrence. The edges are colored by the color of matched keywords that both of the connected groups contain. For example, Figure 7 shows the bipartite view of topic "Election" and keywords "inec" (Independent National Electoral Commission), "medical", "electoral", "health", and "hospital". Mousing over any rectangles on the bipartite view will bring out the tooltip which

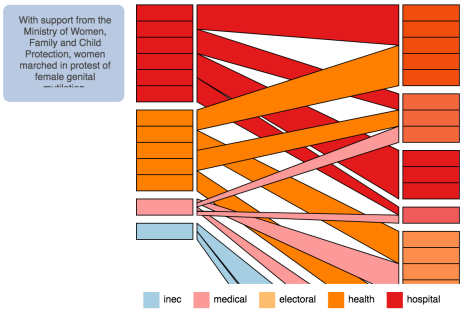


Fig. 7. Example of the Bipartite view. This view shows the events (left side) and the media articles (right side) indicated by the causality lag under analysis and connected by semantically matched keywords.

shows the beginning of the event note/article. For displaying a long list of matches, the Bipartite View is scrollable so that the user can view the details when they do not fit in the area.

6.2.2 *Detail List*

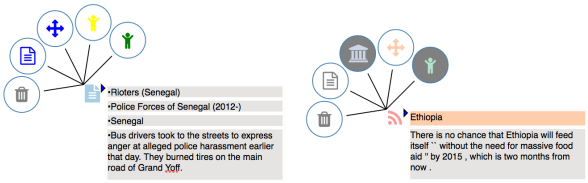
The detailed information that the rectangles in the Bipartite View contain can be explored in the linked detail view on the sides of the Bipartite view. The detail of event records are listed on the left and the detail of media articles are listed on the right (Figure 9, bottom). Each event record has up to two actors, a location, and a note describing the event. The border color of each entry in the lists matches the color of its matched keywords. Different to the color design of the Bipartite View, if there are more than one matched keyword for the entry, there will be multiple borders, each colored by one keyword's color. The order of entries on both lists reflects the order of the rectangles on the bipartite view.

6.2.3 *Entity Annotations*

While exploring the detailed information through the bipartite view, the analysts can externalize their findings to the timeline by double clicking on the media and event rectangles. The selected event will be annotated as a record icon using its keyword's color. The selected media text will be annotated as a feed icon using the color of its keyword. To expand the detailed information of the annotated events and media articles, analysts can click on an icon and the actors, locations, and text information will be available as expanded nodes (Figure 8). From these nodes, the analyst can choose which event attribute to show on the Timeline. The annotation can help analysts to immediately interact with data so that one can flag events that can constitute changes in the underlying equilibrium of these processes.

6.3 Limitations

The proposed media timeline annotation approach has limited scalability due to the screen space and the dense content and causal relationships being displayed. In practice, a 1920×1080 resolution screen would be able to display annotations of 10 to 20 events and media articles without perceptual difficulties. Each application also depends on the actual data and users can easily expand/collapse icons to optimize the space. Furthermore, the color design in the Bipartite View uses color blending to represent multiple matching keywords, which could potentially cause perceptual issues.



(a) Event Annotation (b) Media Article Annotation

Fig. 8. The annotation glyph for event and media articles. The analyst can annotate entities by clicking on the expanded nodes.

7 Case Studies

In this section, we demonstrate our framework through an analysis using a climate change media collection and a social unrest media collection respectively. These datasets will be semantically annotated by the Armed Conflict Event Location Dataset (ACLED) [45] and causal drivers explored. For the case studies, a paired analysis protocol [46] was used in which system features were explained and demonstrated to our partners in political science. Our partners discussed their developing hypotheses and instructed the framework developers driving the system.

7.1 Datasets

ACLED: The ACLED dataset (1997 to present) contains information on the dates and locations of all reported political violence events in over 50 developing countries with a focus on Africa. Each event record contains information on the date, location, event type and actors involved with approximately 6500 events from August to December 2014.

Climate Change Media: The climate change media dataset is composed of RSS feeds from 122 English language news outlets and filtered for relevance by matching against a set of 222 keywords. From August 2014 to December 2014, this collection contains 1245 relevant articles with 9070 sentences which are further coded into one (or none) of 25 framing categories. All articles have been analyzed through entity recognition to extract people, location, and organizations. A more specific description can be found in our previous work [9].

Social Unrest Media: The social unrest media dataset is composed of RSS feeds from 128 English language news outlets collected in March 2015. RSS feeds were scanned hourly and the content of each news article was filtered by a set of 378 social unrest keywords. The LDA topic modeling algorithm [11] was run on these articles and 50 topics were extracted. The following 7 topics were selected based on their relatedness to the ACLED dataset: Election, Economy, Education, Conflicts, Agriculture, Justice, and Energy. All articles have been processed using entity recognition for annotation.

7.2 Climate-induced Unrest During Drought

The drought in Africa has attracted the attention of researchers who want to analyze potential societal impacts that the draught may have [47]. Specifically, the draught has caused widespread agricultural failures and led to famines and political instability. In this case study, the analyst wanted to explore if the GHoA (Great Horn of Africa) drought in 2014 coincided with instances of political violence. Causal relationships are extremely difficult to test when using observational data, and political scientists debate about the relationships between climate change and political

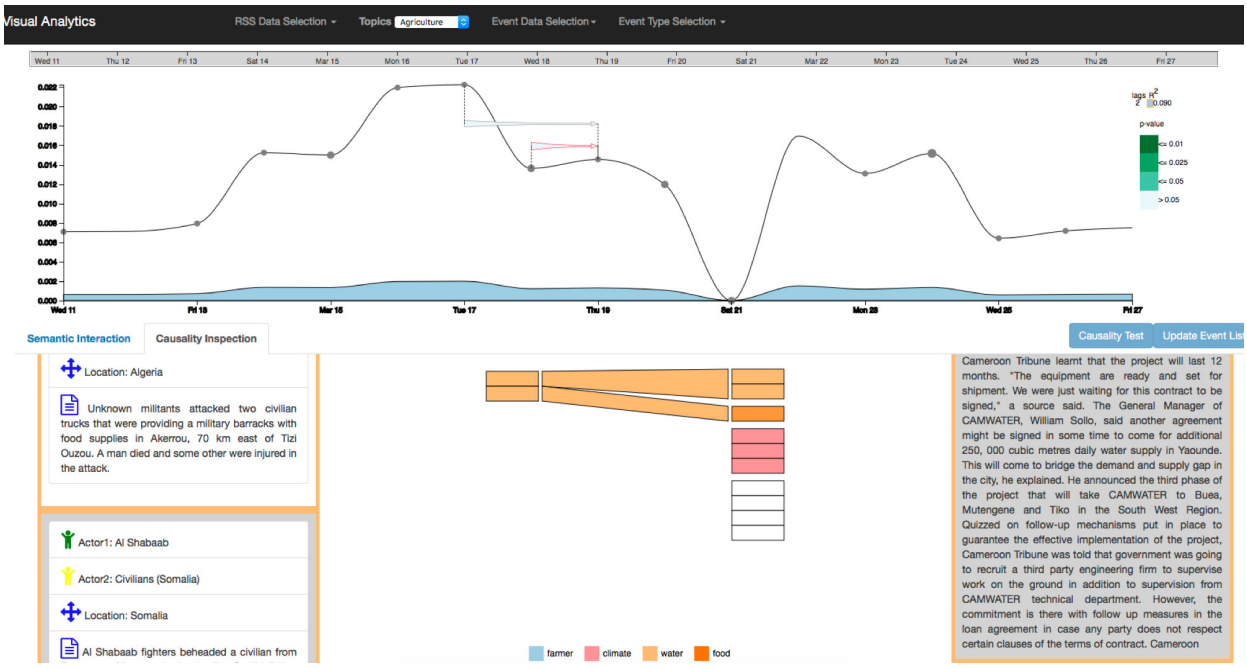


Fig. 9. Investigating the plausibility of climate-induced civilian abuse during the 2014 GHoA Drought. The link between violence events and social unrest RSS have been explored although casualty test shows non-significant result.

violence [48]. Furthermore, the research that examines the role of drought in instances of violent conflict lacks consensus [49], and it remains unclear whether the onset of drought and subsequent observations of violence could be indicative of broader phenomena, such as a governance failure or institutional failure. As such, this case study focuses on the question: Is the 2014 drought in the GHoA linked to reports of social unrest and political violence? To probe this question, the researcher first selected "social unrest RSS", defined the temporal domain to encompass the month of March 2014, and picked agriculture as the topic to explore. Next, events of "violence against civilians" (which ACLED defines as any armed/violent group that attacks civilians") is selected to explore potential drivers between droughts and violence.

To begin, the analyst selected the following words, based on the significance ordering, as most plausibly being related to the March - June 2014 GHoA drought: "water", "food", "farmer", and "climate". These selections are then used to update the relevant events and explore the semantic linkages in the clustering bubble interface displayed below the event timeline. The clustering results reveal that the word "climate", in this semantic mapping, shares words/concepts not related to the concept of agriculture. Instead, words such as "way, order, demand, tension, and control" are found, even after adjusting the similarity threshold to .75. Thus, "climate" is removed and the event list is updated to reflect the change (see Figure 10). After removing "climate", the analyst remains largely satisfied with the clustering for the remaining terms and now creates clustering within the keyword selection container. The causality test is performed and returns the following insignificant model result: lag=2, R2=0.090 (see Figure 9). The insignificant result is not surprising to the analyst since many other factors are also expected to drive the events. However, he also requested to further explore the details of the events and the media posts to identify if there were other keywords or factors he may not have considered. Using the Bipartite View, the analyst briefly evaluated the links between the keywords and the recorded episodes of political violence perpetuated against civilians. His search revealed shared associations between terms related to drought and a recorded event of Al Shabaab beheading a civilian for unstated reasons, (see Figure 9). Based on the exploration, the analyst concludes that the linkage between resource shortages and civilian abuse may be less plausible.

7.3 Food Insecurity and Climate Change Media

Our analyst was also interested in exploring potential drivers of climate change media discussions with respect to ongoing conflict events in Africa. He hypothesized that external drivers, such as riots and protests, may be driving the types of framing being used to discuss climate change. First, the analyst loads the Climate Change media collection and the ACLED dataset. The analyst decides to focus on the food insecurity frame "ProbThreatFood", from October 12th to November 3rd, 2014. Next the analyst chooses to explore "Riots/Protests" from ACLED to annotate the media frame. The analyst first selects keywords "food", "crop", and "agriculture". The analyst adjusts the threshold in the Cluster View and discusses the resulting clusters.

First the analyst chooses several topic keywords related to food insecurity in the climate change media dataset, selecting "crop", "food", and "agriculture". Events are then automatically selected through the semantic keyword processing, and the analyst runs an initial causality test which shows no significant causal correlations. The result of this initial model is shown in Figure 6. The problem is that many events that are marked as semantically similar to these keywords do not match the analyst's meaning of "crop", "food", and "agriculture". Thus, the analyst begins using the Cluster View to group the words into conceptual groups. As such, the analyst explores semantic keywords related to the selection of "crop". However, the analyst finds that events in the secondary dataset matched as semantically similar to "crop" are not embedding

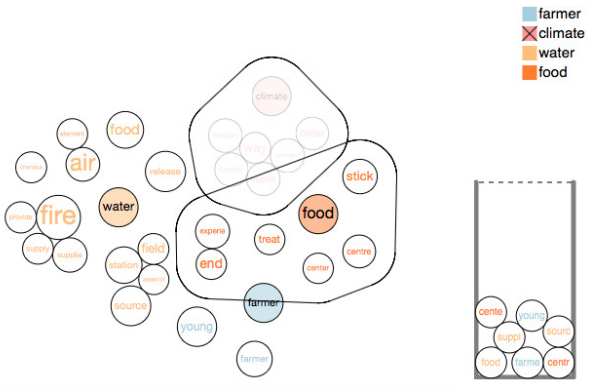


Fig. 10. Semantic word clustering and filtering for the study of climate-induced civilian abuse during the 2014 GHoA drought.

the meaning of "cultivated plants". Instead, the related words, particularly "plant", are referring to industrial plants in the event data, but not cultivated plants. This illustrates the difficulties that a fully automatic approach might face. While it may be reasonable to assume that "plant" as a keyword would represent agricultural concepts, in the ACLED dataset this is not the case. Thus, the analyst removes "crop" from the keyword list so that events semantically linked to this concept of crop are no longer selected. Again, since different domains use slightly different linguistic descriptors, a visual analytics solution can enable analysts to inject domain knowledge and reasoning into the analysis pipeline.

Next, the analyst explores the clusters of semantic keywords linked to "food", as shown in Figure 4, and conceptualizes the meaning of several clusters. Seeing words in the blue cluster, such as "education", "course", and "issue", being clustered together was identified as primarily being a concept of "food for thought". The analyst finds that one cluster contains many "food" related words (e.g. meal, water, food). However, there are still some irrelevant words in this cluster (e.g. treat, centre, board and media). To better select related events to "food", the analyst drags these irrelevant words either into the other cluster or away from both clusters, the result of which is shown in Figure 4 - Step 4, where one cluster now contains "food", "water", "meal" and "beef". Now, the analyst chooses to only semantically link the word "food" to match his conceptual notion.

The analyst then chooses to filter only on certain conceptual cluster groups that he has now created in the cluster view. For example, in the cluster view illustrated in Figure 4(step 6), the cluster with words (food, water, meal, beef) are selected by dragging into the word selection container and used to semantically link the media dataset to events. However, after exploring the event notes, the analyst realizes that "beef" is only used in the context of "beef up", so the analyst removes the word "beef" from the semantic filtering as this meaning does not relate to the current conceptualization of food. The analyst then further filters the events using his agriculture concept group. After filtering events, the best fit causality model at this point is still not significant ($R^2 \approx 0.6$ and $p\text{-}value > 0.1$), Figure 6. From the event list view, the analyst sees that there are many words in the event data that are semantically linked by the media keyword "agriculture" but not related to the meaning in the context of food problems. The analyst again refines the Cluster View to select relevant words for "agriculture" and selects "agriculture" and "fishing" (as Africa has large exports of fish) which is shown in the word selection container in Figure 1. The causality modeling returns three models with lag equals to 2, 3, and 4 respectively. The best fit model (Figure 1) has lag = 4, $R^2 \approx 0.8$ and $p\text{-}value \leq 0.05$.

Switching to the Causality Inspection tab for more details of the relevant media articles and ACLED events, the analyst explores the connections suggested by the causality model with lag = 4 using the Bipartite view to further annotate potential causal relationships. Here, the analyst explores filtered events, annotating nodes with text and exploring lags in the dataset. The final resulting annotated Timeline that was developed is shown in Figure 1. In looking at the peak discussion of food insecurity, we are seeing articles mentioning Food shortages and rising prices have the potential to worsen political, ethnic, class and religious tensions, and "Ordinary people could go hungry if their countries can not produce enough food". Prior to these articles, we are seeing an increased number of riots and protest around unpaid meal allowances and discussions of lack of food. Here, discussions of riots and protests related to food are marked on the Timeline and can provide insight into what external events may cause shifts in media framing, for example, if there are more protests related to food issues, does climate change media pick up on this and begin framing climate change as a food insecurity issue?

7.4 User Feedback

We have collected user feedback from three groups of experts, a political scientist, the collaborators in communication, and our outside partners. The political scientist expert group consists of a co-author on the paper who was interested in developing a framework to enhance his data exploration process. This expert provided detailed feedback about tool usage and details in the case study are from an ongoing analysis. The second group consists of faculty and doctoral students in the Communications Department at Arizona State University. Their interest is in the impacts of framing in social media, which requires the ability to map different lexicons among topics. This group provided detailed feedback on keyword selection and analysis and used the framework to perform social media analyses. The final group consists of experts whose feedback was informally solicited during demo sessions of the tool. Here, experts were walked through the tool usage and case study and discussions were held regarding how such a tool would be useful for their domain applications.

Through using the framework themselves, our political science and communication experts liked the framework overall, indicating that the major advantages were that this "Allows me to read in a way that I can't do manually. This tool allows me to explore my beliefs about the data and events and record details to the point of theory construction." Our experts liked the clustering force directed layout and consider it to be a more intuitive display. They felt it increased their understanding and willingness to engage with the program. The political science analyst also felt that the ability to easily hone the selection of the terms of interest by removing and adding words to more relevant clusters is a striking feature, and the immediate updating of the clustering algorithm within the user-interface after winnowing the clustering to domain-relevant terms is very useful and appropriate. The analyst considers the embedded granger causality testing to be simple to use, and the causality testing helped lead to formalization of actual hypotheses and provides some level of base knowledge about what concepts of interest are most relevant. The visualization of how much variance

the model explained being represented as a filled area under the curve was noted as being highly intuitive.

Along with the above detailed feedback, we have also demonstrated our framework to industrial partners. Feedback from these demonstrations indicated that users like the interface and the approach of semantically linking events to media topics. They think the visual representation of the clustered keywords are quite intuitive and the causality test is easy to understand. They also pointed out some limitations of our work. First, our framework needs the text dataset to be preprocessed, including categorization (e.g. labeling by domain experts or topic modeling) and word similarity calculation. This limitation currently prevents this framework from handling streaming text data. Second, our demonstrations only showed how to link between media posts and conflict events. However, these demonstrations were given to people in vastly different domains who indicated a need to analyze proprietary data sources which may require modifications to the visual design to support domain specific annotations.

Furthermore, many media sources of interest also contain video and images, as such, extracting relevant content becomes difficult. Also, the scalability of the system is a critical issue. In the datasets used by our collaborators for the case studies, scalability was not an issue as some data curating had already been performed and pre-processing of data could take place prior to interactive analysis. However, a key task of causality analysis is to build predictive models. During our demonstrations, requests for real-time model building and updates using streaming data was discussed. Currently, the system is limited by the pre-processing requirements; however, the workflow proposed by the framework is robust to support the causality modeling task but will require the addition of a streaming data processing step.

8 Conclusions and Future Work

This paper presents a framework for semantically annotating media topic discourse through linked datasets. To accomplish this, we have designed a cluster force layout that can facilitate the development of a concept map of keywords to be used for semantically filtering linked events. Relationships between these events and media trends can be analyzed using causality modeling, and model results are interactively displayed on the Timeline. Though the causality modeling cannot guarantee a true cause-effect relationship, results obtained from such models are able to help analysts in their knowledge discovery and hypothesis generation. Analysts may explore suggested connections between media articles and linked events, and articles and events that are linked by multiple concepts are further visualized in the Bipartite View. From the Bipartite View, analysts can annotate events of interest on the Timeline to help inform their given hypotheses.

While our examples focused on media and conflicts in Africa, the tools developed are applicable to a variety of domains and data. However, there are several limitations to this framework. First, the semantic match is constrained to a keyword based approach, i.e., the analyst must choose an initial set of keywords from the document as a starting point. This can limit the matching as other words between the corpuses may serve as more appropriate semantic bridges. Future work will explore new metrics for organizing the keyword list and providing automatic initial suggestions. Second, although we have shown that the knowledge-based semantic similarity methods can be leveraged to connect two textual datasets, other information retrieval metrics could also be explored and compared. Future work will focus on developing an ensemble of methods for stronger semantic matching utilizing links between named entities in the document. Third, the framework has limitations of its scalability and capability of generalizing to streaming data and more complexed data format. The scalability issue exists in both the cluster view and the Timeline as discussed. More advanced techniques in database and similarity calculations are needed to generalize this framework for streaming data and broader datasets.

Acknowledgments

This work was supported by the U.S. Department of Homeland Security's VACCINE Center, Award 2009-ST-061-CI0001 and the National Science Foundation, Grant Nos. 1350573 and 1639227.

References

[1] H. Lee, J. Kihm, J. Choo, J. Stasko, and H. Park, "ivisclustering: An interactive visual document clustering via topic modeling," in *Computer Graphics Forum*, vol. 31, 2012, pp. 1155–1164.

[2] S. Liu, M. X. Zhou, S. Pan, W. Qian, W. Cai, and X. Lian, "Interactive, topic-based visual text summarization and analysis," in *Proceedings of the 18th ACM Conference on Information and Knowledge Management*. ACM, 2009, pp. 543–552.

[3] W. Dou, X. Wang, D. Skau, W. Ribarsky, and M. X. Zhou, "Leadline: Interactive visual analysis of text data through event identification and exploration," in *IEEE Conference on Visual Analytics Science and Technology*, 2012, pp. 93–102.

[4] W. Cui, S. Liu, Z. Wu, and H. Wei, "How hierarchical topics evolve in large text corpora," *IEEE Transactions on Visualization and Computer Graphics*, vol. 20, pp. 2281–2290, 2014.

[5] S. Havre, E. Hetzler, P. Whitney, and L. Nowell, "Themeriver: Visualizing thematic changes in large document collections," *IEEE Transactions on Visualization and Computer Graphics*, vol. 8, pp. 9–20, 2002.

[6] F. Wanner, W. Jentner, T. Schreck, A. Stoffel, L. Sharalieva, and D. A. Keim, "Integrated visual analysis of patterns in time series and text data-workflow and application to financial data analysis," *Information Visualization*, vol. 15, pp. 75–90, 2016.

[7] J. Hullman, N. Diakopoulos, and E. Adar, "Contextifier: Automatic generation of annotated stock visualizations," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2013, pp. 2707–2716.

[8] S. Liu, Y. Chen, H. Wei, J. Yang, K. Zhou, and S. M. Drucker, "Exploring topical lead-lag across corpora," *IEEE Transactions on Knowledge and Data Engineering*, vol. 27, no. 1, pp. 115–129, Jan 2015.

[9] Y. Lu, M. Steptoe, S. Burke, H. Wang, J.-Y. Tsai, H. Davulcu, D. Montgomery, S. R. Corman, and R. Maciejewski, "Exploring evolving media discourse through event cueing," *IEEE Transactions on Visualization and Computer Graphics*, vol. 22, pp. 220–229, Jan 2016.

[10] J. A. Wise, J. J. Thomas, K. Pennock, D. Lantrip, M. Pottier, A. Schur, and V. Crow, "Visualizing the non-visual: spatial analysis and interaction with information from text documents," in *Proceedings of Information Visualization*. IEEE, 1995, pp. 51–58.

[11] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *The Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[12] G. Sun, Y. Wu, S. Liu, T. Peng, J. Zhu, and R. Liang, "Evoriver: Visual analysis of topic coopetition on social media," *IEEE Transactions on Visualization and Computer Graphics*, vol. 20, pp. 1753–1762, 2014.

[13] P. Xu, Y. Wu, E. Wei, T. Q. Peng, S. Liu, J. J. H. Zhu, and H. Qu, "Visual analysis of topic competition on social media," *IEEE Transactions on Visualization and Computer Graphics*, vol. 19, pp. 2012–2021, 2013.

[14] F. Beck, S. Koch, and D. Weiskopf, "Visual analysis and dissemination of scientific literature collections with survis," *IEEE Transactions on Visualization and Computer Graphics*, vol. 22, pp. 180–189, 2016.

[15] F. Heimerl, Q. Han, S. Koch, and T. Ertl, "Citerivers: Visual analytics of citation patterns," *IEEE Transactions on Visualization and Computer Graphics*, vol. 22, pp. 190–199, 2016.

[16] I. Cho, W. Dou, D. X. Wang, E. Sauda, and W. Ribarsky, "Vairoma: A visual analytics system for making sense of places, times, and events in roman history," *IEEE Transactions on Visualization and Computer Graphics*, vol. 22, pp. 210–219, 2016.

[17] W. Ribarsky, D. X. Wang, and W. Dou, "Social media analytics for competitive advantage," *Computers & Graphics*, vol. 38, pp. 328–331, 2014.

[18] D. Fisher, A. Hoff, G. Robertson, and M. Hurst, "Narratives: A visualization to track narrative events as they develop," in *IEEE Symposium on Visual Analytics Science and Technology*, 2008, pp. 115–122.

[19] K. Wongsuphasawat and J. Lin, "Using visualizations to monitor changes and harvest insights from a global-scale logging infrastructure at twitter," in *IEEE Conference on Visual Analytics Science and Technology*, 2014, pp. 113–122.

[20] N. Diakopoulos, M. Naaman, and F. Kivran-Swaine, "Diamonds in the rough: Social media visual analytics for journalistic inquiry," in *IEEE Symposium on Visual Analytics Science and Technology*, 2010, pp. 115–122.

[21] A. Scharl, A. Hubmann-Haidvogel, A. Weichselbraun, G. Wohlgenannt, H.-P. Lang, and M. Sabou, "Extraction and interactive exploration of knowledge from aggregated news and social media content," in *Proceedings of the 4th ACM SIGCHI Symposium on Engineering Interactive Computing Systems*. ACM, 2012, pp. 163–168.

[22] A. Scharl, A. Hubmann-Haidvogel, A. Jones, D. Fischl, R. Kamolov, A. Weichselbraun, and W. Rafelsberger, "Analyzing the public discourse on works of fiction–detection and visualization of emotion in online coverage about hbos game of thrones," *Information Processing & Management*, vol. 52, pp. 129–138, 2016.

[23] Q. Zheng, K. Booth, and J. McGrenere, "Co-authoring with structured annotations," in *Proceedings of the SIGCHI conference on Human Factors in computing systems*. ACM, 2006, pp. 131–140.

[24] L. Hong, E. H. Chi, R. Budiu, P. Pirolli, and L. Nelson, "Spartag. us: A low cost tagging system for foraging of web content," in *Proceedings of the working conference on Advanced visual interfaces*. ACM, 2008, pp. 65–72.

[25] Y. Chen, S. Barlowe, and J. Yang, "Click2annotate: Automated insight externalization with rich semantics," in *IEEE Symposium on Visual Analytics Science and Technology*. IEEE, 2010, pp. 155–162.

[26] T. Gao, J. R. Hullman, E. Adar, B. Hecht, and N. Diakopoulos, "Newsviews: An automated pipeline for creating custom geovisualizations for news," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2014, pp. 3005–3014.

[27] J. Fulda, M. Brehmer, and T. Munzner, "Timelinecurator: Interactive authoring of visual timelines from unstructured text," *IEEE Transactions on Visualization and Computer Graphics*, vol. 22, pp. 300–309, 2016.

[28] N. Elmqvist, A. V. Moere, H.-C. Jetter, D. Cernea, H. Reiterer, and T. Jankun-Kelly, "Fluid interaction for information visualization," *Information Visualization*, 2011.

[29] R. Baeza-Yates and B. Ribeiro-Neto, *Modern information retrieval*. ACM press New York, 1999, vol. 463.

[30] D. Metzler, S. Dumais, and C. Meek, *Similarity measures for short segments of text*. Springer, 2007.

[31] R. Mihalcea, C. Corley, and C. Strapparava, "Corpus-based and knowledge-based measures of text semantic similarity," in *AAAI*, vol. 6, 2006, pp. 775–780.

[32] R. Navigli, "Word sense disambiguation: A survey," *ACM Computing Surveys*, vol. 41, p. 10, 2009.

[33] Z. Wu and M. Palmer, "Verbs semantics and lexical selection," in *Proceedings of the annual meeting on Association for Computational Linguistics*. Association for Computational Linguistics, 1994, pp. 133–138.

[34] G. A. Miller, "Wordnet: a lexical database for english," *Communications of the ACM*, vol. 38, pp. 39–41, 1995.

[35] C. D. Manning, M. Surdeanu, J. Bauer, J. R. Finkel, S. Bethard, and D. McClosky, "The Stanford CoreNLP Natural Language Processing Toolkit." in *ACL (System Demonstrations)*, 2014, pp. 55–60.

[36] S. C. Leman, L. House, D. Maiti, A. Endert, and C. North, "Visual to parametric interaction (v2pi)," *PloS one*, vol. 8, p. e50474, 2013.

[37] T. Sørensen, "A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on danish commons," *Biologiske Skrifter*, vol. 5, pp. 1–34, 1948.

[38] A. Endert, L. Bradel, and C. North, "Beyond control panels: Direct manipulation for visual analytics," *IEEE Computer Graphics and Applications*, vol. 33, pp. 6–13, 2013.

[39] A. Endert, P. Fiaux, and C. North, "Semantic interaction for visual text analytics," in *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 2012, pp. 473–482.

[40] A. Endert, S. Fox, D. Maiti, and C. North, "The semantics of clustering: analysis of user-generated spatializations of text documents," in *Proceedings of the International Working Conference on Advanced Visual Interfaces*. ACM, 2012, pp. 555–562.

[41] H. B. Asher, *Causal modeling*. Sage, 1983.

[42] C. W. Granger, "Investigating causal relations by econometric models and cross-spectral methods," *Econometrica: Journal of the Econometric Society*, pp. 424–438, 1969.

[43] F. X. Diebold, *Elements of forecasting*. South-Western College Publ., 1998.

[44] J. Zhao, M. Glueck, S. Breslav, F. Chevalier, and A. Khan, "Annotation graphs: A graph-based visualization for meta-analysis of data based on user-authored annotations," *IEEE Transactions on Visualization and Computer Graphics*, vol. PP, pp. 1–1, 2016.

[45] "Armed conflict location & event data project," http://http://www.acleddata.com/, accessed: 2016-03-17.

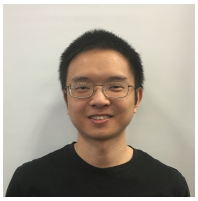
[46] R. Arias-Hernandez, L. T. Kaastra, T. M. Green, and B. Fisher, "Pair analytics: Capturing reasoning processes in collaborative visual analytics," in *Hawaii International Conference on System Sciences*. IEEE, 2011, pp. 1–10.

[47] S. C. Herring, A. Hoell, M. P. Hoerling, J. P. Kossin, C. J. Schreck, and P. A. Stott, "Explaining extreme events of 2015 from a climate perspective," *Bulletin of the American Meteorological Society*, vol. 97, p. Sii, 2016.

[48] H. Buhaug, "Climate change and conflict: Taking stock," *Peace Economics, Peace Science and Public Policy*, vol. 22, pp. 331–338, 2016.

[49] O. M. Theisen, H. Holtermann, and H. Buhaug, "Climate wars? assessing the claim that drought breeds conflict," *MIT Press*, 2011.

Yafeng Lu is a Postdoctoral Research Assistant in the School of Computing, Informatics and Decision Systems Engineering at Arizona State University. Her current research interests are in social media and predictive visual analytics.



Hong Wang is a Ph.D student in the School of Computing, Informatics and Decision Systems Engineering Arizona State University. His research interests include Data Visualization and Human Computer Interaction.

Steven Landis is an Assistant Professor of Political Science at the University of Nevada, Las Vegas. His research interests are focused on the intersections of environmental security, economic development, and political violence.

Ross Meciejewski is an Associate Professor of Computer Science at Arizona State University. His primary research interests are in the areas of geographical visualization, social media mining, and predictive analytics.