Explainable AI Components for Narrative Map Extraction

Brian Keith[1,2,*], Fausto German[2], Eric Krokos[3], Sarah Joseph[3] and Chris North[2]

[1]*Universidad Católica del Norte, Av. Angamos 0610, Antofagasta, 1270709, Chile*
[2]*Virginia Tech, 620 Drillfield Drive, Blacksburg, VA, 24061, USA*
[3]*U.S. Government, Washington, D.C. 20500, USA*

Abstract

As narrative extraction systems grow in complexity, establishing user trust through interpretable and explainable outputs becomes increasingly critical. This paper presents an evaluation of an Explainable Artificial Intelligence (XAI) system for narrative map extraction that provides meaningful explanations across multiple levels of abstraction. Our system integrates explanations based on topical clusters for low-level document relationships, connection explanations for event relationships, and high-level structure explanations for overall narrative patterns. In particular, we evaluate the XAI system through a user study involving 10 participants that examined narratives from the 2021 Cuban protests. The analysis of results demonstrates that participants using the explanations made the users trust in the system's decisions, with connection explanations and important event detection proving particularly effective at building user confidence. Survey responses indicate that the multi-level explanation approach helped users develop appropriate trust in the system's narrative extraction capabilities. This work advances the state-of-the-art in explainable narrative extraction while providing practical insights for developing reliable narrative extraction systems that support effective human-AI collaboration.

Keywords

Explainable AI, Narrative Extraction, Narrative Visualization, Sensemaking, Text Analysis

1. Introduction

Understanding and extracting narratives from large collections of text documents presents significant challenges in natural language processing and visual analytics [1]. As narrative extraction methods become more sophisticated, particularly with the rise of complex Artificial Intelligence (AI) models, there is an increasing need to make these processes transparent and interpretable for users [2, 3]. This is especially crucial in domains such as journalism, intelligence analysis, and digital humanities, where analysts need not only to identify narratives but also to understand how and why specific narrative structures were extracted [4].

Narrative maps—graph-based representations that capture the connections between events in a story—have emerged as a powerful tool for narrative sensemaking [5]. Throughout this work, we consider events—the basic unit of narratives—to be represented by single documents, following a document-based representation of news narratives [1], under the assumption that a single news article usually contains a single main event [6].

Narrative maps are structures that represent events as nodes and their relationships as edges [5], allowing analysts to explore how different parts of a narrative connect and evolve over time. However, the extraction of these maps often relies on complex pipelines involving multiple AI models [4], from embedding generation to clustering and graph optimization. This complexity creates a "black box" effect [7], where users may not understand why certain events are connected or how the narrative structure was determined.

At the fundamental level, narrative extraction involves both low-level text processing and high-level structure generation, requiring explanations at different granularities [4]. The connections

In: R. Campos, A. Jorge, A. Jatowt, S. Bhatia, M. Litvak (eds.): Proceedings of the Text2Story'25 Workshop, Lucca (Italy), 10-April-2025

[*] Corresponding author.

brian.keith@ucn.cl (B. Keith)
https://briankeithn.github.io/ (B. Keith)
0000-0001-5734-8962 (B. Keith)

between events can be based on various factors [8] including temporal sequences, causal relationships, and thematic similarities, making it difficult to explain why specific relationships were identified. Furthermore, narrative extraction typically combines multiple AI models [1], necessitating explanations that bridge different types of processing and help users understand how these components work together to produce the final narrative structure.

In this paper, we present an Explainable AI (XAI) system specifically designed for the task of narrative map extraction. Our approach provides explanations at multiple levels of the extraction process [4] through three main components. First, we leverage topical clusters [9, 10] to generate keyword-based explanation for understanding the low-level document space. Second, we create a connection explanation framework that clarifies why events are linked in the narrative structure using explanations based on SHAP (Shapley Additive Explanations) values [11] and information about shared topics and entities between events. Third, we implement a high-level explanation system that provides descriptive names [12] for storylines and identifies important events.

Our results from the user study demonstrate that providing explanations at multiple levels helps analysts develop appropriate *trust*—understood in the broad sense of reliability, predictability, and efficiency [13]—in extracted narrative structures, leading to more effective human-AI collaboration. This work addresses a critical gap in narrative extraction research by making these complex processes more transparent and interpretable to end users. Our system and data used to extract narratives is available in a public GitHub repository[1].

The remainder of this paper is organized as follows. Section 2 reviews related work. Section 3 presents our methodology, including the XAI components and the user study. Section 4 reports the results of the user study. Section 5 presents the discussion. Section 6 concludes with future directions.

2. Related Work

Research in explainable artificial intelligence for text analysis has emerged as a critical area as natural language processing systems become more complex. Adadi and Berrada [3] provide a comprehensive survey of XAI approaches, categorizing them into model-agnostic and model-dependent techniques. Model-agnostic approaches can be applied to any machine learning model without consideration of internal structure, while model-dependent techniques are tailored to specific architectures. In text analysis, common XAI methods include simplification-based explanations [14], relevance-based explanations [11], and visual explanations [15].

The integration of XAI with visual analytics systems has received particular attention in recent research. Hohman et al. [16] examine how visualization techniques can help reveal the inner workings of deep learning models for text processing. This work demonstrates that visual analytics can bridge the gap between complex AI models and human understanding. Building on this foundation, recent work by Vivacqua et al. [17] specifically addresses XAI in the context of visualizations and sensemaking, showing how transparency in AI models can foster trust between humans and automated systems [18].

In the domain of narrative understanding, XAI faces unique challenges due to the temporal and causal structure inherent in narratives. Abbott [19] establishes that narratives have underlying temporal and causal structures that distinguish them from other forms of text. This structural complexity creates additional requirements for explanation systems. Keith and Mitra [5] introduce the concept of narrative maps as a representation for computational narrative extraction, demonstrating the need for explanations that can address both local event relationships and global narrative structure.

The explanation of narrative structures presents distinct challenges compared to general text analysis. Narrative extraction often relies on event-based models [1], which must capture both the individual events and their interconnections. Traditional XAI approaches for text classification or similarity measurement must be adapted to account for these narrative-specific requirements. Recent work by Keith et al. [8] establishes design guidelines for narrative maps in sensemaking tasks, highlighting the importance of explanations that align with analysts' cognitive processes.

[1] https://github.com/briankeithn/narrative-maps

While existing research has made progress in explaining individual components of narrative analysis systems, there remains a gap in providing comprehensive explanations that span the entire extraction pipeline. Current approaches typically focus on either low-level text processing or high-level structure analysis, but rarely address both in an integrated manner. Our work builds upon these foundations while addressing the challenges of explaining narrative extraction across multiple levels of abstraction.

Finally, we note that, although explainable AI offers significant benefits for narrative extraction systems, we need to acknowledge potential limitations. XAI methods may introduce information overload when explanations are too technical or detailed, overwhelming users rather than helping to understand the AI model [20]. There is also the risk that explanations function merely as proxies for complex underlying processes, potentially providing only a surface-level understanding that may not fully represent the actual computational mechanisms [21]. Additionally, misalignment between the outputs of the algorithm and the corresponding explanations can occur, leading to decreased trust if the behavior of the system contradicts its explanations [22]. In our approach, we address these concerns by designing explanations at appropriate levels of abstraction, focusing on providing general pointers for understanding the big picture while also supporting specific confirmatory tasks, such as verifying connection validity between events.

3. Methodology

3.1. Problem Definition and Assumptions

As narrative extraction systems grow in complexity, users face increasing difficulty understanding both how these systems work and why specific narrative structures were extracted. This paper addresses the specific problem of providing meaningful explanations across multiple levels of abstraction in narrative map extraction systems to enhance transparency, user trust, and effective human-AI collaboration.

In this context, our approach is built on several key assumptions. First, users must understand both the low-level relationship between documents and the high-level narrative structures to develop the appropriate trust in the system. Second, different types of explanations are required for different aspects of the system, such as topical clusters, connections, and storylines. Third, explanations bridging the gap between computational processes and human cognition should enhance trust and usability. Additionally, explanations should balance detail with comprehensibility, avoiding information overload while providing sufficient insight to support user understanding and decision making.

These assumptions guided our development of our multi-level explanation components that address different aspects of the narrative extraction process while maintaining cognitive accessibility.

3.2. Overview of XAI System for Narrative Maps

Our explainable AI system for narrative maps addresses the challenge of providing meaningful explanations across multiple levels of abstraction in the narrative extraction process. We note that our XAI system builds upon previous extraction models and interactive prototypes [5, 4] and thus we do not explain the underlying components or the extraction method of the narrative maps in detail. In particular, the system integrates with a mixed multi-model pipeline [4] that combines low-level continuous spaces for document representation with high-level discrete structures for narrative visualization.

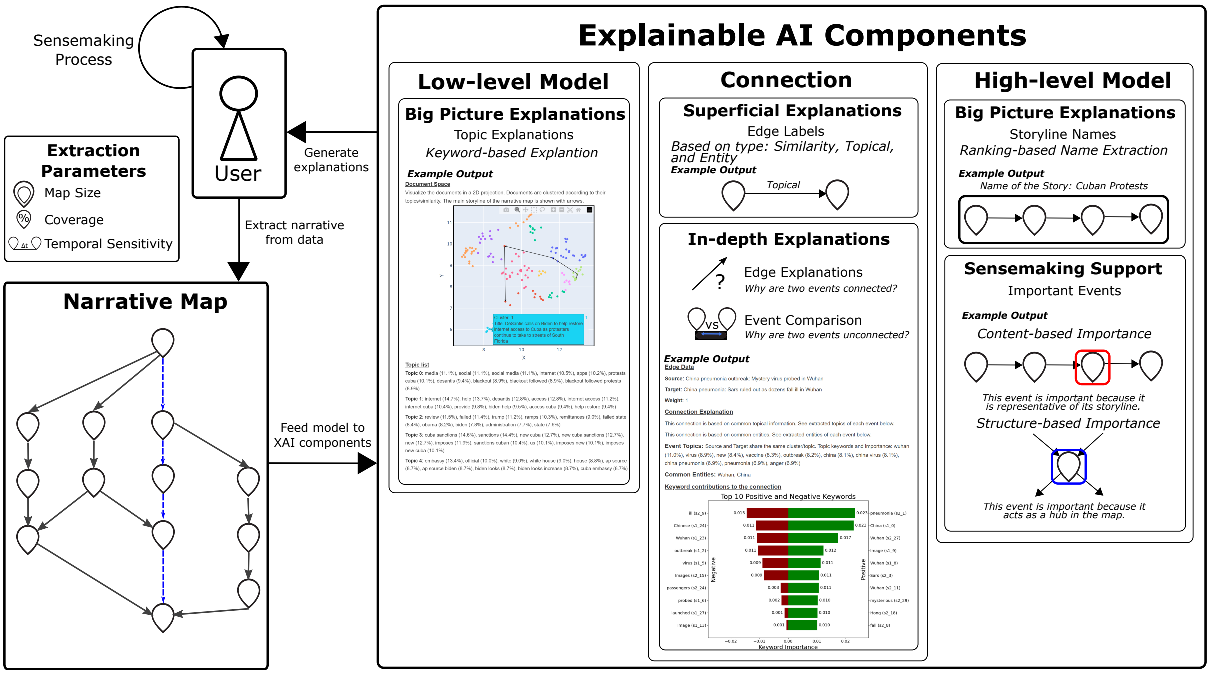
In general, the extraction process takes news articles as input and proceeds in two main phases: extraction and post-processing [4]. During extraction, the system maps articles into an embedding space, computes coherence between events using information about topical clusters and similarity measures, and uses linear programming to build the optimal narrative structure [5]. Post-processing then simplifies this structure, following design guidelines [23] to make it more understandable by removing redundant connections while preserving the core narrative elements.

At the foundation of our implementation lies the principle that explanations must bridge the gap between the computational processes of narrative extraction and the cognitive processes of human analysts. We accomplish this through a three-tiered approach that provides explanations for the document space (a low-level model without structure), the narrative structure (a high-level model that captures the underlying narrative connections), and the *connection* between these two. These explanations are generated through a combination of model-agnostic and model-dependent techniques, carefully selected to maintain computational efficiency while providing meaningful insights.

We show the pipeline of our system in Figure 1. To use the system, the user extracts a narrative map representation from data with user-defined parameters (map size, story coverage, and temporal sensitivity). The narrative representation is then fed to the XAI components. The XAI components focus on each of the previously mentioned tiers (the low-level model, the high-level model, and the connection between these two).

For the low-level model, the XAI system uses keyword representations of topical clusters to provide big picture explanations that seek to capture general patterns in the document space. For the high-level model, the XAI system uses storyline names to provide big picture explanations. The storyline names are extracted from relevant parts of the documents using a ranking-based name extraction algorithm. The XAI system also provides supporting explanations for sensemaking purposes. In particular, the system identifies important events to help users identify highly relevant documents. These important events are identified based on the relevance of their content with respect to their storyline or whether they are relevant to the overall structure of the narrative (e.g., acting as a central node in the graph).

Finally, for the connection between the two models, the XAI system assigns labels to each connection, which provide a super explanation of the type of connection. The types identified by the XAI system are based on a previous taxonomy used by analysts during the narrative sensemaking process [23]. While these labels provide a general view of why two specific events were connected, they do not provide sufficient details. Thus, we expand upon each relevant element of the different types of connections, including a breakdown of the *topic* to which each event belongs, the common *entities* they have, and the specific contributions of each *keyword* towards similarity. The information shown depends on the specific label assigned to the connection (e.g., a connection with an “Entity” label will display the common entities). For the topical information, we specify whether the events share common topics and the keywords that define these topics. Following the same format from the explanations of the low-level space. For common entities, we simply intersect the sets of identified entities and display them accordingly. To find the keywords with the highest contribution (in positive or negative terms), we use the SHAP library [11] with a permutation-based approach.



Document Space

Visualize the documents in a 2D projection. Documents are clustered according to their topics/similarity. The main storyline of the narrative map is shown with arrows.

Topic list

Topic 0: media (11.1%), social (11.1%), social media (11.1%), internet (10.5%), apps (10.2%), protests cuba (10.1%), desantis (9.4%), blackout (8.9%), blackout followed (8.9%), blackout followed protests (8.9%)

Topic 1: internet (14.7%), help (13.7%), desantis (12.8%), access (12.8%), internet access (11.2%), internet cuba (10.4%), provide (9.8%), biden help (9.5%), access cuba (9.4%), help restore (9.4%)

Topic 2: review (11.5%), failed (11.4%), trump (11.2%), ramps (10.3%), remittances (9.0%), failed state (8.4%), obama (8.2%), biden (7.8%), administration (7.7%), state (7.6%)

Topic 3: cuba sanctions (14.6%), sanctions (14.4%), new cuba (12.7%), new cuba sanctions (12.7%), new (12.7%), imposes (11.9%), sanctions cuban (10.4%), us (10.1%), imposes new (10.1%), imposes new cuba (10.1%)

Topic 4: embassy (13.4%), official (10.0%), white (9.0%), white house (9.0%), house (8.8%), ap source (8.7%), ap source biden (8.7%), biden looks (8.7%), biden looks increase (8.7%), cuba embassy (8.7%)

Edge Data

Source: China pneumonia outbreak: Mystery virus probed in Wuhan

Target: China pneumonia: Sars ruled out as dozens fall ill in Wuhan

Weight: 1

Connection Explanation

This connection is based on common topical information. See extracted topics of each event below.

This connection is based on common entities. See extracted entities of each event below.

Event Topics: Source and Target share the same cluster/topic. Topic keywords and importance: wuhan (11.0%), virus (8.9%), new (8.4%), vaccine (8.3%), outbreak (8.2%), china (8.1%), china virus (8.1%), china pneumonia (6.9%), pneumonia (6.9%), anger (6.9%)

Common Entities: Wuhan, China

Keyword contributions to the connection

Figure 1: Implemented pipeline showing the extraction of narrative map and explainable AI components to aid users in understanding the underlying model, including topic clusters, connection labels, and storyline names.

3.3. Low-level Space Explanations

The low-level space explanations seek to help analysts understand the document embedding space and the topical relationships between events. We implement this by identifying topical clusters to reveal the underlying structure of the document space. In particular, we use HDBSCAN clustering [24] to identify coherent groups of documents in the embedding space. This hierarchical clustering method was used because it has proven to work reliably for text data in the context of narrative extraction [5, 4]. Furthermore, it can be used to extract cluster probability vectors (i.e. soft clustering), allowing for a more nuanced assignment of events to each cluster [24].

For each cluster, we generate keyword-based explanations using a modified TF-IDF (Term Frequency and Inverse Document Frequency) representation [25] that incorporates both global and local term importance, allowing us to capture cluster-specific terminology while maintaining context from the broader document collection. Specifically, we compute the importance score S for term t in cluster c as:

$$S(t,c) = TF(t,c) \cdot IDF_{global}(t) \cdot IDF_{local}(t,c) \tag{1}$$

where $TF(t,c)$ represents the term frequency in the cluster, $IDF_{global}(t)$ captures the term's importance across the entire corpus, and $IDF_{local}(t,c)$ measures the term's specificity within the cluster.

The cluster visualization provides a spatial view of the document relationships through a 2D projection using UMAP [26]. This projection preserves both local and global structure, allowing analysts to see how documents relate to each other within and across topic clusters. This visualization includes interactive tooltips that display the cluster membership and key terms for specific documents.

To ensure the explanations remain interpretable, we limit the number of keywords shown for each cluster based on empirical testing with analysts. The system displays the top-k keywords where k is determined dynamically based on the cluster size and keyword importance distribution. This approach prevents information overload while ensuring that sufficient context is provided for understanding each topic cluster. We show the topic explanations in Figure 2(a).

3.4. Connection Explanations

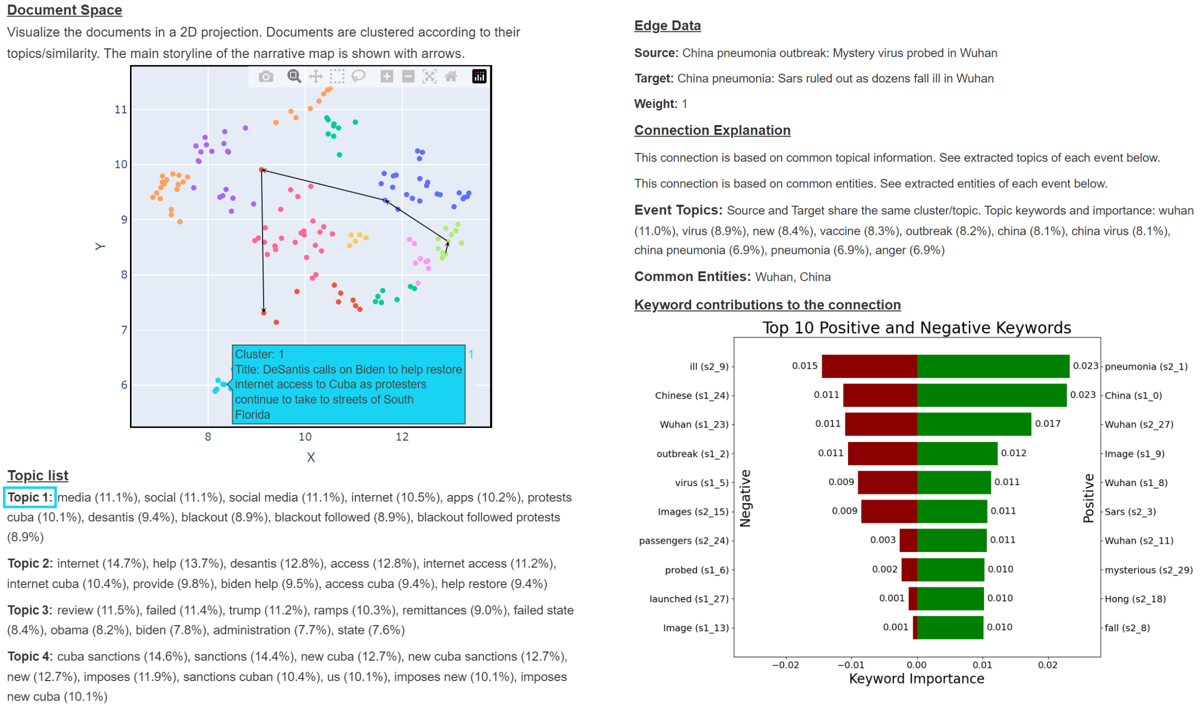
The connection explanation component addresses the critical task of explaining why specific events are connected in the narrative. Our approach generates these explanations through three interconnected processes: connection label generation, detailed explanation generation, and event comparison.

Connection label generation relies on a taxonomy derived from narrative map design guidelines [23]. The system classifies connections into three primary types: similarity-based, entity-based, and topical connections. To determine the connection type, we analyze the components of the coherence measure used in the extraction process. The system computes the relative contribution of clustering similarity versus basic text similarity, assigning the label “Topical” when the clustering component contributes more than 50% to the coherence score, and “Similarity” otherwise. Entity-based connections are identified through named entity recognition, with an additional overlap score based on Jaccard similarity to handle partial entity matches:

$$overlap(e_1,e_2) = \frac{|tokens(e_1) \cap tokens(e_2)|}{|tokens(e_1) \cup tokens(e_2)|} \tag{2}$$

For detailed explanation generation, we implement a model-agnostic approach using SHAP values [11] to identify the most influential terms contributing to event connections. The system generates explanations by analyzing both positive and negative contributions to the connection strength. We modify the standard SHAP implementation to produce interpretable explanations by focusing on the headline and first thirty words of each event, which typically contain the most relevant information in news narratives [6]. We show the connection explanation function in Figure 2(b).

The event comparison functionality extends these explanation capabilities to help analysts understand why certain events are not connected. This component applies the same analysis techniques used for connected events but focuses on explaining the factors that resulted in events remaining unconnected.



Document Space

Visualize the documents in a 2D projection. Documents are clustered according to their topics/similarity. The main storyline of the narrative map is shown with arrows.

Y
11
10
9
8
7
6
8
10
12
X
Cluster: 1
Title: DeSantis calls on Biden to help restore internet access to Cuba as protesters continue to take to streets of South Florida
1

Topic list

Topic 1: media (11.1%), social (11.1%), social media (11.1%), internet (10.5%), apps (10.2%), protests cuba (10.1%), desantis (9.4%), blackout (8.9%), blackout followed (8.9%), blackout followed protests (8.9%)

Topic 2: internet (14.7%), help (13.7%), desantis (12.8%), access (12.8%), internet access (11.2%), internet cuba (10.4%), provide (9.8%), biden help (9.5%), access cuba (9.4%), help restore (9.4%)

Topic 3: review (11.5%), failed (11.4%), trump (11.2%), ramps (10.3%), remittances (9.0%), failed state (8.4%), obama (8.2%), biden (7.8%), administration (7.7%), state (7.6%)

Topic 4: cuba sanctions (14.6%), sanctions (14.4%), new cuba (12.7%), new cuba sanctions (12.7%), new (12.7%), imposes (11.9%), sanctions cuban (10.4%), us (10.1%), imposes new (10.1%), imposes new cuba (10.1%)

Edge Data

Source: China pneumonia outbreak: Mystery virus probed in Wuhan

Target: China pneumonia: Sars ruled out as dozens fall ill in Wuhan

Weight: 1

Connection Explanation

This connection is based on common topical information. See extracted topics of each event below.

This connection is based on common entities. See extracted entities of each event below.

Event Topics: Source and Target share the same cluster/topic. Topic keywords and importance: wuhan (11.0%), virus (8.9%), new (8.4%), vaccine (8.3%), outbreak (8.2%), china (8.1%), china virus (8.1%), china pneumonia (6.9%), pneumonia (6.9%), anger (6.9%)

Common Entities: Wuhan, China

Keyword contributions to the connection

(a) **(b)**

Figure 2: Example explanations using a COVID-19 dataset (not used in the tasks of the user study, shown only as an illustration). **(a)** Topical cluster explanations for the low-level space, including a scatter plot of the space and an overlay of the main storyline. Tooltips with the same color as the corresponding cluster are also included. In this figure, we also highlighted the corresponding description on the topic list. **(b)** Explanation for an edge between two events. This explanation specifies the type of connection (e.g., topical or entity-based), the topics of the events, and the keyword contributions that directly impact their similarity and thus the coherence score.

3.5. High-level Structure Explanations

High-level structure explanations focus on making the global narrative structure interpretable through automated storyline naming and important event detection. These explanations help analysts understand the broader narrative patterns while maintaining connection to the underlying evidence.

3.5.1. Storyline Name Extraction

The storyline naming process builds upon the work of Laban and Hearst [12] in timeline summarization. Our approach identifies candidate names by extracting maximal noun phrases that contain both proper nouns and abstract terms. We show a simplfiied, but illustrative version of the model to score these candidates using a linear combination of factors:

$$Score(name) = \alpha \cdot C_{entity} + \beta \cdot C_{abstract} + \gamma \cdot C_{coverage} - \delta \cdot O_{overlap} \tag{3}$$

where C_{entity}, $C_{abstract}$, and $C_{coverage}$ represent the presence of entities, abstract terms, and coverage of storyline content respectively, while $O_{overlap}$ penalizes redundancy with existing storyline names. The weights α, β, γ, and δ are determined empirically.

3.5.2. Important Event Detection

Important event detection combines both content-based and structural approaches to identify key events in the narrative. For content importance, we compute the similarity between each event's embedding and the centroid of its storyline. Structural importance is assessed through degree centrality in the

Table 1
Post-task questionnaire with Likert scale questions.

Category	Questionnaire Item
General	**Usefulness:** The explanations provided by the system were useful. **Trust:** The explanations made me trust the system more.
Storyline Names	**Correctness:** The storyline names provided by the system were appropriate for each storyline. **Relevance:** The storyline names provided by the system were relevant. **Usefulness:** The storyline names provided by the system made the narrative easier to understand.
Connections	**Labels Correctness:** The connection labels provided by the system about the types of connections made sense. **Labels Usefulness:** The connection labels provided by the system about the types of connections were useful to understand the connections. **Connection Explanations:** The in-depth explanations for connections provided by the system were useful. **Event Comparison:** The event comparison explanations for unconnected events provided by the system were useful.
Important Events	**Relevance:** The important events selected by the system were relevant. **Usefulness:** The important events selected by the system made the narrative easier to understand.

narrative graph, weighted by the coherence values of the connections. The system identifies important events by selecting the top-n events according to each criterion:

$$I_{content}(e) = cos(e, centroid_{storyline}) \tag{4}$$

$$I_{structure}(e) = \sum_{v \in N(e)} coherence(e, v) \tag{5}$$

The explanation system visually marks these important events on the narrative map. Events that score highly on both content and structural measures receive additional emphasis, as they represent key narrative elements that are both thematically central and well-connected.

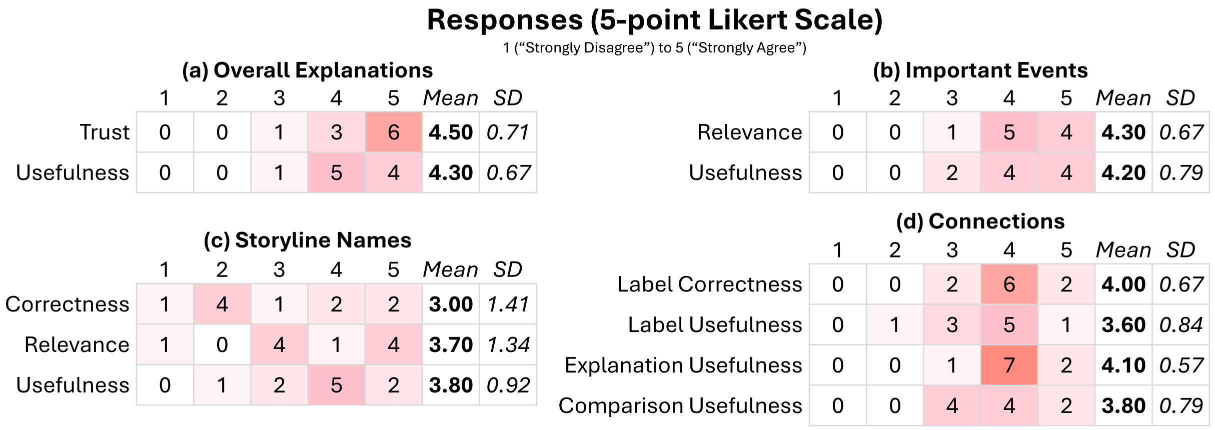
3.6. Evaluation of the System

We evaluated our system through a user study focused on analyzing narratives about the 2021 Cuban protests. Using an insight-based evaluation methodology [27], we assessed how effectively our XAI components supported narrative sensemaking tasks. The dataset used in this user study comprised 160 news articles from diverse sources, providing comprehensive coverage while maintaining manageable computational requirements. We recruited 10 participants with backgrounds in computer science, communications, and national security. The participants reported minimal prior knowledge of the 2021 Cuban protests, with a mean (M) *familiarity* of 1.4 on a 5-point Likert scale and a standard deviation (SD) of 0.52. The participants first received a 15-minute training session using a separate COVID-19 news dataset [23] to familiarize themselves with the features of the system, ensuring that they could focus on evaluating the XAI features rather than learning the mechanics of the system.

The insight-based evaluation required users to obtain as many insights as possible with our system. The insights were then categorized and counted. This open-ended task gives analysts freedom to explore the dataset and provides an approximation of a realistic narrative sensemaking task. Finally, we asked participants to complete a follow-up questionnaire (see Table 1) on their perception of the XAI components to help develop an understanding of the narrative structure and verify their analysis.

4. Evaluation Results

Our evaluation focused on user perception and trust in the system's XAI components. We show the results in Figure 3. The results are reported on a 5-point Likert scale.



Responses (5-point Likert Scale)

1 (“Strongly Disagree”) to 5 (“Strongly Agree”)

(a) Overall Explanations

	1	2	3	4	5	*Mean*	*SD*
Trust	0	0	1	3	6	**4.50**	*0.71*
Usefulness	0	0	1	5	4	**4.30**	*0.67*

(b) Important Events

	1	2	3	4	5	*Mean*	*SD*
Relevance	0	0	1	5	4	**4.30**	*0.67*
Usefulness	0	0	2	4	4	**4.20**	*0.79*

(c) Storyline Names

	1	2	3	4	5	*Mean*	*SD*
Correctness	1	4	1	2	2	**3.00**	*1.41*
Relevance	1	0	4	1	4	**3.70**	*1.34*
Usefulness	0	1	2	5	2	**3.80**	*0.92*

(d) Connections

	1	2	3	4	5	*Mean*	*SD*
Label Correctness	0	0	2	6	2	**4.00**	*0.67*
Label Usefulness	0	1	3	5	1	**3.60**	*0.84*
Explanation Usefulness	0	0	1	7	2	**4.10**	*0.57*
Comparison Usefulness	0	0	4	4	2	**3.80**	*0.79*

Figure 3: A tally of the answers to our survey questions on the *explainable AI* components.

Overall Explanations. Survey results indicated that, in general, the explanation components significantly increased user trust in the system (M = 4.5, SD = 0.71) and were considered useful by the participants (M = 4.3, SD = 0.67). In particular, the participants reported that the explanations of the topical clusters using TF-IDF increased confidence in understanding the clustering of documents.

Important Events. The important events detected by the system were considered relevant (M = 4.3, SD = 0.67) and useful (M = 4.2, SD = 0.79). Compared to other components of the system (story names and connection labels), important event detection proved to be more consistently valuable, helping participants quickly identify key narrative elements.

Storyline Names. These high-level structure explanations showed mixed results in building trust. Automated storyline naming received more variable feedback in terms of correctness, presenting a higher standard deviation and lower mean (M = 3.0, SD = 1.41), and only slightly better in terms of relevance (M = 3.7, SD = 1.34) and usefulness (M = 3.8, SD = 0.92). Paraphrased from a participant: “*Storyline names are not always correct, but they are relevant and useful.*”

Connections. Label-based explanations demonstrated varying levels of trust enhancement among participants. The generated labels were generally considered correct (M = 4.0, SD = 0.67), but slightly less useful (M = 3.6, SD = 0.84). The SHAP-based keyword explanations helped the participants verify connection validity, increasing confidence in the system’s linking decisions, and were considered generally useful (M = 4.1, SD = 0.57). The comparison tool was considered slightly less useful (M = 3.8, SD = 0.79). Analysis of participant feedback revealed that the combination of explanations at multiple levels enhanced trust in the system’s narrative extraction capabilities.

These results suggest that our XAI system effectively supports user trust in narrative analysis systems, particularly when explanations span multiple levels of abstraction. The variable effectiveness of different explanation types indicates opportunities for future refinement of explanation strategies to better align with user expectations and trust-building needs.

5. Discussion and Limitations

5.1. Explainable AI Results

Our results indicate that XAI components substantially support narrative sensemaking tasks when properly integrated into the analysis workflow. In particular, the participants of our user study considered the addition of XAI beneficial and it increased their trust in the underlying models, even if the explanations themselves were not always particularly useful for the assigned task. However, our user evaluation also shows that there is some contention on the usefulness of each individual component, such as storyline names, which were sometimes considered very useful and sometimes discarded as mostly useless, depending on the participant.

For most use cases, XAI provides useful scaffolding and enables users to potentially find more insights. However, developing appropriate XAI methods that capture the intricacies of the underlying models and are useful from a user's perspective is a complex task. In general, based on our results, we recommend designing such explanations with the goal of helping users understand the big picture by providing general pointers and providing support for specific confirmatory tasks, such as checking whether an element of the model is valid (e.g., event connections). In our evaluation, methods that provided too much information or were too technical in nature were generally regarded negatively, as participants were likely overwhelmed by the information overload of the explanations [20].

Our findings provide some practical insights for future narrative extraction systems. In particular, the variable effectiveness of different explanation types demonstrates that explanation strategies should be tailored to specific components of the narrative structure, rather than applying a single approach uniformly. Furthermore, our analysis of results suggests that a design approach that combines multiple explanation modalities while prioritizing cognitive accessibility over technical comprehensiveness. In general, we found that explanations contributed to user-perceived trust, which underscores the importance of transparency in AI-assisted analytical tasks.

5.2. Trust Building with Explainable AI

The proposed explainable AI framework relies on a set of machine learning and data mining algorithms to interpret the semantic representations embedded in the narrative maps. While this method provides a surface-level explanation of the outcomes of the model, it introduces a significant challenge to the reliability and validity of the generated outcomes. Minor adjustments at each stage of the process can yield disparate outputs, thereby raising concerns about the consistency and accuracy of the results. However, from the results of our user study, despite these potential limitations, the users still found the system to be useful for solving sensemaking tasks.

Similarly, XAI methods often function as proxies that attempt to clarify the impact or contribution of various features within the model to increase transparency [21]. For example, in the case of our narrative map explanations, we leverage topical clusters and storyline name extraction to summarize the events and provide a general overview. However, even if these explanations may only capture a surface-level understanding of the processes governing the system, our user study shows that these explanations are still useful, even if they do not fully represent the underlying model.

Moreover, a key concern when implementing XAI methods is the potential risk of providing inaccurate or misleading explanations, which, in turn, could lead to a decrease in user trust in the system [28, 29, 30] and the underlying AI model [22]. This is an issue that should be carefully addressed to ensure that XAI methods are transparent, trustworthy, and reliable. Thus, to ensure that users are aware that the algorithm outputs may not always be perfectly aligned with the XAI components, future versions of the system can incorporate warning mechanisms or indicators that can signal when the explanations provided by the model might be less reliable or accurate.

When it comes to the storyline names, text summarization is an active area of research [31] with many challenges and opportunities, particularly in handling context, maintaining coherence, and ensuring accuracy while condensing large volumes of information. While the current method follows an extractive strategy, a shift towards an *abstractive* strategy might be more effective [32]. This approach would involve creating names with words that might not directly exist in the storylines but accurately describe them. Implementing this could be achieved with generative neural networks [33], which can distill the essential aspects of the storyline.

5.3. Limitations

First, we note that we did not compare the proposed system with a proper baseline in our user study, such as another system from the literature or the same system without XAI. Instead, our focus was on exploring how participants used the features and whether they considered them useful. The findings of this study could help inform future versions of such XAI systems to provide better user support.

Furthermore, we note that participants had no easy way to determine whether the extracted narratives were indeed correct, as the system does not provide evaluation metrics on the factual accuracy of the narratives. However, evaluating the correctness of narratives in general is an open problem, as there are no unified metrics that work in all cases [5].

Scalability remains a consideration for larger datasets. Although our implementation handled the test dataset effectively, the computational complexity of generating comprehensive explanations may become prohibitive for larger narrative collections. Future work could explore hierarchical explanation strategies that can scale more effectively to larger datasets while maintaining explanation quality.

Despite these limitations, our results demonstrate the value of integrated XAI components in narrative analysis systems. The positive user feedback on explanation utility suggests that our approach effectively supports the sensemaking process of analysts. Future work could address these limitations while expanding XAI methods to handle additional types of narrative relationships and explanation needs.

6. Conclusions

This paper presented the evaluation of an XAI system to explain AI-driven narrative extraction systems through multiple levels of abstraction. Our approach bridges the gap between low-level text processing and high-level narrative structures, providing analysts with meaningful explanations that enhance trust in automated narrative analysis. Through empirical evaluation, we have shown that integrated explainable AI components improve user confidence in working with complex narrative structures.

The results of our user study indicate that narrative maps augmented with XAI techniques effectively build user trust in automated analysis systems. The combination of topical cluster explanations, connection explanations, and high-level structure explanations provides analysts with a coherent understanding of system decisions at multiple levels. Our evaluation demonstrates that connection type labels and important event detection enhance user confidence, while more complex explanations require additional refinement to maximize their trust-building potential.

Several promising directions emerge for future research. The development of more sophisticated temporal and causal explanation strategies could further enhance user trust in automated narrative extraction. In addition, the exploration of adaptive explanation approaches that respond to different levels of user expertise and trust requirements presents an important avenue for investigation.

The broader implications of this work extend beyond narrative analysis to the general challenge of building trustworthy AI systems. Our findings suggest that carefully designed explanations at multiple levels of abstraction can effectively support human-AI collaboration by establishing appropriate levels of trust. As narrative analysis systems continue to evolve, the principles and approaches developed in this work can inform the design of future explainable AI systems that users can confidently rely upon for complex analytical tasks.

Acknowledgments

This research was supported by industry, government, and institute members of the NSF SHREC Center, which was founded in the IUCRC program of the National Science Foundation. Brian Keith is also supported by Project 202311010033-VRIDT-UCN.

Declaration on Generative AI

During the preparation of this work, the authors used Grammarly and Writefull integrated with Overleaf to perform grammar and spelling corrections. After using these tools/services, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] B. F. Keith Norambuena, T. Mitra, C. North, A survey on event-based news narrative extraction, ACM Comput. Surv. 55 (2023). URL: https://doi.org/10.1145/3584741. doi:10.1145/3584741.

[2] A. Barredo Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, F. Herrera, Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai, Information Fusion 58 (2020) 82–115.

[3] A. Adadi, M. Berrada, Peeking inside the black-box: A survey on explainable artificial intelligence (xai), IEEE Access 6 (2018) 52138–52160.

[4] B. F. Keith Norambuena, T. Mitra, C. North, Mixed multi-model semantic interaction for graph-based narrative visualizations, in: Proceedings of the 28th International Conference on Intelligent User Interfaces, IUI ’23, Association for Computing Machinery, New York, NY, USA, 2023, p. 866–888. URL: https://doi.org/10.1145/3581641.3584076. doi:10.1145/3581641.3584076.

[5] B. Keith Norambuena, T. Mitra, Narrative maps: An algorithmic approach to represent and extract information narratives, in: Proc. ACM Hum.-Comput. Interact., volume 4, ACM, New York, NY, USA, 2020, p. 33 pages.

[6] B. Keith Norambuena, M. Horning, T. Mitra, Evaluating the inverted pyramid structure through automatic 5w1h extraction and summarization, in: Proc. of the 2020 Computation + Journalism Symposium, C + J 2020, Boston, MA, USA, 2020, pp. 1–7.

[7] J. Wenskovitch, C. North, Interactive artificial intelligence: Designing for the” two black boxes” problem, Computer 53 (2020) 29–39.

[8] B. F. Keith Norambuena, T. Mitra, C. North, Narrative sensemaking: Strategies for narrative maps construction, in: 2021 IEEE Visualization Conference (VIS), IEEE, New Orleans, LA, USA, 2021, pp. 181–185.

[9] D. Kim, A. Oh, Topic chains for understanding a news corpus, in: A. Gelbukh (Ed.), Computational Linguistics and Intelligent Text Processing, Springer Berlin Heidelberg, Berlin, Heidelberg, 2011, pp. 163–176.

[10] R. Churchill, L. Singh, The evolution of topic modeling, ACM Comput. Surv. 54 (2022).

[11] S. M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, in: I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.), Advances in Neural Information Processing Systems, volume 30, Curran Associates, Inc., Long Beach, CA, USA, 2017. URL: https://proceedings.neurips.cc/paper_files/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf.

[12] P. Laban, M. A. Hearst, newslens: building and visualizing long-ranging news stories, in: Proc. of the Events and Stories in the News Workshop, ACL, Vancouver, Canada, 2017, pp. 1–9.

[13] R. Hoffman, S. Mueller, G. Klein, J. Litman, Measuring trust in the xai context (2021).

[14] R. Konig, U. Johansson, L. Niklasson, G-rex: A versatile framework for evolutionary data mining, in: 2008 IEEE International Conference on Data Mining Workshops, IEEE, Pisa, Italy, 2008, pp. 971–974.

[15] M. T. Ribeiro, S. Singh, C. Guestrin, ”why should i trust you?”: Explaining the predictions of any classifier, in: Proc. of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD ’16, ACM, New York, NY, USA, 2016, p. 1135–1144.

[16] F. Hohman, M. Kahng, R. Pienta, D. H. Chau, Visual analytics in deep learning: An interrogative survey for the next frontiers, IEEE trans. on visualization and computer graphics 25 (2018) 2674–2693.

[17] A. S. Vivacqua, R. Stelling, A. C. B. Garcia, L. C. Gouvea, Explanations and sensemaking with ai and hci, in: Proc. of the IX Latin American Conference on Human Computer Interaction, CLIHC ’19, ACM, New York, NY, USA, 2020.

[18] C. C. Stephanidis, G. Salvendy, M. of the Group Margherita Antona, J. Y. C. Chen, J. Dong, V. G. Duffy, X. Fang, C. Fidopiastis, G. Fragomeni, L. P. Fu, Y. Guo, D. Harris, A. Ioannou, K. ah (Kate) Jeong, S. Konomi, H. Krömker, M. Kurosu, J. R. Lewis, A. Marcus, G. Meiselwitz,

A. Moallem, H. Mori, F. F.-H. Nah, S. Ntoa, P.-L. P. Rau, D. Schmorrow, K. Siau, N. Streitz, W. Wang, S. Yamamoto, P. Zaphiris, J. Zhou, Seven hci grand challenges, International Journal of Human–Computer Interaction 35 (2019) 1229–1269.

[19] H. P. Abbott, The Cambridge introduction to narrative, Cambridge University Press, One Liberty Plaza, New York, NY, USA, 2008.

[20] A. N. Ferguson, M. Franklin, D. Lagnado, Explanations that backfire: Explainable artificial intelligence can cause information overload, in: Proc. of the Annual Meeting of the Cognitive Science Society, volume 44, Cognitive Science Society, Sydney, NSW, Australia, 2022.

[21] M. T. Ribeiro, S. Singh, C. Guestrin, "why should i trust you?": Explaining the predictions of any classifier, 2016. arXiv:1602.04938.

[22] J. Zerilli, U. Bhatt, A. Weller, How transparency modulates trust in artificial intelligence, Patterns 3 (2022) 100455. URL: https://www.sciencedirect.com/science/article/pii/S2666389922000289. doi:https://doi.org/10.1016/j.patter.2022.100455.

[23] B. F. Keith Norambuena, T. Mitra, C. North, Design guidelines for narrative maps in sensemaking tasks, Information Visualization 21 (0) 220–245.

[24] L. McInnes, J. Healy, S. Astels, hdbscan: Hierarchical density based clustering, Journal of Open Source Software 2 (2017) 205.

[25] F. Debole, F. Sebastiani, Supervised term weighting for automated text categorization, in: Proceedings of the 2003 ACM Symposium on Applied Computing, SAC '03, Association for Computing Machinery, New York, NY, USA, 2003, p. 784–788. URL: https://doi.org/10.1145/952532.952688. doi:10.1145/952532.952688.

[26] L. McInnes, J. Healy, J. Melville, Umap: Uniform manifold approximation and projection for dimension reduction, arXiv preprint arXiv:1802.03426, 2018.

[27] C. North, P. Saraiya, K. Duca, A comparison of benchmark task and insight evaluation methods for information visualization, Information Visualization 10 (2011) 162–181.

[28] D. Manzey, J. Reichenbach, L. Onnasch, Human performance consequences of automated decision aids, Journal of Cognitive Engineering and Decision Making 6 (2012) 57–87. doi:10.1177/1555343411433844.

[29] M. T. Dzindolet, S. A. Peterson, R. A. Pomranky, L. G. Pierce, H. P. Beck, The role of trust in automation reliance, International Journal of Human-Computer Studies 58 (2003) 697–718. doi:10.1016/S1071-5819(03)00038-7.

[30] B. J. Dietvorst, S. Bharti, People reject algorithms in uncertain decision domains because they have diminishing sensitivity to forecasting error, Psychological Science 31 (2020) 1302–1314. doi:10.1177/0956797620948841.

[31] A. P. Widyassari, S. Rustad, G. F. Shidik, E. Noersasongko, A. Syukur, A. Affandy, D. R. I. M. Setiadi, Review of automatic text summarization techniques & methods, Journal of King Saud University - Computer and Information Sciences 34 (2022) 1029–1046. URL: https://www.sciencedirect.com/science/article/pii/S1319157820303712. doi:https://doi.org/10.1016/j.jksuci.2020.05.006.

[32] H. Lin, V. Ng, Abstractive summarization: A survey of the state of the art, Proceedings of the AAAI Conference on Artificial Intelligence 33 (2019) 9815–9822. URL: https://ojs.aaai.org/index.php/AAAI/article/view/5056. doi:10.1609/aaai.v33i01.33019815.

[33] Y. Gao, Y. Wang, L. Liu, Y. Guo, H. Huang, Neural abstractive summarization fusing by global generative topics, Neural Computing and Applications 32 (2020) 5049–5058. URL: https://doi.org/10.1007/s00521-018-3946-7. doi:10.1007/s00521-018-3946-7.