A Framework for Comparative Analysis of News Content: A Model-Based Approach

Bahareh Fatemi[1(✉)], Fazle Rabbi[1], Yngve Lamo[2], and Andreas L. Opdahl[1]

[1] University of Bergen, Bergen, Norway
{bahareh.fatemi,fazle.rabbi,andreas.opdahl}@uib.no
[2] Western Norway University of Applied Sciences, Bergen, Norway
yngve.lamo@hvl.no

Abstract. In the digital age, the volume of news data available from diverse sources is vast and continually growing. On the one hand, the quantity of information can overwhelm reporters and on the other hand, news reporting is further complicated by the inherent complexities of multifaceted events that evolve over time, as well as the biases and perspectives that different reporters and media outlets bring to their coverage. Despite such challenges, journalists must report on events in a timely and ethical manner. However, there is a lack of computational methods for analyzing massive news streams in an explainable and responsible way. In this paper, we propose a content based news analysis framework based on news comparison that enables modeling various analytical tasks such as analyzing the perspectives of news publishers, monitoring the progression of news events from various perspectives, exploring the evolution patterns of events over time and analyzing news article variants and for uncovering underlying story-lines. Our approach utilizes a knowledge graph to represent key concepts in the news domain, such as events and their contextual information, across various dimensions. This facilitates a multi-dimensional and comparative analysis of news article variants. We demonstrate the practical applicability of our method through a running example. By adopting a model-based approach, our framework offers the flexibility needed to represent a broad spectrum of domain concepts.

Keywords: Category theory · Content analysis · Model-based framework · Knowledge graph · Natural language processing · Computational journalism

1 Introduction

In every human community, individuals share news to keep one another informed about significant events and developments. News plays a crucial role in this exchange of information, and journalists are tasked with the responsibility of turning facts into engaging and informative stories [21]. Good journalism is grounded in core ethical principles such as *trust and accuracy*; *independence*; *fairness and impartiality*; *humanity*; and *accountability* [2]. However, it is important to recognize that completely unbiased journalism is an ideal that is rarely achieved in practice [9,21]. The subjective nature of news reporting and the presence of bias are inherent aspects of the media landscape.

F. José Domínguez Mayo et al. (Eds.): MODELSWARD 2024, CCIS 2547, pp. 3–22, 2026.
https://doi.org/10.1007/978-3-031-96841-9_1

One of the key challenges for journalists is to account for these various perspectives and present a balanced view of events, despite the subjective nature of news. A significant challenge for journalists is to navigate these diverse perspectives and offer a balanced portrayal of events. Journalists also need to keep track of ongoing events worldwide, and carefully analyze their dynamics to inform their audience about the changing world. As global media coverage grows more abundant, journalists and researchers face the challenge of distilling complex and evolving narratives from vast amounts of data. The need to analyze event progression stems from the desire to uncover how events unfold, identify emerging trends, and understand the dynamics and causal relationships among different phases of an event. However, this task is complicated by the dynamic nature of news coverage and the vast amount of information from diverse sources. A major challenge in this process is managing and interpreting data from numerous publishers, each offering unique and sometimes conflicting narratives. This need for effective analysis extends beyond journalism to other disciplines, including sociology, history, political science, and information science, where professionals also rely on news articles for various forms of research and insight. In this paper we present a model-based framework that employs a diverse range of models to represent knowledge from news articles and uses computational methods for the analysis of news events. This paper is an extended version of our previous work [15], where the foundational concepts of a multidimensional meta-model for news content analysis were introduced. The extended framework integrates the following components:

- state-of-the-art natural language processing technique for parsing content from news articles;
- a multi dimensional meta-model allowing data to be arranged into hierarchical groups and a knowledge graph schema for structuring event related information;
- a content comparison method based on category theory;
- a logical framework for capturing the content-based dynamic patterns of events; and
- a statistical analysis method for analyzing news article variants.

We utilize knowledge graphs to represent news events, incorporating key information such as the source article, publication date, involved persons, involved countries. To enhance the clarity and consistency of our representations, we also annotate news events using IPTC (International Press Telecommunications Council) Media Topics and store it as an important dimension. IPTC is a global standardization organization that provides metadata standards for the news industry. The hierarchical structure of IPTC Media Topics enables the extraction of news events across various levels of abstraction. By combining different attributes and relationships of news events along with the domain ontology in IPTC Media Topics, the framework allows users to extract different views of news events from a knowledge graph.

The framework integrates a computational model based on category theory which allows us to analyze news events at a higher abstraction level, for example, to compare and categorize events and to analyze flow of progression of events. We present novel application areas of category theory for analyzing events stored in a knowledge graph and how they progress.

In Sect. 2, we present a method for extracting structured information about news events from news articles using large language models (LLMs). We present a running

example while describing the proposed method. In Sect. 3, we present our model-based framework for content analysis. In Sect. 4, we provide a discussion about the proposed method and provide a comparison with existing works.

2 Harvesting News Events Knowledge Graph with a Pre-trained LLMs

Harvesting news events into a knowledge graph is an important topic that has been investigated in several projects to support various tasks within the news domain. Opdahl et al. [14] provide a comprehensive review on the use of semantic knowledge graphs in news production, distribution, and consumption, highlighting their potential to integrate heterogeneous information across the news industry. The Global Database of Events, Language, and Tone (GDELT) is a Google-sponsored project that monitors news media from all over the world and provides a real-time update of events in every 15 min [3]. Rospocher et al. present a method to automatically build Event-Centric Knowledge Graphs from news articles using NLP techniques, such as Entity Linking and Semantic Role Labeling [16]. Liu et al. introduce a domain-specific knowledge graph called the "news graph" that incorporates collaborative relations between entities and topic context information for news recommendations [13]. Berven et al. study the harvesting of news events into a knowledge graph by presenting a knowledge graph platform for newsrooms [5]. They propose an event detection technique that identifies potentially newsworthy events from clusters of news items according to named entities, topics, and location.

To structure the information about news events we propose to use a dimensional meta-model (Fig. 1 top) which allows storing events with contexts along various dimensions in a hierarchical model. The lower section of Fig. 1 illustrates a knowledge graph schema designed to structure events and their contextual information, including the event's location, involved countries, and entities. This knowledge graph is further enriched with IPTC Media Topics, enabling access to hierarchical information through the *:HAS_PARENT* relationships. A Neo4j graph database has been used to store news events and their relationships with other entities. The information model is centered around *Event* which also allows us to preserve the epistemic view of individual publishers. For example, if two publishers publish 2 news articles about a certain event, we will be storing 2 instances of $Event$ (along with their contextual information) in our knowledge graph.

In our proposed technique, we take input from GDELT every 15 min. The input includes web addresses to news article texts. These articles are parsed for analysis using pre-trained large language models (LLMs). Specifically, we use GPT-3.5 Turbo to extract information from the news articles and harvest news event related information. LLMs have shown their effectiveness in annotating natural language text based on predefined ontologies[20]. In our previous study we explored the effectiveness of GPT language model in the classification of news articles in IPTC news ontology [8]. Particularly GPT 3.5 Turbo model in zero-shot setting was 82% and 61% successful respectively in first and second level classification of news articles according to IPTC ontology. We explored two prompting strategies namely **simultaneous classification** in

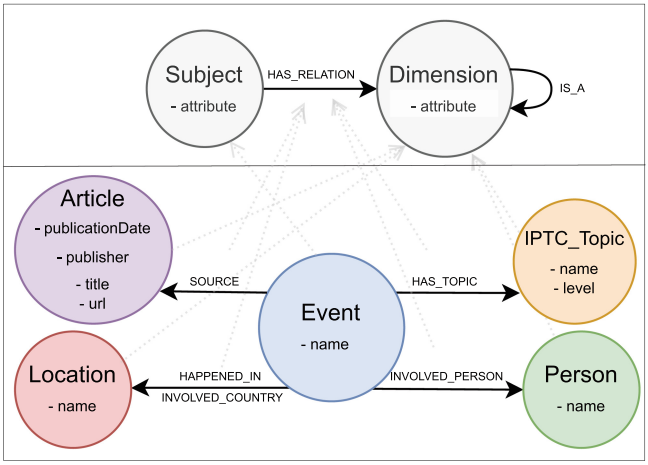


Fig. 1. Dimensional meta-model (top) and Knowledge Graph Schema (bottom) for structuring event related information.

which we provided the model with the entire ontology and tasked it with the simultaneous classification of a news article into one of the Level-1 categories and its corresponding Level-2 subcategories, and **hierarchical classification** in which we initially provided the model with Level-1 categories and requested it to classify the news article accordingly. After determining the Level-1 category, we provided the corresponding subcategories belonging to the chosen category and tasked the model with classifying the news article into a specific subcategory. The experiments showed that in the hierarchical approach some issues related to misclassification and hallucination was effectively resolved or mitigated. Although by fine-tuning GPT model they can perform more effectively, for this study we relying on their ability to annotate data according to IPTC news codes in zero-shot setting.

Figure 2 illustrates the general structure of the prompts we have used to extract structured information from news article texts.

The proposed method in this paper is demonstrated with a running example which includes a knowledge graph of news events about $Niger$ and $Gabon$ extracted from the news articles published by 6 media outlets (*aljazeera.com, theguardian.com, reuters.com, independent.co.uk, nytimes.com, and washingtontimes.com*) from July 28^{th} to September 2^{nd} 2023. The knowledge graph consists of news events in $Niger$ and $Gabon$ about two coups that took place during the above-mentioned period.

3 Model-Based Framework for Content Analysis

We propose a novel model-based framework for news content analysis that incorporates techniques for multidimensional comparative analysis. This framework enables the examination of different perspectives on news content, capturing temporal event patterns, and analyzing the progression of events at various levels of abstraction. It allows the user to select an appropriate dimension and abstraction level. For instance, a user might be interested in comparing the perspectives of different publishers over

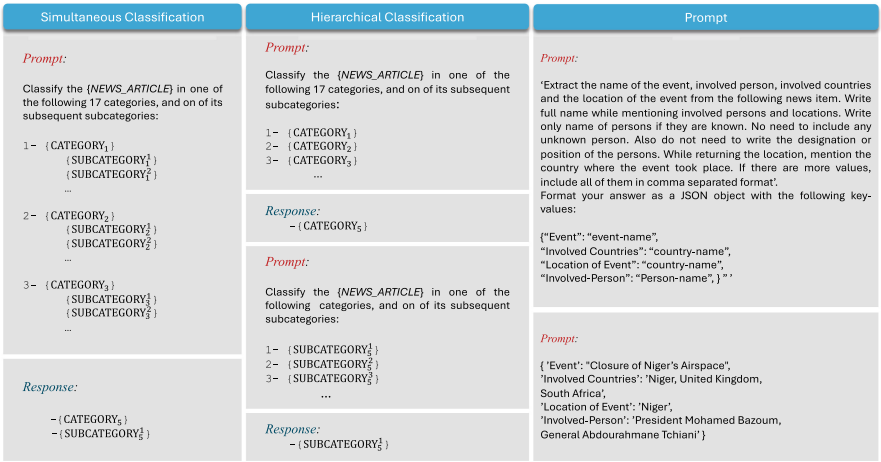


Fig. 2. Prompts for extracting event related information [8, 15].

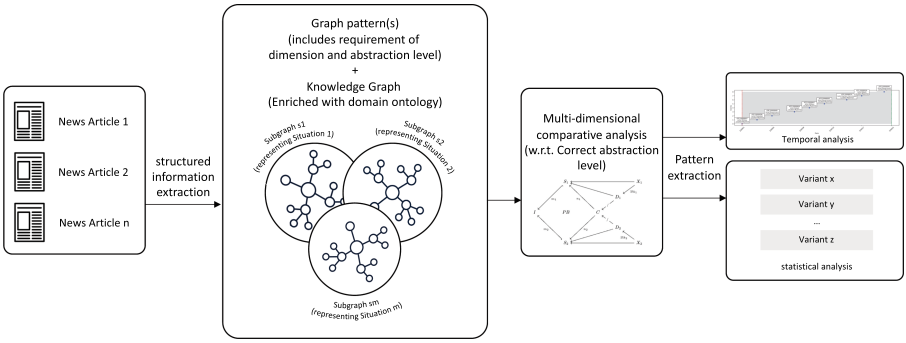


Fig. 3. Model-based framework for multi dimensional comparative analysis of news contents.

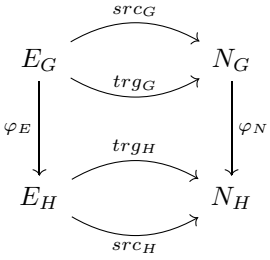
a certain period of time or the progression of events at a certain level of abstraction. The knowledge graph incorporates events and their contextual information across various hierarchically organized dimensions. For instance, the *IPTC* Media topic ontology structures topic names in a hierarchy, where the highest level of abstraction (level 1) includes 17 media topic names. This hierarchical organization allows for the selection of dimensions and abstraction levels to extract information from the knowledge graph, which is then used for comparative analysis. The results of this analysis are utilized to identify patterns of variants. We propose a semi-automated approach to variant analysis, involving human input to enhance the process. Figure 3 illustrates the model-based framework which employs models for representing computational methods for the analysis of news events. Graph patterns are used to specify search criteria. We propose to use categorial operations to perform comparative analysis over the search results (i.e., subgraphs). Category theory allows us to deal with abstract structures and relationships between them. It allows us to study the news content from high levels of abstraction

and thereby enables us to gain deeper insights into media contents. In this paper we focus on the perspective comparison, progression of events, temporal and variant analysis. The model-based framework is adaptive to new dimensions with more contextual information, for example numbers of casualties, sentiments, proximity, news angles, etc.

3.1 Preliminaries

In this section, we present a brief overview of the fundamental concepts of category theory [4] and the diagram predicate framework [19], which will aid in understanding the subsequent discussions and analyses.

Graphs and Category Theory. A **graph** G denoted by $G = (N, E, src^G, trg^G)$ is a collection of nodes N, edges E and maps $src^G, trg^G : E \to N$ which assign a source and a target node to each edge E. A **graph homomorphism** $\varphi : G \to H$ is a pair of maps $\varphi_N : N_G \to N_H$ and $\varphi_E : E_G \to E_H$ that preserves the sources and targets:



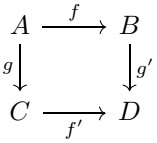
(1)

A **category** C is a structure in which the following elements participate: a set of objects, denoted as $A, B, C \cdots$, a set of morphisms, denoted as $f, g, h, \cdots$, and a relation that associates to each morphism a pair of objects, which is denoted as $f : A \to B$, where A and B represent the domain and co-domain of the morphism f, respectively. The composition of the morphisms is written as $g \circ f$ and has two properties

1. Associativity: If $f : A \to B, g : B \to C$, and $h : C \to D$ then $h \circ (g \circ f) = (h \circ g) \circ f$,
2. Identity: For every object X, there exists an identity morphism $1_X : X \to X$, such that for every morphism $f : A \to B$, we have $1_B \circ f = f = f \circ 1_A$

The **category graph** has graphs as objects and its morphisms are graph homomorphism [17,19].

In category theory, a maximal **pullback** object A (see Diagram 2) is a construction that captures the common elements and relationships between objects B and C in a category, and a minimal **pushout** D is the result of gluing two objects along a common sub-object [6]:



(2)

Table 1. A sample signature $\Sigma_1 = (\Pi^{\Sigma_1}, \alpha^{\Sigma_1})$ used for conflict modelling.

Name ($P \in \Pi^{\Sigma_1}$)	Arity ($\alpha^{\Sigma_1}(P)$)	Semantic Interpretation ($[\![P]\!]$)
`[IPTC_Topic]`	$E \xrightarrow{f} I$	$\exists e \in E, (\lvert f(e)\rvert \geq 0)$
`[Location]`	$E \xrightarrow{g} C$	$\exists e \in E, (\lvert g(e)\rvert \geq 0)$
`[Person]`	$E \xrightarrow{h} P$	$\exists e \in E, (\lvert h(e)\rvert \geq 0)$

Table 2. The atomic constraints from Table 1 and their graph homomorphisms.

Name ($P \in \Pi^{\Sigma_1}$)	Arity ($\alpha^{\Sigma_1}(P)$)	$\delta(\alpha^{\Sigma_1}(P))$
`[IPTC_TOPIC]`	$E \xrightarrow{f} I$	$Event \xrightarrow{\text{HAS\_Topic}} IPTC\_Topic$
`[Location]`	$E \xrightarrow{g} C$	$Event \xrightarrow{\text{HAPPENED\_IN}} Country$
`[Person]`	$E \xrightarrow{h} P$	$Event \xrightarrow{\text{INVOLVED\_PERSON}} Person$

The pullback object can be seen as a generalized intersection of two objects over a common third object, while a pushout can be seen as a disjoint union (where the common part is preserved) of them.

Diagram Predicate Framework (DPF). DPF [19] is a meta-modeling framework founded on the mathematical principles of category theory and graph theory. It offers a graphical notation for defining models and their interrelationships, making it an effective tool for representing complex systems and concepts. DPF enables the specification of models across various levels of abstraction and facilitates the definition of structure and constraints through the use of predicates. These predicates can be applied to a model or a segment of a model to delineate properties or conditions that the model must fulfill.

The relationship between knowledge graphs and the DPF lies in their shared use of graph-based representations to model complex systems and their interconnections. Knowledge graphs map entities and their relationships using nodes and edges, but DPF takes this a step further by introducing constraints and predicates that enforce logical consistency. Both frameworks provide hierarchical modeling and abstraction, breaking down intricate concepts into manageable layers of detail. DPF's predicates add a semantic layer similar to ontologies in knowledge graphs, setting rules and properties that validate the model. Moreover, with its roots in category theory, DPF can bring a solid mathematical foundation to the expressive nature of knowledge graphs.

In DPF, a model or specification $G = (S, C^G : \Sigma)$ comprises an underlying graph S along with a set of constraints C^G, which are defined by a predicate signature $\Sigma = (\Pi^{\Sigma}, \alpha^{\Sigma})$. A predicate signature encompasses a collection of predicates, each of which has a name and an arity (also referred to as a shape graph). A constraint involves a predicate from the signature in conjunction with the sub-graph of the model's base graph that is impacted by the constraint. The semantics of a predicate is the set of all graphs that satisfy the predicate denoted as $[\![P]\!]$, called its set of valid instances. Table 1

shows three predicates defined on the model in Fig. 1. In Sect. 3.4, we will apply these predicates to our temporal pattern analyses.

3.2 Perspective Comparison

Understanding news coverage requires more than just knowing what events occurred; it involves analyzing how different sources report on these events. Perspective comparison is a crucial aspect of news analysis because it uncovers the varied ways in which different publishers interpret, emphasize, and present information about the same events.

In this section, we present a method for perspective comparison that leverages the knowledge graph to analyze how various publishers report on events. In our proposed method, we compare the perspectives across various dimensions of these events. For instance, we examine the types of events that were reported by different publishers during a specific time period while they were covering a particular event and its subsequent development.

To effectively compare the perspectives of different publishers on the same news events, we propose leveraging category theory operations, particularly pullbacks and commutative diagrams. This approach enables a structured and formal analysis of how different sources report on the same events and helps identify commonalities and differences in their reporting. Figure 4 gives an overview of the proposed method for perspective analysis. All news article-related information is represented as a graph database, denoted as I. It contains comprehensive information about news events, including details such as event locations, event types, and involved countries and individuals.

To compare the perspectives of different publishers, we represent their individual reports as subgraphs of I. Specifically, S_1 and S_2 in the figure represent the reports from two different publishers. These subgraphs can be computed by querying the graph database using Cypher queries [1], which extract fragments of the graph that correspond to the local perspectives of the publishers. For example, if we are interested in comparing how two publishers cover the same event, we can extract subgraphs that contain the media topics and event details reported by each publisher. The objective is to analyze the extent to which the media topics used by the two publishers align or differ in their coverage of specific events.

The pullback object C in Fig. 4 which is computed from the following two morphisms: $S_1 \xrightarrow{m_1} I$ and $S_2 \xrightarrow{m_2} I$ is central to our method for perspective comparison. It captures the information about the events from the perspectives of both S_1 and S_2. From the pullback object, we can figure out the perspectives of different publishers as shown in Fig. 4 by object D_1 and D_2.

Here the proposed method is illustrated through a running example. The focus is on computing and comparing the perspectives of two publishers regarding their news stories covering events in *Niger* from July 28th to September 2nd. While the pullback object can be computed programmatically using general-purpose programming languages (e.g., Python with the Neo4j library), this paper demonstrates how a Cypher query can be employed to perform this computation. Cypher queries are utilized to extract relevant data from the Neo4j graph database. Cypher, a query language for graph databases, allows queries to be expressed as graph patterns involving variables. These queries retrieve specific subgraphs from the entire graph database that represent

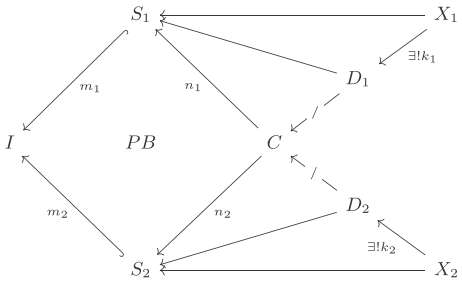


Fig. 4. Pullback object (C) computes the commonality between S_1 and S_2; D_1 and D_2 objects are used to compute the dissimilarities between S_1 and S_2 [15].

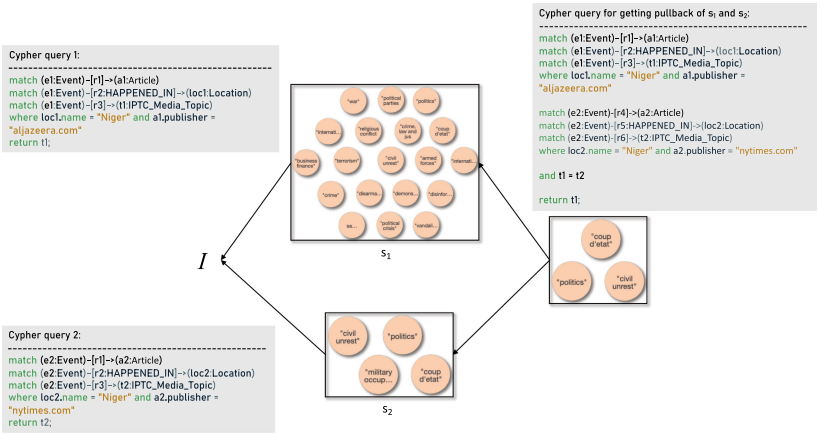


Fig. 5. Computing pullback with Cypher query [15].

the news stories covered by the two publishers. Specifically, Fig. 5 presents a Cypher query that combines two separate queries to compute the pullback object of the two subgraphs extracted from the graph database. To ensure that the diagram commutes, the condition $t1 = t2$ is specified in the query. Since the two subgraphs S_1 and S_2 include only nodes of type $IPTC\_MediaTopic$, we include $IPTC\_MediaTopic$ nodes in the result pullback object. Figure 5 shows a cypher query expression to compute the pullback object of $S_1 \rightarrow I$ and $S_2 \rightarrow I$. The perspectives of the publishers are computed from the difference of the subgraphs S_1 and S_2 with the pullback object. Here we have demonstrated the perspective analysis with respect to *IPTC* media topics but the other dimensions can also be used for perspective analysis.

3.3 Analyzing the Progression of Events

Analyzing the progression of events is a crucial task in computational journalism, essential for understanding how news stories develop and change over time. As global media coverage grows more abundant, journalists and researchers face the challenge of dis-

tilling complex and evolving narratives from vast amounts of data. The need to analyze event progression stems from the desire to uncover how events unfold, identify emerging trends, and understand the dynamics and causal relationships among different phases of an event. However, this task is complicated by the dynamic nature of news coverage and the vast amount of information from diverse sources. A major challenge in this process is managing and interpreting data from numerous publishers, each offering unique and sometimes conflicting narratives. There is a lack of tool support in computational journalism to systematically record events and analyze their progression to extract meaningful insights. We propose (1) to use features such as names, locations and *IPTC* topics to group news articles covering stories about closely related topics and, then, (2) to use category theory to analyze the progression of events by means of analyzing contents in news articles. We reuse the concept presented in Fig. 4 where we adapt S_1 and S_2 with a selection of events capturing situations from $time_{x1} - time_{y1}$ and $time_{x2} - time_{y2}$ respectively. From S_1 and S_2 we systematically compare the evolution of events from $time_{x1} - time_{y1}$ to $time_{x2} - time_{y2}$. For example, S_1 and S_2 may represent the *IPTC* media topics being used to cover the news events about *Niger* from July 31 to August 6 and from August 7 to August 13, respectively. From these subgraphs, we compute the emerging *IPTC* media topics in the reports published during August 7 to August 13. This comparative analysis allows journalists to get an overview of the progression of events.

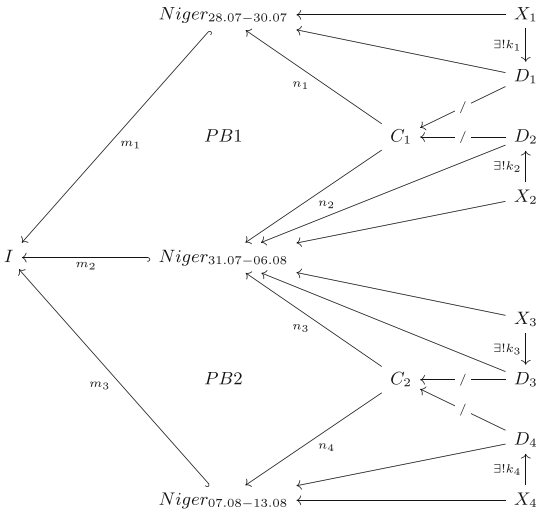


Fig. 6. Capturing the progression of events with pullback operation [15].

The progression of events can be represented as a transformation of *IPTC* media topics being covered by the publishers. Let us consider that in Fig. 6, $Niger_{28.07-30.07}$, $Niger_{31.07-06.08}$ and $Niger_{07.08-13.08}$ are representing the *IPTC* media topics being used to cover the news events in $Niger$ for periods July 28 to July 30, July 31 to August

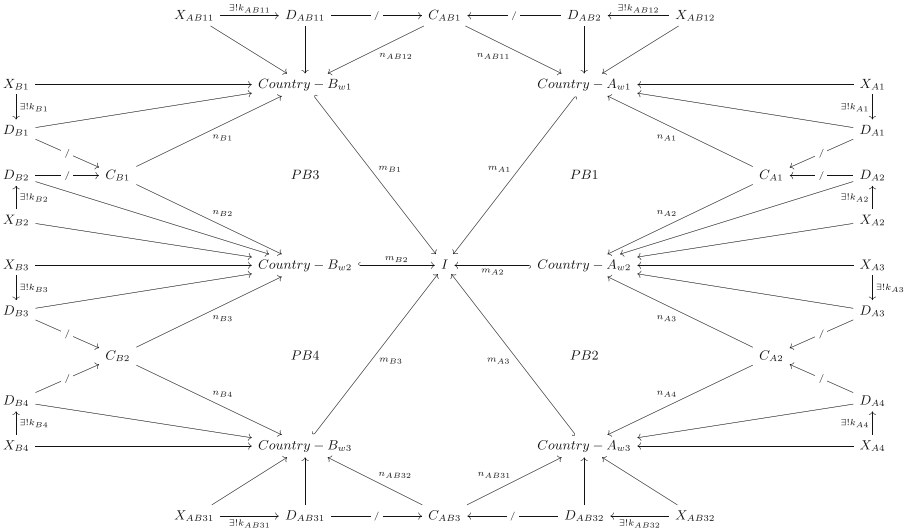


Fig. 7. Comparison of progression of events [15].

6 and August 7 to August 13 respectively. The pullback object C_1 and C_2 represents the commonality of the events (with respect to *IPTC* media topics) in $Niger_{28.07-30.07}$, $Niger_{31.07-06.08}$ and $Niger_{31.07-06.08}$, $Niger_{07.08-13.08}$ respectively. The object D_1 would capture the media topics being removed from the reporting during July 31 to August 6; D_2 would capture the media topics being newly added during July 31 to August 6. Similarly, D_3 would capture the media topics being removed from the reporting during August 7 to August 13 and D_4 would capture the media topics being added during August 7 to August 13.

Similar categorical operations can be employed to analyze the progression of events across two different countries. For instance, consider the task of analyzing the weekly progression of events in *Niger* and *Gabon* since the start of coups in these two countries. Figure 7 illustrates a computational model for such analysis. The pullback object C_{AB1} captures the commonality in the progression of events between the two countries $Country - A$ and $Country - B$, where $Country - A_{w1}$ and $Country - B_{w1}$ represent contextual information of events (such as *IPTC* media topics or involved countries or individuals) reported in the first week. For brevity we did not show C_{AB2} (pullback object between $Country - A_{w2}$ and $Country - B_{w2}$) in the diagram. By examining the pullback objects C_{AB1}, C_{AB2}, C_{AB3}, and so forth, common patterns in the event progression between the two countries can be identified.

Figure 8 illustrates a computation model for the comparison of progression of events at a higher level of abstraction. $\alpha_1, \alpha_2, \beta_1, \beta_2$ represents contextual information of events specified at a certain abstraction level j; In our running example we only have a hierarchical data model for *IPTC* Media topics, therefore, all the *IPTC* Media topics in $\alpha_1, \alpha_2, \beta_1, \beta_2$ are at level j in the *IPTC* Media topic ontology. $\alpha_1', \alpha_2', \beta 1', \beta 2'$ represents contextual information of events specified at a higher level of abstraction. The

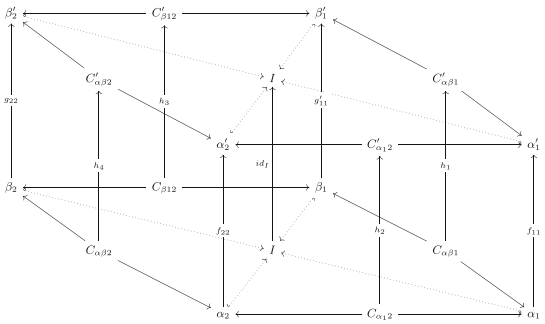


Fig. 8. Comparison of progression of events at a higher level of abstraction [15].

pullback objects $C_{\alpha\beta i}$ (where $i = 1, 2$) in the bottom layer represent the commonality of the progression of events. The arrows between layers represent graph homomorphisms between corresponding elements from lower to higher levels of abstraction in the knowledge graph I.

Theorem: For any non-empty pullback object $C_{\alpha\beta i}$ (where $i = 1, 2$) at level j, the corresponding pullback objects $C'_{\alpha\beta i}$ at level $k < j$ is non-empty.

Proof Sketch: Consider a non-empty pullback object $C_{\alpha\beta i}$ (where $i = 1, 2$) at level j; this would require at least one element $n_a \in \alpha_i$ and one element $n_b \in \beta_i$ where n_a and n_b are mapped to the same element in the knowledge graph. If n'_a (with level k) is a parent of n_a, and n'_b (with level k) is a parent of n_b, then n'_a and n'_b must also map to the same element in the knowledge graph. The pullback objects $C'_{\alpha\beta i}$ should at least contain an element that maps to n'_a and n'_b and therefore cannot be empty.

3.4 Temporal Analysis

The temporal aspect of events in journalism is crucial as events are continuously evolving. Journalistic events are often interconnected, with causal relationships influencing their progression. For instance, a *political crisis* can increase the likelihood of subsequent *armed conflict*. Additionally, events can evolve over time, involving new entities such as international organizations or humanitarian groups in ongoing conflicts or crises. Therefore, capturing the temporal dynamics of events provides valuable insights for journalists.

Linear Temporal Logic (LTL) offers a framework to capture these temporal aspects. We propose enhancing LTL in two ways. First, we incorporate the concept of duration, building on our previous work [7]. Second, we integrate the Diagram Predicate Framework (DPF) and category theory into our LTL definition, thereby enabling a richer representation of the temporal and structural properties of events.

Linear temporal logic (LTL) [10] is a modal temporal logic with modalities referring to (temporal) order of events. We propose LTL_{DGP} (Dynamic Graph Patterns) that models the temporal aspects of the events in our news framework incorporating

category theoretical concepts. A well-formed LTL_{DGP} formula with time, ϕ, is therefore recursively defined by the BNF formula below:

$$\begin{aligned}\phi &:= \top \mid \bot \mid U \mid \neg\phi \mid \phi \wedge \phi \mid \phi \vee \phi \mid \\ &\quad \bigcirc_{(TimeInterval)}\ \phi \mid \Diamond_{(TimeInterval)}\ \phi \mid \\ &\quad \Box_{(TimeInterval)}\ \phi \mid \phi \cup_{(TimeInterval)}\ \phi \\ U &:= P.n = Atom\end{aligned}$$

where P denotes an arbitrary predicate listed in Table 1. Table 2 illustrates how these predicates are mapped to the underlying model. We borrow the notation $P.n$ from the object oriented programming in order to refer to a node n from the extracted portion of a graph based on the predicate P.

$$\begin{aligned}TimeInterval &:= < Time \mid \leq Time \mid \\ &\quad > Time \mid \geq Time \mid = Time \\ Time &:= INT\ second \mid INT\ minute \mid INT\ hour \mid INT\ days \mid \\ &\quad INT\ week \mid INT\ month \mid INT\ year \mid Time\ and\ Time\end{aligned}$$

We also define the time difference between event instances:

$$\begin{aligned}&diff_{Time}\big((S_i, m_i), (S_{i+1}, m_{i+1})\big) = \\ &InstanceTime(S_{i+1}, m_{i+1}) - InstanceTime((S_i, m_i)\end{aligned} \tag{3}$$

where $InstanceTime(x)$ is a function that returns the time associated with event instance x.

Definition 1. *Given a formula ϕ, and an arbitrary path $\pi_E = \big((S_1, m_1), (S_2, m_2), (S_3, m_3)\dots\big)$ π encompasses a sequence of all reports on an arbitrary event E. Each S_i along with its corresponding graph morphism m_i, represents an instance of the schema shown in Fig. 1 at time i. The satisfaction relation $\models$ is defined as follows:*

- $\pi \models \top$
- $\pi \not\models \bot$
- $\pi \models U$ *iff for* $P \in \Pi^{\Sigma_1}$, $m_1^* \in [\![P]\!]$ *the following diagram commutes, and* $O_1^*.n = Atom$*:*

$$\begin{array}{ccc} P & \xrightarrow{\ \delta\ } & I \\ {\scriptstyle m_1^*}\uparrow & & \uparrow{\scriptstyle m} \\ O_1^* & \xrightarrow[\ \delta^*\ ]{} & S_1 \end{array}$$

- $\pi \models \neg\phi$ *iff* $\pi \not\models \phi$
- $\pi \models \phi_1 \wedge \phi_2$ *iff* $\pi \models \phi_1$ *and* $\pi \models \phi_2$
- $\pi \models \phi_1 \vee \phi_2$ *iff* $\pi \models \phi_1$ *or* $\pi \models \phi_2$
- $\pi \models \phi_1 \rightarrow \phi_2$ *iff* $\pi \models \phi_2$ *whenever* $\pi \models \phi_1$

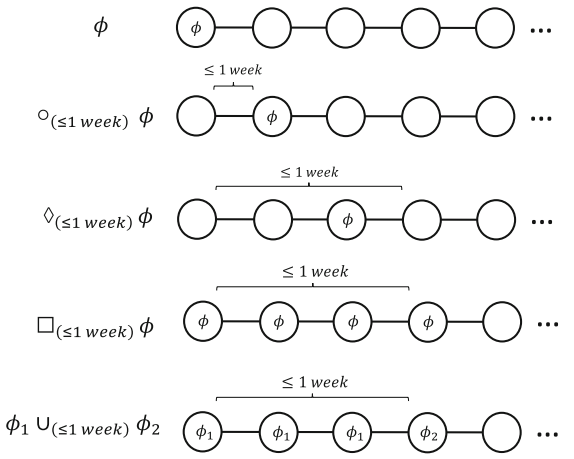


Fig. 9. Example of paths with events and time intervals [7].

- $\pi \models \bigcirc_{(TimeInterval)} \phi$ *iff* $\pi^2 \models \phi$ *and* $diff_{Time}(S_1, S_2)$ *complies with* $TimeInterval$
- $\pi \models \Box_{(TimeInterval)} \phi$ *iff, for all* $i \geq 1, \pi^i \models \phi$ *and* $diff_{Time}(S_1, S_i)$ *complies with* $TimeInterval$
- $\pi \models \Diamond_{(TimeInterval)} \phi$ *iff, there is some* $i \geq 1$ *such that for* $j \leq i$, $\pi^j \models \phi$, *and* $diff_{Time}(S_1, S_i)$ *complies with* $TimeInterval$
- $\pi \models \phi_1 \cup_{(TimeInterval)} \phi_2$ *iff, there is some* $i \geq 1$ *such that* $\pi^i \models \phi_2$ *and for all* $j = 1, 2, ...i - 1$ *we have* $\pi^j \models \phi_1$ *and* $diff_{Time}(S_1, S_i)$ *complies with* $TimeInterval$.

Figure 9 visualizes the semantics of some operations of our proposed temporal logic. The enhanced expressiveness of LTL_{DGP} empowers us to formulate a diverse range of formulas, allowing for the identification of patterns of interest within the event paths. In the following few example are presented.

– "Eventually, in less than 4 d, there will be an event where an *armed conflict* (*IPTC*=20000056) occurs involving *protesters*." In simpler terms, this formula asserts that within the time interval of 4 days or more, there should be an occurrence of an event classified under the *IPTC* media topic "armed conflict" and involving "protesters". Figure 10a shows an event-set that is satisfied by this formula. The gray area represents the 4 day time interval.

$$\Diamond_{(\geq 4 \text{ days})}((\texttt{[IPTC]}.IPTC\_Topic = \text{"20000056"}) \wedge (\texttt{[Person]}.Person = \text{"}Protesters\text{"})) \tag{4}$$

– "Always, in the next 8-day interval, if *political dissent* (*IPTC*=20000648) occurs at a specific point of time, then *political process* (*IPTC*=20000649) must occur." In other words, this formula asserts that within an interval of 8 days , if there is

an event classified under the *IPTC* media topic 20000648, it must be followed (or accompanied) by an event classified under the *IPTC* media topic (*IPTC*=20000649). Figure 10b shows an event-set that is satisfied by this formula.

$$\begin{aligned}\Box_{(\geq\ 8\ \text{days})}((&[\texttt{IPTC}].IPTC\_Topic = \text{"20000648"}) \\ &\rightarrow ([\texttt{IPTC}].IPTC\_Topic = \text{"20000649"}))\end{aligned} \tag{5}$$

– "If at any time we encounter a *political crisis* (*IPTC*=20000647), then it should eventually be followed by an *armed conflict* within a duration of six days." The formula is a conjunction of 2 terms, The first term in the conjunction ensures that at some point in the future, a *political crisis* must be reported. This term uses the diamond operator to assert the eventual occurrence of the event, making it a necessary condition for the subsequent implications to be meaningful. This prevents the formula from holding vacuously by ensuring that the condition of a *political crisis* being reported is an unavoidable event. The second term in the conjunction specifies the temporal relationship that must hold once a *political crisis* occurs. The box operator indicates that the enclosed condition must always hold true for every instance of time following the initial event. Specifically, it states that whenever a *political crisis* is reported, it must always be followed by an *armed conflict* within the next six days. The gray area represents the 4 day time interval.

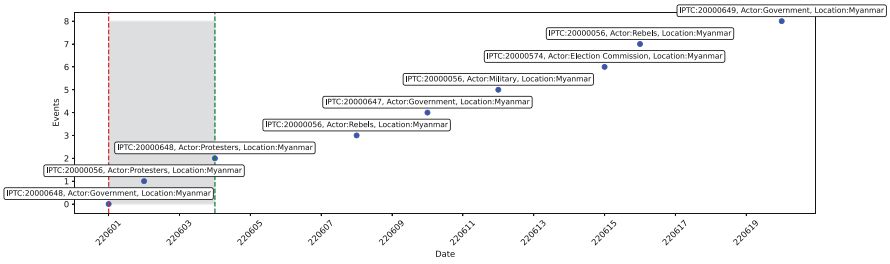
$$\begin{aligned}&(\Diamond_{(\geq\ 0\ \text{seconds})}([\texttt{IPTC}].IPTC\_Topic = \text{"20000647"})) \wedge \\ &(\Box_{(\geq\ 0\ \text{seconds})}([\texttt{IPTC}].IPTC\_Topic = \text{"20000647"} \rightarrow \\ &\qquad \Diamond_{(\leq\ 6\ \text{days})}([\texttt{IPTC}].IPTC\_Topic = \text{"20000056"})))\end{aligned} \tag{6}$$

– "If at any point in the future a *political crisis* is reported in *Myanmar*, then within the following 6 days, there must be a report involving *Government*." This formula asserts that if there is an event located in Myanmar and it is classified under the *IPTC* media topic 20000647, then within the next 6 days, there must be an event involving the government.

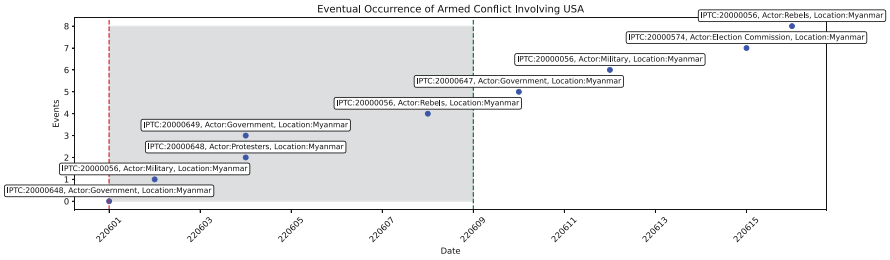
$$\begin{aligned}\Box(\geq\ 0\ \text{seconds})((&[\texttt{Location}].Location = \text{"}Myanmar\text{"} \wedge \\ &[\texttt{IPTC}].IPTC\_MediaTopics = \text{"20000647"}) \rightarrow \\ &\Diamond_{(\leq\ 6\ days)}([\texttt{Person}].Person = \text{"}Government\text{"}))\end{aligned} \tag{7}$$

3.5 Variant Analysis

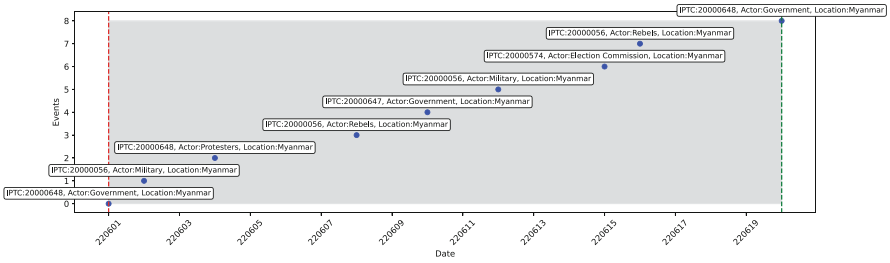
In this section, we introduce a technique for variant analysis based on the computation results from Sect. 3. We apply statistical methods to detect variations in news articles. Section 3.3 detailed techniques for retrieving data from a knowledge graph across various dimensions and abstraction levels. We use this data selection process to identify variants by applying statistical methods. Here, we present *Exploratory Data Analysis* (EDA) for identifying trends in time and space, which serves as the foundation for our variant analysis.



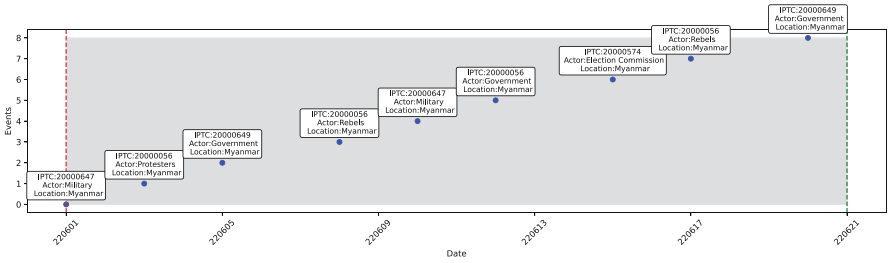
(a) A sample event path satisfied by Pattern 4.



(b) A sample event path satisfied by Pattern 5.



(c) A sample event path satisfied by Pattern 6.



(d) A sample event path satisfied by Pattern 6.

Fig. 10. Timeline Visualization of Event Sequences Satisfying LTL_{DGP} Formulas.

To identify trends in reporting across different topics, a dimension and an abstraction level are first selected, and relevant data is then extracted from the knowledge graph. For instance, to analyze trends in the reporting of *civil unrest* in *Niger* from

August 1 to August 20, 2023, events corresponding to the *civil unrest* topic in the *IPTC* Media topic ontology are retrieved from the knowledge graph. The extracted data is subsequently used for statistical analysis, such as frequency distribution, and for visualizing trends over time using a timeline. Through the visualization of these events on a timeline, patterns of reporting by different publishers are illustrated, and the reporting activity throughout the specified period is tracked.

Figure 11 highlights the duration of engagement of individual publishers (*aljazeera.com*, *theguardian.com*, *reuters.com*, *independent.co.uk*, *nytimes.com*, *washingtontimes.com*, and *cnn.com*) in reporting about *civil unrest* in *Niger*. In the figure, the background represents the co-limit, which is a categorical representation of the union of all events reported by these publishers on the topic of *civil unrest* in *Niger*. From this figure, various aspects of the news coverage can be analyzed, such as the identification of similarities between publishers. For example, it may be observed that *Independent.co.uk* and *WashingtonTimes.com* display similar reporting patterns during the period of *civil unrest* in *Niger*. Such similarities may reveal insights into the approaches of different publishers to the same topic. Beyond the comparison of reporting patterns for a specific topic, the proposed method can be adapted to explore other dimensions of the news dataset. For instance, trends related to the involvement of specific countries in conflicts can be identified by retrieving and analyzing data on different types of conflicts involving various countries. We leverage ontological hierarchies to ensure that we extract events at the appropriate level of abstraction. For example, using the *IPTC* Media topic ontology, we can gather data on coups in African nations and identify common trends in how foreign countries are involved in these events.

4 Discussion and Future Work

In this paper, we have introduced a model-based framework for content analysis in computational journalism. By leveraging knowledge graphs and category theory, our approach enables detailed comparative and temporal analysis of news content across various dimensions and abstraction levels.

In the landscape of news content analysis, various systems such as GDELT [12] have been developed for identifying and organizing news events from vast data streams in structured formats. While GDELT efficiently aggregates and quantitatively analyzes vast volumes of news data, a new approach is needed to enable researchers to dive deeper into individual news events, one which also holds the potential to promote transparency and accountability in news analysis in order to foster more responsible journalism practices.

Our framework for content analysis introduces a novel approach that goes beyond traditional text mining and semantic technologies [11], [18]. In this paper, our primary focus has been on the analysis of various reports pertaining to a specific event, particularly in terms of perspectives. One can furthermore include the intricacies of opinions, reporting angles, tones, and the framing of articles, enriching our understanding of news narratives. We also presented a logical framework that leverages Linear Temporal Logic (LTL) in the context of knowledge graphs to capture and analyze temporal patterns within news storylines. By formulating formulas, we can represent and query

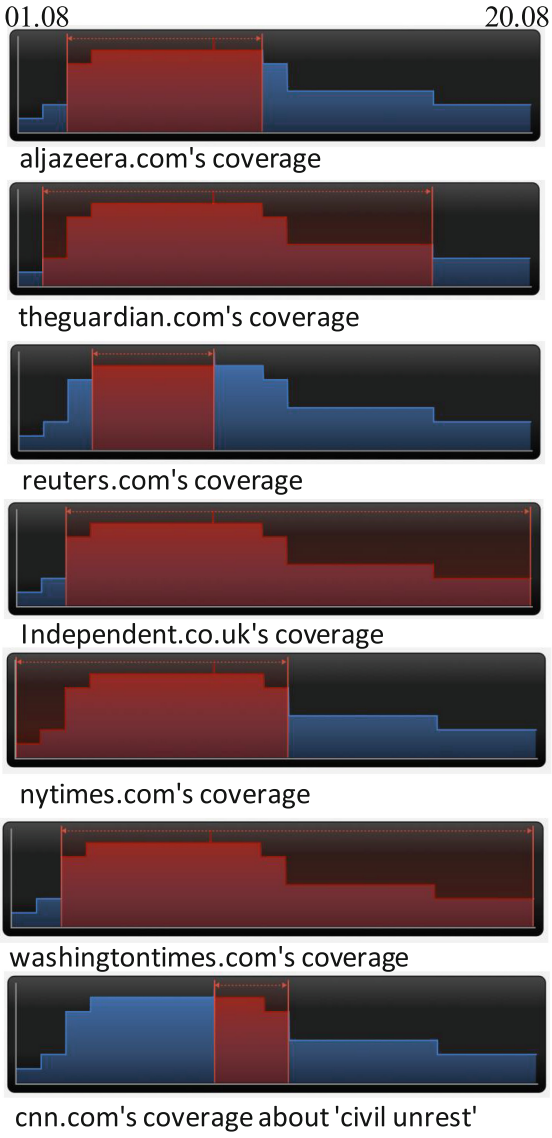


Fig. 11. Timeframe showing the engagement of news publishers in reporting about *civil unrest* in *Niger* [15].

various temporal aspects of event progression. Additionally, we have employed a systematic approach to track the evolution and progression of these events over time which provides insights into how events unfold and transform over time.

While we have presented an analysis technique using category theory, there is much more to explore and develop in this field. We believe that the integration of generative

AI and category theory can contribute to the evolution of journalism in the digital age, fostering transparency, accountability, and enriched news content for both journalists and readers. Particularly, our approach has the capacity to assist in tasks that involve the comparison of news items. For instance, it can be particularly useful in multilingual news comparison, where it can facilitate cross-cultural analysis of news events by overcoming language barriers. It can also be utilized in fact checking and verification, aiding in the assessment of news source credibility. Additionally, it is well-suited for bias and framing analysis, enabling the exploration of different perspectives presented in the media. By addressing these future directions, we hope to enhance the capabilities of the news analysis frameworks and contribute to the advancement of media studies.

Acknowledgement. This research is funded by SFI MediaFutures partners and the Research Council of Norway (grant number 309339).

References

1. Cypher query language (2023). https://neo4j.com/developer/cypher/. Accessed 25 Sept 2023
2. Ethical journalism network (2023). https://ethicaljournalismnetwork.org/who-we-are. Accessed 26 Sept 2023
3. GDELT (2023). https://www.gdeltproject.org/data.html. Accessed 12 Sept 2023
4. Barr, M., Wells, C.: Category Theory for Computing Science. Prentice-Hall Inc., Upper Saddle River (1990)
5. Berven, A., Christensen, O.A., Moldeklev, S., Opdahl, A.L., Villanger, K.J.: A knowledge-graph platform for newsrooms. Comput. Ind. **123**, 103321 (2020). https://doi.org/10.1016/j.compind.2020.103321. https://www.sciencedirect.com/science/article/pii/S0166361520305558
6. Ehrig, H., Ehrig, K., Prange, U., Taentzer, G.: Fundamentals of Algebraic Graph Transformation. Monographs in Theoretical Computer Science. An EATCS Series, Springer (2006)
7. Fatemi, B., Rabbi, F., MacCaull, W.: A validated learning approach to healthcare process analysis through contextual and temporal filtering, pp. 108–137. Springer, Heidelberg (2024). https://doi.org/10.1007/978-3-662-68191-6_5
8. Fatemi, B., Rabbi, F., Opdahl, A.L.: Evaluating the effectiveness of gpt large language model for news classification in the iptc news ontology. IEEE Access **11**, 145386–145394 (2023). https://doi.org/10.1109/ACCESS.2023.3345414
9. Fatemi, B., Rabbi, F., Tessem, B.: Fairness in automated data journalism systems. NIKT: Norsk IKT-konferanse for forskning og utdanning (2023). https://doi.org/10.13140/RG.2.2.30374.19522. https://www.researchgate.net/publication/365127564_Fairness_in_automated_data_journalism_systems
10. Gabbay, D.M., Hodkinson, I., Reynolds, M.A.: Temporal logic: mathematical foundations and computational aspects (1994)
11. Leban, G., Fortuna, B., Brank, J., Grobelnik, M.: Event registry: learning about world events from news. In: Proceedings of the 23rd International Conference on World Wide Web, pp. 107–110. ACM (2014). https://doi.org/10.1145/2567948.2577024
12. Leetaru, K., Schrodt, P.A.: Gdelt: global data on events, location, and tone, 1979–2012. In: ISA Annual Convention, vol. 2, pp. 1–49. Citeseer (2013)
13. Liu, D., Bai, T., Lian, J., Zhao, X., Sun, G., Wen, J.R., Xie, X.: News graph: an enhanced knowledge graph for news recommendation. In: KaRS@ CIKM, pp. 1–7 (2019)

14. Opdahl, A.L., Al-Moslmi, T., Dang-Nguyen, D.T., Gallofré Ocaña, M., Tessem, B., Veres, C.: Semantic knowledge graphs for the news: a review. ACM Comput. Surv. **55**(7), 1–38 (2022)
15. Rabbi, F., Fatemi, B., Lamo, Y., Opdahl, A.L.: A model-based framework for news content analysis. In: MODELSWARD, pp. 99–107 (2024)
16. Rospocher, M., et al.: Building event-centric knowledge graphs from news. J. Web Semant. **37**, 132–151 (2016)
17. Rossini, A.: Diagram predicate framework meets model versioning and deep metamodelling (2011)
18. Rudnik, C., Ehrhart, T., Ferret, O., Teyssou, D., Troncy, R., Tannier, X.: Searching news articles using an event knowledge graph leveraged by wikidata. In: Companion Proceedings of the 2019 world Wide Web Conference, pp. 1232–1239 (2019)
19. Rutle, A.: Diagram predicate framework: a formal approach to mde (2010)
20. Savelka, J., Ashley, K.D.: The unreasonable effectiveness of large language models in zero-shot semantic annotation of legal texts. Front. Artif. Intell. **6** (2023)
21. Schudson, M.: Journalism: Why it Matters. Polity Press, London (2020)