IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2023.0322000

Event Graph-Based News Clustering: The Role of Named Entity-Centered Subgraphs

BASAK BULUZ KÖMEÇOGLU[1], AND BURCU YILMAZ[1]
[1]Institute of Information Technologies, Gebze Technical University, Gebze, Kocaeli, 41400, Turkey
Corresponding author: Basak Buluz Komecoglu (e-mail: bbuluz@gtu.edu.tr).

ABSTRACT In an era of exponential growth in online news sources, the need for intelligent digital solutions capable of efficiently analyzing and organizing large amounts of news content has become crucial. This paper presents a graph-based methodology designed to enhance Topic Detection and Tracking (TDT) tasks in natural language processing by efficiently clustering news events into coherent stories. The proposed approach leverages a novel event graph model that captures not only the characteristics of individual news events but also their collective narrative context. Using Named Entity Centred Frequent Subgraphs, the model excels in identifying recurring patterns of events and thus provides a framework for learning a robust, language-independent, and structured representation for structuring news stories, which represents a significant advance in the refinement of traditional clustering algorithms. Empirical experiments using a multilingual benchmark dataset, the News Clustering Dataset, highlight the superior clustering performance of our approach compared to state-of-the-art monolingual document clustering techniques, particularly in English and the competitive results in Spanish. To underline the adaptability of the methodology to low-resource languages, the Turkish 'Story-Based News Dataset' developed specifically for this study also promises to serve as an important resource for a wide range of natural language processing tasks.

INDEX TERMS Frequent subgraph mining,low-resource language, natural language processing, text clustering.

I. INTRODUCTION

THE term "news" in the English language, as the plural form of "new," refers to the presentation of new information through various communication channels, including written, visual, or auditory mediums [1]. News texts delivered to readers via online or offline platforms are designed to address fundamental journalistic questions (who, what, when, where, and why) in the first few sentences, with the aim of quickly conveying information about the main event. These responses depict the characteristics of the main event and provide a description [2], [11]. Unlike other scientific disciplines, it is not necessary to discuss causal relationships with sub-events related to the ongoing periods of the event under consideration. Even if a news text portrays a long process, the event it covers is typically defined as the present or recent past [3]. Consequently, when dealing with news about an ongoing event, readers often face challenges in linking current news to previous articles, which hampers the analysis of event development.

The exponential increase in the number of news sources in the digital age has led to an increasing demand for intelligent news aggregation systems capable of tracking events in news content and transforming them into stories. These systems, which aim to collect various headlines from reputable global and/or national news sources, group news texts related to the same event, and provide insights based on features extracted from the data, require solutions that focus on Topic Detection and Tracking (TDT) tasks in natural language processing research [4], [5]. The process carried out in research focusing on this task traditionally begins with event detection and ultimately aims to generate a storyline [6], [7]. Although many different approaches have been proposed in the literature, clustering solutions are commonly used after news texts are represented as vectors in multidimensional space [6], [8], [9].

Current news aggregation systems adopt search-based event-tracking approaches. These approaches fail to effectively combine the discrete fragments of a story that occur at different times, organize them into a coherent timeline, decompose conflicts, and ultimately establish a coherent narrative structure consistent with the cognitive processes of the human brain [10]. However, as in almost all natural language processing problems, most approaches focus on English [30]. The developed multilingual approaches, on the other hand, are highly dependent on language-specific features, exhibit

poorer performance compared to approaches focusing on a single language, and are difficult to extend to low-resource languages such as Turkish [13], [30]. Based on these limitations, we propose a graph model capable of representing the features of the main event in a news text regardless of language, as well as capturing the overall story context by combining common patterns that often occur in similar events. Utilizing this proposed graph model, we focus on the task of clustering events in news content based on the story chains to which they belong without detaching them from their context.

News text is represented by an unweighted directed graph model that combines words in the content and identifies elements that answer basic journalistic questions, such as named entities, timestamps, and publication dates. Events that form the basic parts of a story are expected to have similar recurring patterns and common elements, such as person/organization, location, and time. Therefore, frequent subgraphs based on named entities were identified, and event graphs were extended to include these graphs. In this way, the ability of the graph embedding representing the events to represent the story to which it belongs with a holistic approach has been increased, while on the other hand, its separability from the events in different stories has been improved. Each news text modelled with the graph structure was vectorized by averaging the Node2Vec node embeddings, and the event clusters were successfully obtained using traditional clustering algorithms. To evaluate the impact of the proposed event-graph model on the clustering performance, a multilingual benchmark dataset for news clustering [16] was used together with a dataset curated by researchers for the resource-limited Turkish language. This study introduces several novel contributions to the field, which can be summarized as follows:

- We propose an event graph model that goes beyond representing the features of an event in a news text, and is able to capture the overall context of the story by combining common patterns that often occur in similar events. The proposed graph model allows news texts to be structurally represented independently of the language.
- We propose an approach for the discovery of Named Entity-Centered Frequent Subgraphs. This approach improves the performance of traditional clustering algorithms by identifying recurring patterns that involve named entities.
- We introduce a story-based news clustering dataset created in Turkish, a low-resource language. This dataset can be used for a wide variety of tasks in natural language processing, particularly text clustering. This provides a valuable resource for researchers and practitioners in this field.
- We compare our approach with state-of-the-art monolingual document clustering approaches and show results with higher clustering performance for the News Clustering dataset in English and Spanish, and competitive results for the German language.

These contributions collectively contribute to the field of TDT by providing innovative approaches, valuable datasets, and enhanced modeling techniques for understanding and clustering news events based on their storyline and contextual patterns.

The rest of this paper is structured as follows: Section 2 presents a summary of related work in the literature. Section 3 defines the problem, technical concepts, and the methodology applied. Information regarding the experimental environment, presentation of the experimental results, analyses, and discussions of the methodological approaches are presented in Section 4. Finally, Section 5 provides conclusions, including a summary and key findings.

II. RELATED WORKS

The primary goal of the Topic Detection and Tracking (TDT) task, first described in the literature by Allan et al. in 1998 [5], is to organize a set of news articles or news streams into groups of events called "stories". The document-clustering problem requires unsupervised methods that aim to identify clusters based on similarities between items in a given collection of documents. These methods attempt to represent similar documents by forming dense clusters in a semantic vector space, whereas dissimilar documents tend to be located further apart from each other. In this respect, the representation method used in news texts conveying event content plays an important role in selecting an appropriate clustering method.

In text-based topic detection and tracking, the most widely used representation model is the Vector Space Model (VSM), which uses words in a document and their corresponding term weights [17]–[19]. In vector space models, only the term weights are considered important, without considering the order of words, their interrelationships, or semantic information [17]. Consequently, TDT approaches using VSM are referred to as term-based event-detection techniques. Based on a review of the studies [4], [12], [18] in which these methods have been applied, the approach can be broadly described as follows: First, word and keyword extraction is performed on the documents through a series of preprocessing steps. Then, the documents are subjected to statistics-based vector space models, such as Term Frequency-Inverse Document Frequency (TF-IDF) [22], [23], Named Entity Recognition (NER) [20], [24], and Bag of Words [15], which are frequently used in the TDT task, and the task is completed by means of a clustering algorithm. These approaches, which were very popular in the past, have taken a back seat with the emergence of deep learning-based methods with high performance in almost all text representations and natural language processing tasks [14], [17].

Different levels of embedding techniques, such as words, sentences, and document embeddings, ensure that semantically similar items are closely located in vector space [27]. Several studies [14], [28], [29] have shown that state-of-the-art word, sentence, and document embedding techniques

perform better than VSM models in TDT tasks. They can also avoid the curse of sparsity and dimensionality problems in text representation [17], while the discovery of multilingual representation methods has been noted as a significant improvement over representation methods that depend on language-specific features [13], [30], [31]. Another innovative perspective is to represent events in news documents in a predetermined structural form and to apply clustering approaches specific to this structural form. Graphs, which have a high ability to represent the complex relationships of elements, are ideal for event representation, and event clusters can be formed directly using attribute or structure-based graph clustering approaches [21], [32]. In attribute-based graph clustering, the main challenge is to utilize node context information and capture the interrelationships of structural elements. State-of-the-art deep learning-based graphical representation methods have been developed to solve this problem. They hold great promise for TDT tasks as they preserve the temporal, spatial, and nonlinear relationships between the temporal, spatial, and actor components of events within complex graph structures [33], [34]. Graph embeddings are simple, can be used as direct inputs to traditional clustering algorithms, and achieve high-level results [35], [36].

In recent years, several approaches have emerged that specifically aim to learn node or graph embeddings while preserving important properties of the underlying graph. These approaches include matrix factorization-based methods [37], [38], skip gram-based methods [39], [40], and deep neural network-based methods [25], [26] found in the literature.

In a graph topology represented by an adjacency matrix, a row or column vector of size N (where N is the number of nodes) is typically used to represent each node in the graph. Matrix factorization techniques applied to adjacency matrices have been widely used to find a low-dimensional space to represent a graph [41], [42]. The Word2Vec approach [27], which has made significant contributions to natural language processing, has also inspired graph-embedding techniques. The fundamental concept of Word2Vec is to create a vector representation for a word based on the vectors of its neighboring words, determined by their co-occurrence patterns. Similarly, in graph embedding, the co-occurrence patterns of random walk models and node neighborhoods are computed by treating each node in the graph as a word in a sentence [43]. DeepWalk [40] and Node2Vec [39] are popular skip gram-based methods. These methods leverage deep neural network models, which are well known for providing end-to-end solutions to complex problems, including graph-embedding tasks.

Studies focusing on obtaining event representations through graph embedding methods have primarily explored graph models or encoders capable of representing individual event features and capturing event network neighborhoods while preserving graph topology and node semantics. One study leveraging the widely used Node2Vec node-embedding approach on a network generated from event features, another approach maps each news feature to a vector, and represents the event as the sum of these features [44]. Zeng et al. proposed a novel framework that encodes an event network using a graph encoder at the corpus level, thereby providing a general context for event representation [45]. In addition, in an innovative study, researchers proposed a dynamic word graph designed to detect and track significant events [46].

Typically, journalists present newsworthy events to the public in the form of textual documents. Therefore, the TDT task has been addressed as a document-clustering problem in many studies. Traditional clustering algorithms are widely applied to different types of dense vectors representing news documents, and their advantages and disadvantages have been previously discussed. Among these algorithms, k-means, hierarchical clustering, and DBSCAN are the most frequently used approaches in the literature [7], [31], [47]. Most traditional clustering algorithms initially assumed that the number of clusters ('k') is a mandatory parameter. However, it is difficult to determine the number of clusters for real-time news systems or systems operating in large batches of news.

In this study, a new event graph model is proposed for event representation, which differs from the previous literature. This model covers all components of an event and captures the common patterns shared by other events. This approach ensures that nodes representing common patterns that are frequently observed across multiple events have a weighted influence on event representations. The proposed graph model is represented using the state-of-the-art Node2Vec node embedding approach to evaluate its contribution to the performance of traditional clustering algorithms, BIRCH, which does not require the 'k' cluster number parameter, and HDBSCAN clustering algorithms, which combine the advantages of hierarchical clustering and density-based DBSCAN clustering algorithms. Thus, the algorithmic complexity of structure-based graph clustering algorithms is avoided and on the other hand, approaches that do not require the 'k' parameter are utilized.

III. METHODOLOGY

The story-based event clustering problem can be defined as follows: Given a set of graphs $n_1, n_2, \ldots, n_n \epsilon\ N$ representing a specific event(news) , our objective is to generate subsets $s_1, s_2, \ldots, s_m \epsilon\ S$ from these graphs, where each subset contains graphs that belong to the same story. The challenge lies in computing the event-graph model and its embedding with high decomposition and representativeness, thereby enabling accurate detection of these subsets. Figure 1 illustrates the overall framework of the five-step methodology suggested in this study to overcome this problem, the main steps of which are as follows: (1) Extracting Graph Nodes from News Text, (2) Discovering Named Entity-Centered Frequent Subgraphs, (3) Constructing an Event Graph Model, (4) Event Graph Embedding Learning with Node2Vec and (5) Story-based Event Clustering.

Definition 1 (Event). An event refers to a primary action or phenomenon that is the subject of the news article. Each news text was considered as an event in the context of this study. Each event is represented by a graph, denoted as

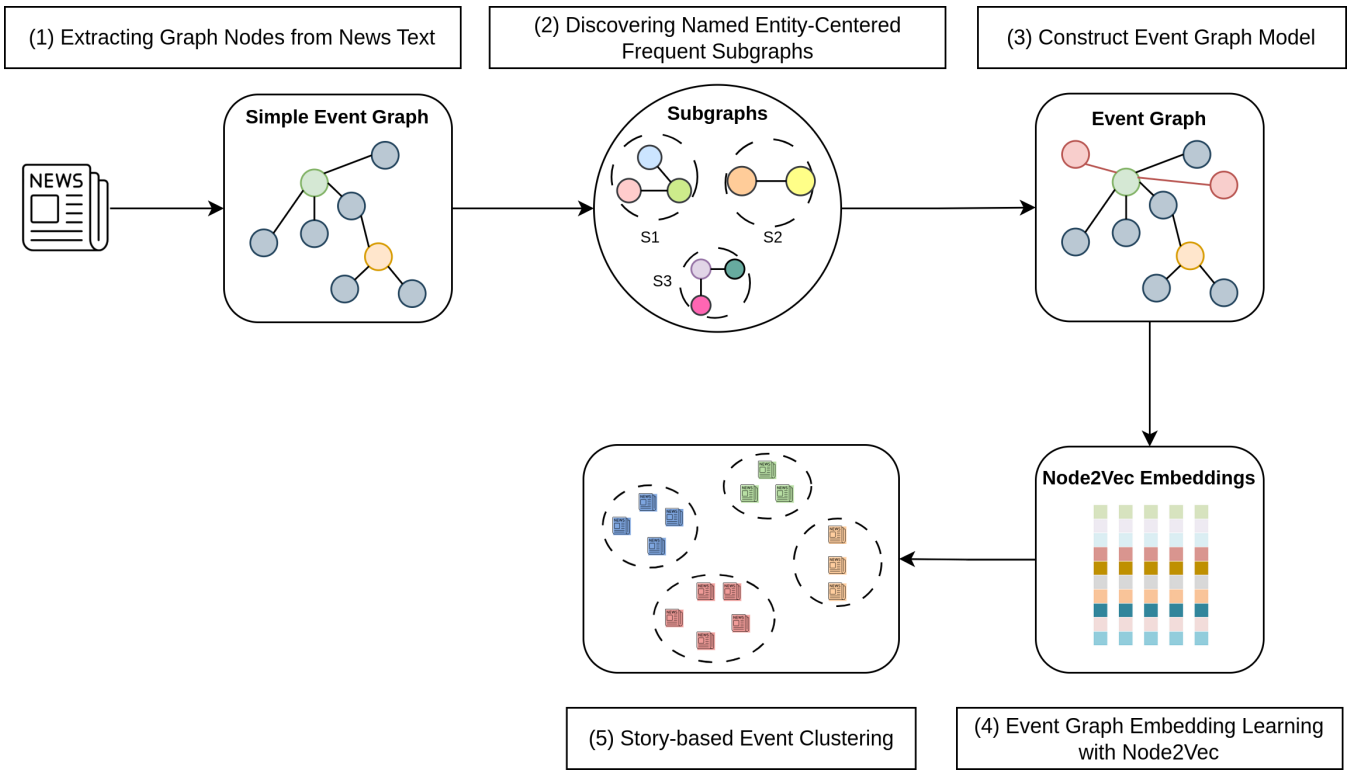

FIGURE 1. Overall structure of Event Graph-Based News Clustering Approach: (1) Extracting the information required for the Simple Event Graph, which is accepted as the baseline, from the unstructured news text and constructing the graph. (2) Creating paths between the named entities with the 'NEXT ENTITY' edges added between the named entities in the Simple Event Graph and thus obtaining Named-Entity Centred Subgraphs using gSpan. (3) Constructing the Event Graph Model by adding the relevant nodes and edges. (4) Learning the embedding of the Event Graph using Node2Vec. (5) Performing the Story-based Event Clustering task with the BIRCH clustering algorithm.

$n_1, n_2, \ldots, n_n \epsilon\ N$, where the hyper-node in this graph model represents a unique ID number that identifies the event.

Definition 2 (Story). A story is a collection of related events that have been published in a news source, occurring at a specific time $t \epsilon\ T$ following an event $n_i \epsilon\ N$. A story is represented by a set of events, denoted as $s_1, s_2, \ldots, s_m \epsilon\ S$, where the cardinality of the set $s(s_i)$ is greater than or equal to two. It is crucial to highlight that each element in the story set is unique and contains distinct information compared with other elements within the same story.

A. EXTRACTING GRAPH NODES FROM NEWS TEXT

The TDT task uses a corpus of real-world events that have had an impact on public opinion and requires the organization of highly interrelated events into story clusters. In this task, an event is represented by triples <location, time, people involved>, which gives rise to a series of news articles over time [7]. The accurate and complete discovery of these triples hidden in news articles presented in textual form by natural language processing solutions and their integration as nodes in event-graph models are crucial for the creation of embeddings capable of representing the event. The procedure for obtaining graph nodes from a sample news text in proposed method is illustrated in Figure 2.

Named Entity Extraction: The identification of entities such as persons, organizations, and locations in textual content is defined by the task Named Entity Recognition (NER) in the natural language processing research area. The Bidirectional Encoding Representations from Transformers (BERT) language model were used for the Turkish-specific tests. Specifically, the BERTurk model, pre-trained for the Turkish language, was fine-tuned using the Turkish NER dataset, which includes named entity tags for persons, organizations, and locations [48]. The trained BERT-based NER model achieved an F1 score exceeding 96%. To the best of our knowledge, this model represents the highest-performing approach to named-entity detection in Turkish. CPU-optimized pipelines in the Spacy library were used in the tests for the English, German, and Spanish languages. Multi-CNN models pre-trained with OntoNotes for English [49], TIGER and WikiNER corpuses for German [50], and UD Spanish AnCora and WikiNER for Spanish [51] are presented, and their performance in the NER task is over 85% F1 score.

Definition 3 (Named entities of an event). In the context of an event $n_i \epsilon\ N$, named entities are identified from text using the NER model. These named entities consist of person entities $p_1, p_2, \ldots, p_i \epsilon\ P$, location entities $l_1, l_2, \ldots, l_k \epsilon\ L$, and organization entities $o_1, o_2, \ldots, o_x \epsilon\ O$.

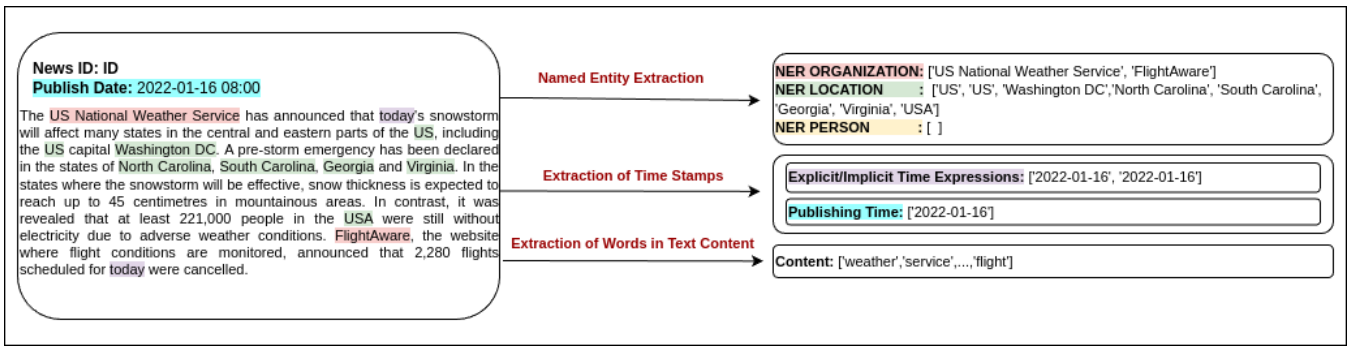

FIGURE 2. Extracting graph nodes from a sample news text. The text was randomly selected from the Story-Based News Dataset. The presented text content and results were translated into English. The models and methodology employed for the three fundamental steps of named entity extraction, time stamp extraction, and word extraction from text content are elucidated in detail in the sub-heading "EXTRACTING GRAPH NODES FROM NEWS TEXT" under "METHODOLOGY".

Extraction of Time Stamps: In datasets comprising news articles and events that occurred at specific time points in the past, the notion of "time" is crucial for understanding the temporal relationships between different news articles and events. Two types of timestamps were considered in the analysis of the news texts:

- *Explicit or Implicit Time Expressions:* Explicit/Implicit Time Expressions are defined as time markers within news content that semantically differentiate one event from another. In this study, the multilingual temporal expression tagger HeidelTime, developed by the Database Systems Research Group at the University of Heidelberg, was used to extract four types of temporal expressions: date, time, duration and repeated time. This rule-based open-source time tagger extracts temporal expressions from text and represents them according to the TIMEX3 standard. It includes manually generated resources for 13 different languages, including English, German, and Spanish, and automatically generates resources for more than 200 languages. HeidelTime can identify temporal expressions with an F1 score of over 90% for English, Spanish, and Turkish and over 85% for German [52], [53].
- *Time of Publication of the News:* This refers to the timestamp indicating when a news article is published. The time of publication is crucial piece of information, particularly for determining the temporal sequence of events that form a story. The datasets used in this study includes news text, sources, and date information, allowing for the extraction of each news article's publication time. By considering these two types of timestamps, this study captures both the temporal aspects of news content and temporal ordering of events through the publication dates of news articles.

Definition 4 (Explicit/Implicit Time Expressions). In the context of an event $n_i \epsilon\ N$, the time expressions are identified in the text. These time expressions are denoted as $t_1, t_2, \ldots, t_q \epsilon\ T$ and are determined using a rule-based temporal expression tagger. They are classified as "date" type and adhere to TimeML standards, which provide a standard representation of temporal information.

Definition 5 (Publishing Time). The publishing time of event $n_i \epsilon\ N$ in the news source is represented by dates $t_1, t_2, \ldots, t_q \epsilon\ T$. These dates were classified as "date" types, similar to explicit/implicit time expressions. They are identified using the same methodology as the temporal expression tagger, which assigns appropriate labels to indicate the publishing dates of the events.

Extraction of Words in Text Content: The words present in the news text form the fundamental elements that convey information about the event. To incorporate the content of each news item, a preprocessing pipeline was applied, which involved steps such as removing stop words, hyperlinks, punctuation marks, and numbers. The resulting words are eliminated according to TF-IDF score (a constraint of 500 words with the highest TF-IDF score in dataset was applied) and each of the remaining words acts as a node in the event graph and are converted to lowercase letters for standardization. To establish the sequential relationship between these words, an edge labelled as 'NEXT' is added based on their order of appearance in the text. This graph model, used to represent news content, is referred to as "SnakeGraph" [54]. SnakeGraph captures the structural and sequential aspects of news text, enabling the analysis and visualization of word-level relationships within the event. By incorporating this graph model, the important linguistic components of news content are integrated into the overall event representation.

Definition 6 (Content). The content of an event $n_i \epsilon\ N$ is represented by a set of unique words $w_1, w_2, \ldots, w_z \epsilon\ W$. These words encompass textual information that conveys the details and characteristics of an event.

B. DISCOVERING NAMED ENTITY-CENTERED FREQUENT SUBGRAPHS

Event texts obtained from news sources at different times are expected to be similar to related events that occur in previous and/or subsequent periods. For example, the news on the first day of the great earthquake that occurred in Turkey in

1999 triggered events related to the earthquake that continued for several days, and the news texts related to this event contained similar elements directly or indirectly. Therefore, we aim to increase the similarity of the representations of event graphs containing the same patterns by detecting of frequent subgraphs in the clustering of events belonging to the same story.

While news texts belonging to the same story tend to contain common patterns in terms of person/institution, location, and time stamps, which are considered the basic elements of the event, it is not a common approach to realize a narrative using the same word sequence in the news content. Therefore, instead of focusing on text content, we focus on the discovery of frequently recurring subgraphs based on named entities.

Simple Event Graph Model: In study, a model was built for each individual event using an unweighted directed graph, which served as the basis for the 'Event Graph Model'. This simple graph model contains a hypernode (NEWS), that represents a unique event ID. In addition, the graph contains various types of nodes, such as consecutive unique words (WORDS) derived from the textual content of the event as well as the person, organization, and location terms (ENTITY) defined by the NER model and time stamps (DATE). The hyper-node connects to various nodes through seven edge types: TOKENS, NEXT, PERSON, LOCATION, ORGANISATION, PUBLISH DATE, and CONTAINS DATE. The unique event ID hyper-node (NEWS) is linked to the first WORDS node in the graph using the TOKENS relationship in the order of appearance in the text. The remaining WORDS nodes are connected to each other by the SnakeGraph structure using the NEXT relationship, which indicates the order of the terms in the news text. In addition, the NEWS hypernode is connected to named entity nodes (ENTITY) using edges such as PERSON, LOCATION, and ORGANISATION, depending on the entity type. Furthermore, two different edge types, PUBLISH DATE and CONTAINS DATE, connect the publication date of the news text and explicit/implicit time expressions in the NEWS node content. Figure 3 shows a simple event graph structure.

Frequent Subgraph Mining: Frequent subgraphs appear more frequently than the user-specified threshold (minsup) within a collection of graphs. In this study, a graph set was formed using graph representations of the news texts in the dataset. The frequent subgraphs extracted from this graph set serve as patterns and provide insights into the characteristic features shared by news texts within a cluster. Because each news text in the dataset consists of word sequences of varying lengths, the corresponding graphs may have different numbers of nodes and edges. However, each news text is represented by a graph, which necessitates the use of algorithms capable of efficiently finding frequent subgraphs from numerous small graph sets. The gSpan algorithm proposed by Yan et al. was the preferred choice for this task [55]. It employs Depth First Search (DFS) and addresses important challenges such as efficient runtime, avoidance of unnecessary candidate generation, and mitigation of the computational complexity

Algorithm 1: Simple Event Graph Construction

```
Input: News Text (news_text)
Output: Simple Event Graph representation of the
        news content (G_SE)
Function PreprocessNewsText(news_text):
    cleaned_text ← RemoveStopWords
     (RemoveHyperlinks (RemovePunctuation
     (RemoveNumbers(news_text))))
    tokens ← TokenizeText(cleaned_text)
    normalized_tokens ←
     ConvertToLowercase(tokens)
    return normalized_tokens
Function
 CreateEventGraph(normalized_tokens,
 named_entities, time_stamps, publication_date):
    Initialize EventGraph as an empty graph
    AddNode(EventGraph, "NEWS", event_id)
    AddNode(EventGraph, "WORDS",
     normalized_tokens[0])
    AddEdge(EventGraph, "NEWS",
     normalized_tokens[0], label='TOKENS')
    for i from 1 to length(normalized_tokens) - 1 do
        AddNode(EventGraph, "WORDS",
         normalized_tokens[i])
        AddEdge(EventGraph,
         normalized_tokens[i-1],
         normalized_tokens[i], label='NEXT')
    foreach entity, entity_type in named_entities do
        AddNode(EventGraph, entity_type, entity)
        AddEdge(EventGraph, "NEWS", entity,
         label=entity_type)
    foreach time_stamp in time_stamps do
        AddNode(EventGraph, "DATE", time_stamp)
        AddEdge(EventGraph, "NEWS", time_stamp,
         label="CONTAINS DATE")
    AddNode(EventGraph, "DATE",
     publication_date)
    AddEdge(EventGraph, "NEWS",
     publication_date, label="PUBLISH DATE")
Function Main(news_text):
    normalized_tokens ←
     PreprocessNewsText(news_text)
    named_entities ←
     ExtractNamedEntities(news_text)
    time_stamps ← ExtractTimeStamps(news_text)
    publication_date ←
     ExtractPublicationDate(news_text)
    EventGraph ←
     CreateEventGraph(normalized_tokens,
     named_entities, time_stamps, publication_date)
    return EventGraph
```

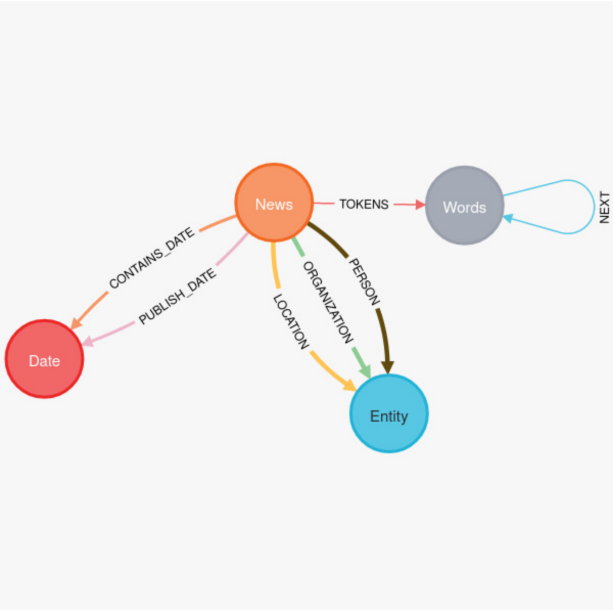

FIGURE 3. Simple Event Graph Structure

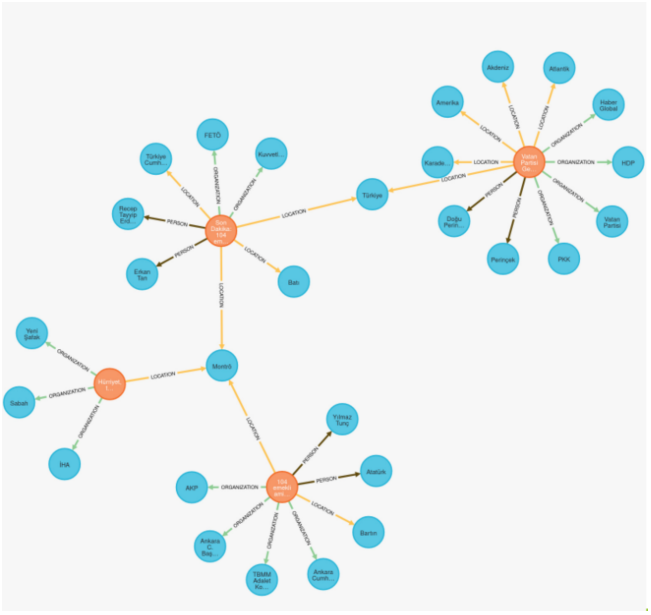

FIGURE 4. Star Graph Structure Example

associated with graph isomorphism. The "minsup" parameter is determined based on the minimum number of news texts within each cluster. This parameter sets the threshold for the minimum occurrence count of a subgraph to be considered frequent, thereby allowing the discovery of meaningful patterns in the dataset. A 2-step filtering was applied for the discovered frequent subgraphs. Firstly, a two-stage ranking was performed according to the graph size and then according to the support value. Accordingly, the first 200 frequent subgraphs with the highest values were accepted and the others were eliminated.

Named Entity-Centered Approach to Frequent Subgraph Mining: In the Simple Event Graph model, the key elements that define and differentiate an event from another are named entities (people, locations, and organizations) and temporal information. The presence of repeated named entities and time tags in similar order across different texts indicates that these texts share the same or similar news content. In the graph model, all ENTITY and DATE nodes are directly connected to a unique event ID (NEWS) hypernode with different edge types. However, there is no direct path between the ENTITY or DATE nodes in the Simple Event Graph model. This means that it is not possible to form subgraphs that include both ENTITY and DATE nodes. This limitation is referred to as the "Star Graph Problem". Figure 4 shows an example of a star-graph structure. The subgraphs obtained using the gSpan algorithm in this model consist mainly of sequential and repetitive word sequences from news text content. Although these word sequences may belong to the same story, they are unlikely to they occur frequently across news texts. To address the Star Graph Problem, an approach is proposed to create a path between the ENTITY and DATE nodes by introducing an edge called 'NEXT ENTITY'. This addition allows for the formation of more meaningful and separable patterns that include both named entities and time tags, which are considered the core elements of an event.

C. CONSTRUCT EVENT GRAPH MODEL

In this study, a graph model called the "Event Graph" or "Extended Event Graph with Named Entity-Centered Subgraphs" was used for the story-based event clustering approach. This graph model is an extension of the baseline event graph model and incorporates Named Entity-Centered Subgraph nodes.

The Event Graph consists of five different node types: NEWS, WORDS, ENTITY, DATE, and SUBGRAPH. These nodes represent the unique event ID, words in the news text, named entities, temporal information, and subgraphs. The graph also includes nine different edge types: TOKENS, NEXT, PERSON, LOCATION, ORGANIZATION, PUBLISH DATE, CONTAINS DATE, CONTAINS SUBGRAPH, and CONTAINS VERTEX. These edges represent the relationships between nodes, such as the order of words, connections between named entities and the event ID, and associations between the event ID and the temporal information. The node types in the Event Graph are formally defined as follows:

Definition 7 (NEWS). The NEWS node in the Event Graph represents the unique ID number that distinguishes the news text, referred to as event $n_i \epsilon\, N$, from other news texts.

Definition 8 (WORDS). The WORDS node in the Event Graph represents a set of unique words $w_1, w_2, \ldots, w_z \epsilon\, W$ contained in the news text content and filtered by TF-IDF score. The WORDS node captures textual content that describes and provides information about the event.

Definition 9 (ENTITY). The ENTITY nodes in the Event

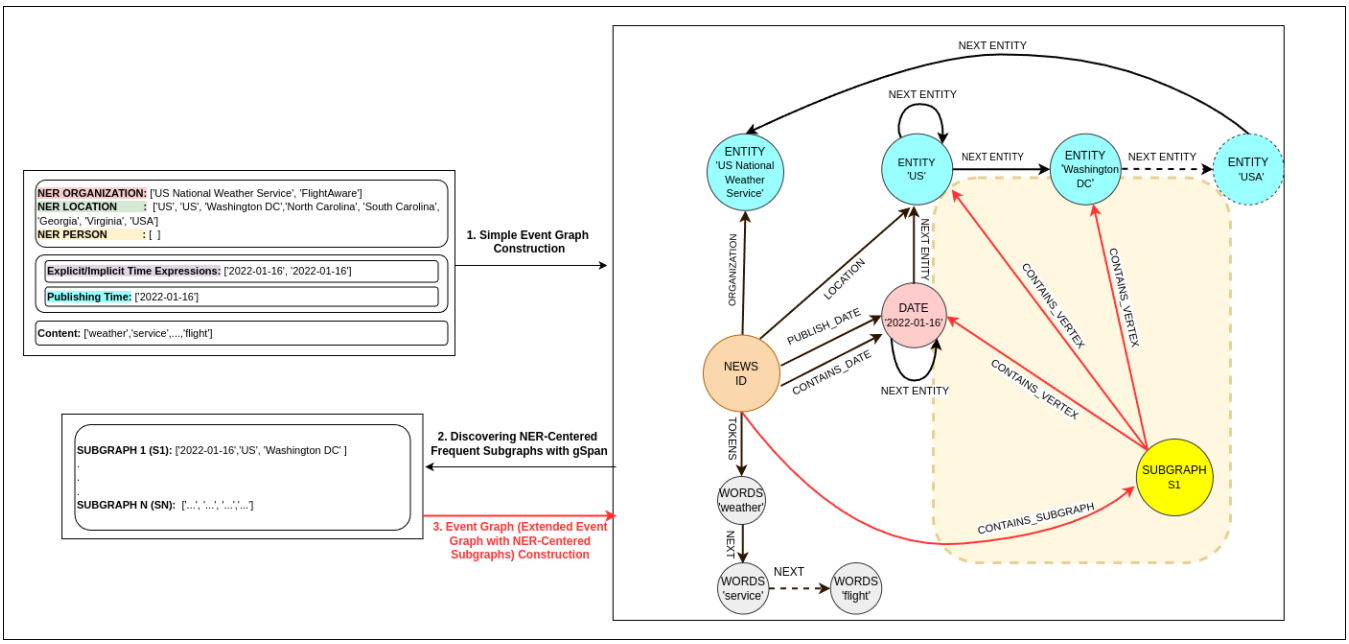

FIGURE 5. **Event Graph construction from a sample news text. The edges in black are constructed in the first step (Simple Event Graph Construction), while the edges in red and the SUBGRAPH (S1) node represent the edges added in step 3 to create the event graph, which is an extended graph with named entity-centered subgraphs.**

Graph represent named entities identified by the NER model from the text of event $n_i \epsilon\, N$. These named entities can include persons $p_1, p_2, \ldots, p_j \epsilon\, P$, $l_1, l_2, \ldots, l_k \epsilon\, L$ locations, and organizations $o_1, o_2, \ldots, o_x \epsilon\, O$.

Definition 10 (DATE). The DATE nodes in the Event Graph represent the publication date of the news text (referred to as the event $n_i \epsilon\, N$) as well as any time expressions in the text. These time expressions were determined using a rule-based temporal expression labeler and categorized as "date" based on TimeML standards.

Definition 11 (Named Entity-Centered Frequent Subgraph). Frequent patterns structured by nodes representing named entities (ENTITY) and/or time expressions (DATE). The gSpan algorithm is used to detect these patterns that exceed a predefined minimum support (minsup) threshold through a 2-step filtering process. Each identified pattern is assigned a unique identification number, which is stored in SUBGRAPH nodes. The minsup threshold serves as a user-defined parameter that determines the minimum number of occurrences required for a subgraph to be considered frequent.

In the extended Event Graph model, a unique identification number is assigned to each undirected subgraph that is detected by the gSpan algorithm based on the minsup threshold and 2-step filtering. This identification number was used to create a SUBGRAPH node representing the subgraph within the baseline graph structure of each event. A CONTAINS SUBGRAPH edge is then created between SUBGRAPH node and the NEWS hypernode, indicating that the subgraph is part of the overall event. In addition, CONTAINS VERTEX edges are established between the ENTITY and/or WORDS nodes, which comprise the structure of the detected subgraph and the corresponding SUBGRAPH node. This allows for a connection between individual elements (entities or words) and the larger subgraph to which they belong. By incorporating these additional nodes and edges, the Event Graph model was expanded to capture the specific subgraph patterns within each event, as depicted in Figure 6. In addition, the methodology for the construction of an event graph from a sample news text is illustrated in Figure 5.

D. EVENT GRAPH EMBEDDING LEARNING WITH NODE2VEC

Each node u in a directed and unweighted event graph $G = (V, E)$ is represented in a low-dimensional space by the node2Vec model with embedding vector $ev(u)$. The learning of each entity (news, words, entity, date, and subgraph) in the event graph model, represented by a node in graph G, is formulated as a maximum likelihood optimization problem. The dimension of the embedding vector (feature representation) $f : V \rightarrow R$ is defined as the mapping function from the nodes that we intend to learn for the embedding vectors. For each source node $u \epsilon V$ on the event graph, $N_S(u) \subset V$ is the network neighborhood of node u generated by the neighborhood sampling strategy S. We aim to optimize the objective function given by Equation 1 with respect to the feature representation given by f :

$$\max_f \sum_{u \epsilon V} \log Pr\left(N_s\left(u\right) | f\left(u\right)\right) \tag{1}$$

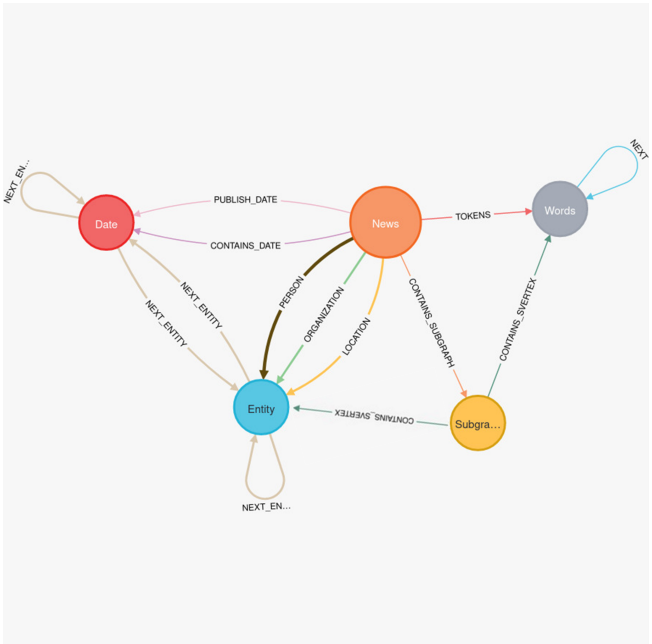

FIGURE 6. Event Graph - Extended Event Graph with Named Entity-Centered Subgraphs

To learn the representation of a given node u in graph G, which is specialized to organize it according to its network role and/or the community to which it belongs, random walks of length l with bias were generated to explore its different neighborhoods. Given that π_{vx} represents the unnormalized transition probability between nodes v and x, while Z denotes the normalization constant, and c_i is the i-th node starting with node $c_0 = u$, the distribution in Equation 2 is used to generate a random walk:

$$P(c_i = x | c_{i-1} = v) = \begin{cases} \frac{\pi_{vx}}{Z} & : \ if \ (v,x) \in E \\ 0 & : \ otherwise \end{cases} \tag{2}$$

A biased walk of the second order is defined based on the turn parameter p and in-out parameter q. The selection of the next node was performed using static edge weights $\pi_{vx} = w_{vx}$. Event graph G is unweighted, where w_{vx} is equal to 1. The unnormalized transition probability for determining the next walking step is $\pi_{vx} = \alpha_{pq}(t,x).w_{vx}$ when a random walk passes the edge (t,v) in the event graph G and is currently at node v. It is essential to note that d_{tx} must be either 0, 1, or 2, representing the shortest path distance between nodes t and x, as seen in Equation 3:

$$\alpha_{pq}(t,x) = \begin{cases} \frac{1}{p} & : \ if \ d_{tx} = 0 \\ 1 & : \ if \ d_{tx} = 1 \\ \frac{1}{q} & : \ if \ d_{tx} = 2 \end{cases} \tag{3}$$

To obtain the event embedding $ev_{event}(G)$ for an event graph G, we computed the arithmetic mean of the Node2Vec node embeddings of all nodes as shown in Equation 4, where the total number of nodes in the graph is denoted by $|V|$:

$$ev_{event}(G) = \frac{1}{|V|} \sum_{u \epsilon V} ev(u) \tag{4}$$

This process enables the capture of collective information and characteristics of the nodes in the event graph, resulting in a representative embedding for the entire event. It is noteworthy that event graphs with common words, entities, times, or subgraph nodes have similar vector representations. This similarity in vector representations indicates semantic and structural relationships between the corresponding event graphs.

E. STORY-BASED EVENT CLUSTERING

The proposed event graph model and the embedding of event $ev_{event}(G)$ in the study were tested on two different traditional clustering algorithms, namely BIRCH and HDBSCAN, with the aim of analyzing their ability to represent the fundamental characteristics of events and the stories to which these events belong, as well as grouping vectors with similar features to define clusters of relevant stories.

The Balanced Iterative Reducing and Clustering using Hierarchies (BIRCH) algorithm, which is specifically designed to handle large datasets, addresses the computational complexity inherent in clustering tasks. It performs this using a Cluster Feature Tree (CF Tree), a tree-like structure constructed to encapsulate important clustering features without relying on a distance matrix. This hierarchical tree enables direct clustering operations, facilitating efficient and scalable clustering of large volumes of data. BIRCH uses parameters such as the branching factor $'B'$ and the threshold $'T'$ to hierarchically organize clusters. CF tree construction involves mathematical representations as shown in Equation 5:

$$CF = (N, LS, SS) \tag{5}$$

where N denotes the data points in a cluster, LS denotes the linear sum, and SS denotes squared sum of data point values. By adopting an incremental learning approach, BIRCH continuously scans the input data, updates the CF tree, merges close microclusters and organizes clusters hierarchically, making it well-suited to text clustering tasks.

Hierarchical Density-Based Spatial Clustering of Noisy Applications (HDBSCAN) is an evolution of the DBSCAN algorithm that uses hierarchical clustering principles and is characterized by its ability to detect clusters of different shapes and densities. It uses parameters such as the minimum number of points (*MinPts*) and minimum size of the cluster to calculate the density. The HDBSCAN clustering algorithm first constructs a mutual reachability graph that encodes the pairwise distances between data points. This graph forms the basis of a minimum spanning tree (*MST*), which is further refined by robust single-link clustering to form a hierarchical structure. A stability measure derived from the hierarchical structure was used to assess cluster stability by estimating the robustness of the reachability distances. This metric measures the probability that a point belongs to a cluster and

the distribution of mutual reachability distances, resulting in a stability score. HDBSCAN's adaptability to complex structures in high-dimensional spaces is a powerful tool for clustering tasks.

News texts are clustered into stories of varying sizes depending on the impact and continuity of the event they contain. Therefore, this study deals with datasets with very different cluster densities, although the number of clusters is large and the dataset size is not. Therefore, clustering algorithms that are sensitive to differences in the cluster shape and density are unsuitable for this task. Considering all these constraints as well as the time and memory complexity, the BIRCH and HDBSCAN algorithms, which do not require the number of clusters as an input parameter, are preferred. Benchmark tests were performed on two different algorithms to evaluate the ability of the proposed event graph model and vector representation to represent the event, to distinguish the story to which it belongs from other stories independently of the algorithm, and to provide a versatile perspective. When the test results were analyzed, it was observed that the BIRCH clustering algorithm outperformed HDBSCAN in terms of the B-cubed $F1$ score evaluation metric for detecting story clusters. Moreover, the BIRCH algorithm has $O(n)$ time complexity whereas HDBSCAN has $O(n^2)$ time complexity. Considering the clustering performance, time and space complexity, the BIRCH algorithm is recommended as the preferred method in this study.

IV. RESULTS AND DISCUSSION

In this section, various experimental studies and their results are presented on the clustering task to evaluate the representation capability of the event graph, which is extended with named entity-centered frequent subgraphs as proposed in the study. The evaluation datasets used in the experiments, evaluation criteria and strategies, experimental settings, and environment are explained in detail. In Table 7, we provide a clear comparison that demonstrates that our graph model and the clustering algorithm we used perform superiorly compared to state-of-the-art approaches independently of language, on the same evaluation dataset and task. In additional, a meticulous investigation was conducted to distinguish the effects of different node and edge connections in a simple event graph and their impact on the ability of the graph model to represent an event. This research is crucial for assessing the effectiveness of named entity-centered frequent subgraph nodes in capturing the significant features and relationships of an event and its associated story.

The evaluation of the Event Graph-Based News Clustering approach focuses on addressing the following research questions:

- RQ1. What information should be stored in the nodes of a Simple Event Graph that models the event, and what are their impacts on clustering performance?
- RQ2. To what extent does an Event Graph model extended with Named Entity-Centered Frequent Subgraph Nodes enhance clustering performance?
- RQ3. What is the performance comparison between the Event Graph-Based News Clustering approach and the state-of-the-art methods?

A. EXPERIMENTAL SETUP

The experiments conducted in this study used the Story-based News Dataset and the News Clustering Dataset was used as a benchmark for the multilingual TDT task. All experiments were performed on a computer with 64 GB memory, Intel(R) Xeon(R) CPU E5-2680 v3 2.50GHz (24 cores), and an NVIDIA RTX A5000 (24 GB) graphics card. The operating system installed in the experimental environment was Windows 10 Pro and the programming language was Python 3.9.15.

Language	Train		Test	
	#Documents	#Story	#Documents	#Story
English	12233	593	8726	222
German	4043	377	2101	118
Spanish	4527	416	2177	149
Slovenian	-	-	37	3
French	-	-	61	2
Italian	-	-	88	2
Chinese	10	1	440	9
Russian	-	-	231	1
Croatian	-	-	13	2

TABLE 1. Statistics for the training and testing partitions of the News Clustering Dataset

B. DATASETS

1) STORY-BASED NEWS DATASET

The "Story-Based News Dataset" was specifically created to address the clustering of news texts that represent events from the same or different stories, published at different times and from various news sources. The dataset comprises 2031 events, which are news texts written in Turkish and sourced from online news platforms. These events were organized into 168 stories. Each story contained a minimum of two events and a maximum of 44 events.

Table 1 presents two case studies that showcase different stories in a dataset. The topics covered in these stories were manually selected by the researchers, focusing on events that have impacted the public agenda since 2021. To gather news data, the researchers used the Google News platform and developed a script that automatically retrieved Turkish news articles related to each topic. The dataset underwent a rigorous verification process conducted by two independent human validators with expertise in natural language processing and practical experience in developing related applications. The creation of the "Story-Based News Dataset" enables researchers to investigate and analyze the clustering

of news texts within the context of specific stories. The dataset provides valuable resources for exploring various natural language processing tasks and serves as a foundation for developing and evaluating novel approaches for text clustering.

In the initial phase, a purification process was conducted to filter out news texts that were irrelevant to the story, contained incomplete or incorrect content, or were duplicates. The preprocessed story sets were sorted based on their publication dates. Time windows were created, spanning up to three hours after the first news article was published. It was anticipated that news texts within the same story set and time window might contain similar content but originate from different news sources.

To address potential duplications or repetitions in the events of a story, a second verification process is conducted using human validators. The objective was to ensure that the dataset did not contain any redundant data. Among the news texts that fit this prediction, only the longest texts in terms of character counts were retained in the story set, whereas other repetitive news texts were eliminated. This approach allows the identification of events that provide distinct information about a story. The fact that the dataset was specifically created in Turkish holds significance for researchers working in less-resourced languages such as Turkish. In this regard, the "Story-Based News Dataset" designed for Turkish natural language processing researchers represents a pioneering effort within the literature. This offers a valuable resource for exploring the various aspects of news analysis and clustering in the context of specific stories.

2) EVALUATION DATASETS

In order to evaluate the methods proposed in this paper, we use "Story-Based News Dataset" and a news dataset [16] commonly used in multilingual text clustering research.

- **Story-Based News Dataset (SBN).** The full Turkish dataset contains 168 distinct stories and 2031 events (news articles).
- **News Clustering Dataset (NC).** The dataset used as a benchmark in the multilingual TDT task contained documents in English, Spanish, German, Chinese, Russian, French, and many other languages [16]. Table 2 presents the statistics of the dataset. In the experiments conducted in this study, test sets presented in three basic languages (English, German, and Spanish) were used.

C. PERFORMANCE EVALUATION

1) EVALUATION METRICS

Two primary types of evaluation criteria are commonly utilized in clustering tasks: intrinsic and extrinsic. Intrinsic metrics gauge the proximity of items within the same cluster and their segregation from other clusters. Conversely, extrinsic metrics leverage labels of data samples or analogous external ground truths, referred to as a human-made gold standard, to evaluate the quality and efficacy of clustering results [66]. In this study, gold standard datasets prepared using human annotators were employed. Hence, the utilization of external evaluation metrics to appraise the performance of the proposed clustering algorithm was deemed appropriate. B-cubed is the only external clustering metric that satisfies all the formal constraints that evaluation metrics should be intuitive, clarify limitations, be formally provable, and distinguish between families of metrics grouped on mathematical principles [56]. It also serves as a standard evaluation metric documented in the literature for the news clustering problem considered in our research [7].

Let $L(e)$ and $C(e)$ represent the category and set of elements e, respectively. The accuracy of the relationship between e and e' in the distribution, as well as other B-Cubed metrics, is defined as shown in Equation 6, Equation 7 and Equation 8 [57], [58]:

$$Correctness(e, e') = \begin{cases} 1 & : \textit{iff } L(e') \leftrightarrow C(e) = C(e')) \\ 0 & : \textit{otherwise} \end{cases} \tag{6}$$

$$Precision\,Bcubed = Avg_{e'}\left[Avg_{e'.C(e)=C(e')}\left[Correctness(e, e')\right]\right] \tag{7}$$

$$Recall\,Bcubed = Avg_{e'}\left[Avg_{e'.L(e)=L(e')}\left[Correctness(e, e')\right]\right] \tag{8}$$

In the experiments conducted on the evaluation datasets, B-Cubed Precision, B-Cubed Recall, and B-Cubed F-scores were utilized to evaluate the impact of different event graphs and named entity-centered frequent subgraph nodes on clustering performance.

2) EVALUATION METHODOLOGY AND PARAMETER SETTINGS

In this study, two traditional clustering algorithms, BIRCH and HDBSCAN, designed to efficiently process large datasets, requiring a few parameters and not requiring predetermination of the number of clusters, but completely different from each other in time complexity and working methodology, were used. The use of two different clustering algorithms in the experiments provides the opportunity to evaluate the effects of graph models created with different node and edge types on the clustering performance of the event graph modelling and the effect of Named Entity-Centered Frequent Subgraph nodes on the representation ability of the event graph independently of the clustering algorithms.

BIRCH, which garnered the 2006 SIGMOD Test of Time Award, is recognized as one of the fastest integrated hierarchical clustering algorithms to date. Its significance lies in its profound influence on subsequent clustering methodologies, particularly due to its effectiveness in handling large-scale data [63]. Central to BIRCH's mechanism is the cluster feature tree (CF Tree), wherein the cluster features (CF) serve

	Story 1 (Cluster 1)
Event 1	"Breaking news: The plane skidded off the runway at Trabzon Airport" The Pegasus passenger plane, which was flying from Ankara to Trabzon, veered off the runway upon landing at Trabzon Airport. According to initial assessments, the passengers onboard the plane were reported to be in good condition, and cooling procedures were conducted. All passengers from the Ankara-Trabzon flight were safely evacuated. Trabzon Governor stated that a technical and administrative investigation was carried out, and he confirmed, "There are no injuries."
Event 2	"The fate of that plane in Trabzon became clear after 200 days" - During the Ankara-Trabzon flight, the dismantling process of the Pegasus Airlines plane named 'Zeynep', which veered off the runway during landing at Trabzon Airport on January 13th, has commenced. The removal operation, which lasted approximately 20 hours, has begun. It is reported that the dismantling of the aircraft will be completed within a month, and all components will be sent to England.
Event 3	"The plane leaving the runway in Trabzon will be a pita lounge"- In a written statement, Yomra Mayor Mustafa Bıyık announced that the plane wreckage will be made available to the entire Black Sea region without any financial burden on the municipality. Bıyık stated that the plane will be transformed into a pita lounge as part of an upcoming project. He further explained that the plane will be placed in the tea garden owned by the municipality in the Sancak District and operated by tenants. The installation and construction costs, totaling 4 million liras, will be fully covered by the operators.
	Story 2 (Cluster 2)
Event 1	"Rize-Artvin Airport opens to service at the end of the year"- Governor Çeber emphasized that the new airport, with a capacity to accommodate 3 million passengers annually, will not only serve as a transportation hub but also contribute significantly to the development of Rize, Artvin, the Black Sea region, and the country as a whole in terms of trade, tourism, agriculture, industry, and culture. Highlighting the unique climate and geographical characteristics of Rize, Çeber stated, "Although we occasionally face extraordinary circumstances such as natural disasters, we anticipate opening the airport by the end of this year if there are no unforeseen events. President RTE closely monitors the progress, receiving regular updates on a daily or weekly basis."
Event 2	"Rize-Artvin Airport changed Turkey's surface area"- Rize-Artvin Airport, designed by the Ministry of Transport and Infrastructure, was constructed on a 1000-hectare area in Yeşilköy, located in the Pazar district of Rize. The foundation of the airport was laid by President RTE on April 3, 2017. This significant project, with an estimated annual capacity of 3 million passengers and an infrastructure cost of 1 billion 78 million liras, has transformed the landscape of Turkey. Approximately 100 million tons of stones were used for the sea fill, resulting in an expansion of the country's surface area by 2.8 square kilometers. The land created through this filling process is clearly visible in satellite imagery, showcasing the scale of this remarkable achievement.
Event 3	"Rize-Artvin Airport, which is illuminated brightly at night, offers a visual feast in the shape of a teacup."- The construction of the Rize-Artvin Airport project, which will be the second airport in Turkey and the Black Sea Region to be built on a sea fill, is progressing steadily. Following the completion of the sea fill on a vast area of 2.8 million square meters, which began with the laying of its foundation on April 3, 2017, and involved the use of 100 million tons of stone, the terminal buildings featuring regional architecture, the tea glass tower, and the entrance ornament in the shape of tea leaves have been finished. One of the standout features of the airport is its teacup-shaped tower, which mesmerizes with a stunning visual display at night, showcasing the Turkish Flag through its illuminated exterior lights.

TABLE 2. **Story-based News Dataset**

HDBSCAN				
Parameters	Datasets			
	SBN (Turkish)	NC (English)	NC (German)	NC (Spanish)
Min Samples	1	4	3	3
Min Cluster Size	2	5	5	4
BIRCH				
Branching Factor	30	34	27	28
Threshold	5	6	5	5
N Cluster	None	None	None	None

TABLE 3. **Parameter Settings for Clustering Algorithms**

as nodes. The architecture of this tree is shaped by three core parameters: threshold, branching factor, and number of clusters [62]. In this framework, genuine clusters are represented by the leaf nodes of the CF Tree. The threshold parameter, symbolized as T, sets the limits on the number of data points or the maximum diameter a leaf node entry can encompass [66]. A data point is added to an existing cluster only if the cluster's radius remains within this threshold; exceeding it necessitates the formation of a new, initially empty cluster. This parameter, therefore, plays a crucial role in controlling the size of the clusters. The branching factor, indicated as Br, determines the maximum number of CF subsets each internal node can hold, effectively constraining the number of children a node can support [63]. If the addition of a new cluster causes the number of children to surpass the branching factor, the node is required to split. To preserve structural balance within the CF Tree, such splitting may need to be recursive at higher nodes. Furthermore, if a predefined number of clusters parameter is set, it allows for the merging of neighboring clusters to optimize cluster quality. Ultimately, the algorithm can be configured either to produce a predetermined number of clusters at completion or to output intermediate clusters without executing a final clustering step, by setting the number of clusters parameter to 'none' [63], [66]. This flexibility in defining the end state of clustering is a key aspect of BIRCH's design, allowing for adaptability based on specific

data or operational requirements.

HDBSCAN has emerged as a leading algorithm in the field of density-based clustering due to its capacity to effectively identify nested clusters of varying densities [59]. In contrast to other density-based clustering methods, HDBSCAN does not necessitate a substantial set of critical input parameters. Instead, it is sensitive to two heuristic parameters: "min cluster size" and "min samples (*MinPts*)". The "min cluster size" is the core parameter of HDBSCAN as it defines the minimum number of data points required to form a cluster [60]. Consequently, as this parameter increases, the resulting clusters tend to be larger, potentially reducing the total number of clusters by merging smaller clusters [61]. Data points that do not meet the minimum cluster size criterion are classified as noise. The parameter "min samples (*MinPts*)" is an optional parameter for the HDBSCAN algorithm and specifies the minimum number of neighbouring data points that a core point must contain within a given radius, called the eps structure diameter, to form a dense region [60]. Increasing the value of *MinPts* raises the threshold for the identification of high-density regions, while increasing the number of data points categorised as noise concentrates clusters into increasingly dense regions. Conversely, a low setting for *MinPts* increases the algorithm's sensitivity to local density variations, which can result in large clusters being broken into many small clusters. The default value for min samples is equal to the min cluster size parameter, but this may vary depending on the dataset, especially in real-life data [60].

To ensure optimal parameter selection, a Grid Search was performed using four different datasets for both the algorithms. The optimal parameter values were determined and are listed in Table 3.

D. RESULTS

1) RQ 1: INFLUENCE OF DIFFERENT SIMPLE EVENT GRAPH MODELS ON THE CLUSTERING PERFORMANCE

News articles are event contents represented by <location, time, related persons> triples. To address RQ1, four different fundamental graph models have been designed that incorporate text content, named entities, and time stamps:

- **Content Graph (G_C):** Given a set of unique words $\{w_1, w_2, ..., w_n\}$ that comprise the event text and a unique event identifier *NEWS*, graph $G_C = (V_C, E_C)$ is formulated in Equation 9:

$$G_C = (\{w_1, w_2, ..., w_n, NEWS\}, \{e_{NEXT}(w_i, w_{i+1}), e_{TOKENS}(w_1, NEWS)\}) \quad (9)$$

- **Named Entity and Time Stamps Graph (G_{NerT}):** Given the named entities $\{en_1, en_2, ..., en_m\}$ present in the event text, time stamps $\{d_1, d_2, ..., d_k\}$, and a unique event identifier *NEWS*, the graph $G_{NerT}(V_{NerT}, E_{NerT})$ is formulated in Equation 10:

$$(G_{NerT}) = (\{en_1, en_2, ..., en_m, d_1, d_2, ..., d_k, NEWS\}, \{e_{entityType}(NEWS, en_j), e_{dateType}(NEWS, d_l)\}) \quad (10)$$

The entityType can be 'PERSON', 'LOCATION', or 'ORGANIZATION', and the dateType can be 'CONTAINS DATE' or 'PUBLISH DATE'.

- **Content Summarized with NER Graph (G_{SNer}):** G_{SNer} is a graph model that has the same node and edge structure as G_C, but focuses solely on sentences that contain at least one named entity label. Therefore, it is expressed as $G_{SNer} \subseteq G_C$.
- **Content, Named Entities and Time Stamps Graph (Simple Event Graph- G_{SE}):** The comprehensive and high-dimensional graph model includes all unique words from the event text, named entities identified by the NER model, time tags within sentences, and the news release date. With this structure, G_{SE} is formulated as $G_{SE} \cup G_{NerT}$.

Experiments were conducted on the text clustering problem to analyze the rich information derived from the text that should be stored in the nodes to model an event represented in a news article with a high representational capability graph. These experiments were conducted using evaluation datasets in four different languages: the Story-Based News Dataset (Turkish), the News Clustering Dataset (English, Spanish, and German), and two different clustering algorithms. The results were measured in terms of B-cubed Precision, Recall, and F1 evaluation metrics, and are presented in Table 4. Upon examining the findings for all datasets and both clustering algorithms, the following observations can be made:

- The G_C graph model, which is constructed solely with unique words from the event text, shows a significant improvement in clustering performance of up to 5 % in terms of the B-cubed F1 metric, when expanded with additional nodes representing named entities and temporal information.
- When considering the G_{NerT} graph model, which consists solely of named entities and time stamps, there is a noticeable decline of up to 10 % in the B-cubed F1 score compared to the G_C graph model.
- The G_{SNer} graph, which is created by summarizing the event content with sentences containing at least one named entity, leads to a significant reduction in the number of graph nodes and edges compared to the G_C graph model, but does not cause a significant decline in clustering performance.

Considering all the results obtained from Table 4 and listed above, it can be observed that although the elements of place, time, person, and organization, which are considered descriptors for events, form a strong foundation for the construction of the graph model, a graph of an event created solely with these elements, independent of text content, is not sufficient in terms of representational capability.

The experiments conducted to answer this research question also revealed the analyses performed during the creation of the G_{SE} Simple Event Graph Model. However, the outcomes of these analyses pave the way for establishing a baseline that will underpin the creation of the 'Event Graph'

HDBSCAN Clustering Algorithm												
	Story_based News Clustering (Turkish)			News Clustering (English)			News Clustering (German)			News Clustering (Spanish)		
	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1
G_C	0.727	0.768	0.745	0.814	0.786	0.800	0.768	0.788	0.778	0.785	0.794	0.789
G_{NerT}	0.664	0.661	0.661	0.722	0.687	0.704	0.691	0.683	0.687	0.706	0.688	0.697
G_{SNer}	0.705	0.799	0.745	0.819	0.826	0.823	0.760	0.824	0.790	0.776	0.830	0.802
G_{SE}	0.773	0.793	**0.783**	0.861	0.859	**0.860**	0.815	0.838	**0.826**	0.833	0.844	**0.838**
BIRCH Clustering Algorithm												
	Story_based News Clustering (Turkish)			News Clustering (English)			News Clustering (German)			News Clustering (Spanish)		
	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1
G_C	0.839	0.749	0.789	0.881	0.844	0.862	0.829	0.849	0.839	0.842	0.859	0.850
G_{NerT}	0.755	0.679	0.714	0.790	0.751	0.770	0.745	0.754	0.750	0.757	0.763	0.760
G_{SNer}	0.909	0.670	0.772	0.879	0.823	0.850	0.849	0.826	0.838	0.862	0.836	0.849
G_{SE}	0.841	0.808	**0.822**	0.921	0.917	**0.919**	0.859	0.914	**0.886**	0.872	0.925	**0.898**

TABLE 4. The effect of different simple graph models on the clustering performance. The highest F1 score is highlighted in bold.

model, which is expanded with Named Entity Frequent Subgraphs and ultimately proposed in the study.

2) RQ 2: ANALYZING THE EFFECT OF NAMED ENTITY-CENTERED FREQUENT SUBGRAPH NODES ON THE CLUSTERING PERFORMANCE

Events with news value, and thus societal significance, are primarily shared with the public through texts before losing their immediacy. Other texts that are part of the same composition (story), such as developments related to the event or other events triggered by it, become part of the news flow during the ongoing periods. Disparating texts that form different parts of the same event tends to contain common patterns. Building on this, the study presents an approach focused on detecting frequently recurring sub-patterns of named entities, identified in experiments designed for the RQ1 research question, which is of particular importance in event representation.

With the Simple Event Graph Model, referred to as G_{SE}, a performance exceeding 78 % has been achieved in traditional clustering algorithms for almost every language. It has been observed that including named entities as separate nodes in the graph model, along with the news text, enhances the distinguishability of events, thereby leading to a significant improvement in clustering performance.

To address RQ2, considering G_{SE} as a baseline, three additional graph models were developed to investigate the impact of incorporating named entity-centered frequent subgraph nodes into this graph model on the clustering performance. The experimental results for the evaluation dataset in the four different languages are presented in Table 5 in terms of the B-cubed Precision, Recall, and F1 evaluation metrics.

- **Simple Event Graph with Frequent Subgraph Nodes ($G_{SE^{+FS}}$):** An event graph containing any one or several of the frequent subgraphs derived from the G_{SE} graph set was expanded by adding these subgraphs as nodes, resulting in the creation of the $G_{SE^{+FS}}$ graph model. In the G_{SE} graph model, named entities and time stamps are directly connected to the *NEWS* hypernode, and there are no edges between them; therefore, frequent subgraphs are derived only from repeating words in the news content. With $\{sg_1, sg_2, ..., sg_p\}$ being the nodes of frequent subgraphs, $G_{SE^{+FS}}$ $(V_{SE^{+FS}}, E_{SE^{+FS}})$ is formulated in Equation 11:

$$(G_{SE^{+FS}}) = (\{en_1, en_2, ..., en_m, d_1, d_2, ..., d_k, NEWS\}, \{e_{entityType}(NEWS, en_j), e_{dateType}(NEWS, d_l)\}) \quad (11)$$

- **Simple Event Graph with "NEXT ENTITY" edges ($G_{SE^{+NEXTENTITY}}$):** In the baseline G_{SE} graph model, all named entities and time stamps are connected to the *NEWS* hypernode with an edge, and there are no edges among them. As a result, because there is no path between named entities or time stamps, frequent subgraphs can only be derived from text content. To address this issue, the $(G_{SE^{+NEXTENTITY}})$ $(V_{SE^{+NEXTENTITY}}, E_{SE^{+NEXTENTITY}})$ graph model was designed by adding *NEXTENTITY* edges between nodes representing entities or time stamps that characterize the event. The graph model is formulated in Equation 12:

	HDBSCAN Clustering Algorithm											
	Story_based News Clustering (Turkish)			News Clustering (English)			News Clustering (German)			News Clustering (Spanish)		
	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1
G_{SE}	0.773	0.793	0.783	0.861	0.859	0.860	0.815	0.838	0.826	0.833	0.844	0.838
G_{SE+FS}	0.729	0.793	0.756	0.833	0.827	0.830	0.742	0.766	0.754	0.762	0.801	0.781
$G_{SE+NEXTENTITY}$	0.715	0.807	0.768	0.843	0.826	0.835	0.833	0.858	0.845	0.805	0.839	0.822
G_E	0.828	0.823	**0.825**	0.912	0.904	**0.908**	0.858	0.923	**0.889**	0.891	0.9505	**0.920**
	BIRCH Clustering Algorithm											
	Story_based News Clustering (Turkish)			News Clustering (English)			News Clustering (German)			News Clustering (Spanish)		
	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1
G_{SE}	0.841	0.808	0.822	0.921	0.917	0.919	0.859	0.914	0.886	0.872	0.925	0.898
G_{SE+FS}	0.804	0.852	0.815	0.942	0.874	0.907	0.820	0.849	0.834	0.844	0.894	0.869
$G_{SE+NEXTENTITY}$	0.825	0.850	0.834	0.919	0.924	0.922	0.829	0.883	0.855	0.838	0.935	0.884
G_E	0.892	0.890	**0.890**	0.957	0.941	**0.949**	0.931	0.940	**0.935**	0.949	0.923	**0.936**

TABLE 5. **The effect of Named Entity-Centered Frequent Subgraph Nodes on the clustering performance. The highest F1 score is highlighted in bold.**

$$(G_{SE+NEXTENTITY}) = (\{V_{SE}\}, \{E_{SE} \cup e_{NEXTENTITY(en_i, en_{i+1}|d_j)}\}) \quad (12)$$

- **Simple Event Graph with Named Entity-Centered Frequent Subgraph Nodes (Event Graph- G_E):** The $G_{SE+NEXTENTITY}$ graph model establishes a foundation by creating paths between named entities or time stamps, enabling the derivation of patterns they form. The Event Graph G_E proposed in this study can be defined as an updated variation of the G_{SE+FS} graph model, enhanced with frequent subgraphs obtained from $G_{SE+NEXTENTITY}$, as shown in Equation 13:

$$(G_{SE+NEXTENTITY}) = (\{V_{SE+FS}\}, \{E_{SE+FS} \cup e_{NEXTENTITY(en_i, en_{i+1}|d_j)}\}) \quad (13)$$

When examining the clustering performances on all datasets and with two different algorithms using the graph models designed to address RQ2, the following results were obtained:

- The G_{SE+FS} graph model is an expanded variation of the G_{SE} graph model, augmented with nodes representing patterns that frequently recur within the news text corpus. However, these nodes did not have a positive impact on the clustering performance.
- The $G_{SE+NEXTENTITY}$ graph model was designed to evaluate the contribution of adding edges between named entities and time nodes that do not have a connection between them in the simple event graph to the representational capability, independent of the nodes from frequent subgraphs. This graph model did not show any significant benefit in terms of enhancing the representational capability of the event in any of the datasets.
- In the news corpus, subgraphs formed by patterns that repeat in different news texts with exactly the same sequence of words have low support values and graph sizes. Upon examining the subgraphs with high support values (frequently recurring), it is understood that these patterns are not sequences of words that can distin-

	SBN (Turkish)	NC (English)	NC (German)	NC (Spanish)
G_{SE+FS}	1572	1817	1969	1293
G_E	1831	2003	2224	1494
#NER	259	186	255	201
Percentage	14.14%	9.28%	11.46%	13.45%

TABLE 6. **Counts of Named Entity-Centered Frequent Subgraphs in Datasets (without 2-step filtering). The values in the first row represent the numbers of Frequent Subgraphs (excluding Named-Entity Centered Frequent Subgraphs) obtained from the SBN (Turkish), NC (English), NC (German), and NC (Spanish) datasets without 2-step filtering. The second row shows the total number of Frequent Subgraphs, including Named-Entity Centered Frequent Subgraphs. By subtracting the values of the first row from those of the second row, the counts of Named-Entity Centered Frequent Subgraphs are obtained, which are presented in the third row. The last row indicates the percentage of Named-Entity Centered Frequent Subgraphs relative to the total number of Frequent Subgraphs.**

	Story_based News Clustering (Turkish)			News Clustering (English)			News Clustering (German)			News Clustering (Spanish)		
	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1
Miranda et al.* [14]	-	-	-	0.942	0.902	0.923	0.989	0.889	0.936	0.964	0.872	0.916
Staykovski et al.* [6]	-	-	-	0.951	0.936	0.944	-	-	-	-	-	-
Linger et al.* [13]	-	-	-	0.941	0.935	0.938	0.951	0.943	**0.946**	0.937	0.900	0.917
Saravanakumar et al.* [7]	-	-	-	0.942	0.952	0.947	-	-	-	-	-	-
Santos et al.* [30]	-	-	-	0.927	0.921	0.924	0.976	0.900	0.937	0.950	0.862	0.903
Ours	0.892	0.890	**0.890**	0.957	0.941	**0.949**	0.931	0.940	0.935	0.949	0.923	**0.936**

TABLE 7. Clustering performance comparison of the proposed method vs the SOTA methods on News Clustering dataset. The highest F1 score is highlighted in bold. * denotes results reported by Santos,et al. [30].

guish one event from another or show similarities rather, they are groups of words that are standard in the page structure of online news sources and/or are commonly repeated in all news from the same source.

- The G_E graph model had the highest performance compared to the other graph models across all datasets and the clustering algorithms used in the experiments, independent of language. In the experiments conducted with the BIRCH clustering algorithm, which is recommended for use in the clustering task in this study, a B-cubed F1 score of 89% was achieved on the Story-based News Clustering (Turkish) dataset, 94.9% on the News Clustering (English) dataset, 93.5% on the News Clustering (German) dataset, and 93.6% on the News Clustering (Spanish) dataset.
- The G_E graph model showed a significant performance increase compared to the G_{SE} graph model, with up to 8.2% improvement in the B-cubed F1 score for the HDBSCAN clustering algorithm, and up to 6.8% for the BIRCH clustering algorithm.
- As shown in Table 6, the proportion of frequently recurring subgraphs that are Named Entity-Centered Frequent Subgraphs constitutes 14.14% for the Story-based News Clustering (Turkish) dataset, 9.28% for the News Clustering (English) dataset, 11.46% for the News Clustering (German) dataset, and 13.45% for the News Clustering (Spanish) dataset. These density percentages are directly proportional to the improvement in the performance of traditional clustering algorithms according to the B-cubed F1 score when comparing the G_E and G_{SE} graph models. In particular, for the BIRCH Clustering Algorithm, there was an approximately 3% performance increase for the News Clustering (English) dataset, compared to approximately a 6.8% increase for the Story-Based News (Turkish) dataset.

3) RQ 3: CLUSTERING PERFORMANCE COMPARISON AGAINST THE SOTA METHODS

To answer the RQ3 research question, the clustering performance of the method proposed in this study was tested on the multilingual News Clustering (NC) benchmark dataset and compared with the performance of the most advanced (State-Of-The-Art, SOTA) methods tested on the same dataset in the literature. The News Clustering dataset contains news texts from nine languages. However, there are training sets for English, German, Spanish, and Chinese (the training set for Chinese consists of only 10 news items for a single story). As seen in Table 2, except for German, Spanish, and English, the test sets for all other languages contained between one and nine stories. Therefore, SOTA text clustering studies in the literature that use the NC dataset have conducted experiments on the test datasets of English [6], [7], [13], [14], [30], German, and Spanish [13], [14], [30] and presented clustering performance comparisons based on B-cubed performance metrics.

Table 7 presents the monolingual text clustering performances of the News Clustering dataset in the literature on test sets for English, German, and Spanish languages. In the last row of the table, the news clustering performance obtained using the Event Graph G_E model, which was proposed for use in news representation in this study, on the BIRCH clustering algorithm is presented in terms of the B-cubed evaluation metric.

Miranda et al. [14] represent news texts with a vector of various TF-IDF sub-vectors containing words, phrases, and named entities. Comparing TF-IDF-weighted sparse vector representations with doc2vec-based dense representations and their combinations, Staykovski et al. [6] address the news clustering problem by re-implementing the well-known NewsLens [67] approach and present evaluation results for the English language only. However, although NewsLens is one of the pioneering works in the research area of news clustering in particular, it has several critical limitations such as language dependency and scalability due to its batch processing methodology. Miranda et al. [14] and Staykovski et

al. [6] have argued for the inefficiency of dense features for clustering documents in the same language. In the literature, monolingual representations for each document have been widely used, where each document is represented by a TF-IDF-weighted bag-of-words sub-vector of entities and words in the title and content of each document. Following this work, Linger et al. [13] proposed an innovative extension of the NewsLens algorithm, with a particular focus on language dependency. In order to cluster news articles into local topics, they adopt the approach of constructing an undirected graph whose nodes represent news articles and whose edges are weighted by their similarity to each other. This similarity score is determined by the cosine similarities between each of the 9 TF-IDF-based representation sub-vectors of news articles. None of these approaches has the ability to exploit the characteristics of the story to which the event belongs, allowing for in-depth analysis of contextual and narrative aspects, which is a key contribution of our study. Contrary to claims in previous work, Saravanakumar et al. [7] empirically demonstrated that dense embeddings improve clustering performance when augmented with task-specific fine-tuning, external information, and a combination of sparse and temporal representations. Here they propose a clustering approach based on a non-parametric k-means algorithm and utilize a shallow neural network approach to make cluster formation decisions. In this study, which only presents evaluation results for the English language, the small dimensionality of the input space for the neural network, which takes the similarity values between the news text and the clusters as input, increases the risk of overfitting. Recent work by Santos et al. [30] utilizes multilingual document representations that overcome the dependency on the use of specific individual models for each language. Thus, unlike the approach proposed by Saravanakumar et al. [7], this solution does not remain confined to the English language. It should be noted that this paper does not analyse on the applicability of the proposed approach to low-resource languages. In the context of training multilingual document representations, it is important to consider the disadvantageous position of low-resource languages.

Unlike all the studies reviewed, in this study, instead of traditional text representation methods, a graph model is designed that includes the 'event' structure of news texts and is also capable of representing the features of the story to which the event belongs. As in other studies, a richly related event representation was created by utilizing not only the elements in the news text content, but also the explicit/hidden entities, time elements and frequent repetitive patterns in the texts. This representation is capable of capturing not only discrete events but also the story context to which the events belong. Furthermore, the effective performance of the language-independent nature of this approach on the clustering task is demonstrated by experiments on the English, German, and Spanish test sets of the News Clustering dataset. On the other hand, it also stands out in terms of its applicability to low-resource languages. The results are presented in Table 7. The proposed approach achieves state-of-the-art (SOTA) performance in English and Spanish and performs competitively in German.

E. LIMITATIONS

Named Entity-centered Frequent Subgraphs are patterns that frequently occur in event graphs and are specific to event items. These subgraphs can be obtained by introducing edges between named entities and temporal nodes in the event graphs. The number of subgraphs can vary depending on the order in which elements are connected. To address this issue, one approach is to add edges between all elements; however, this leads to an increase in the path length and time complexity of the random-walk algorithms. To overcome this limitation, an edge has been added from the "NEXT ENTITY" connection between timestamps, people, locations, and organizations, starting from the publication date that should be present in each news text. The order in which these elements are connected considers the frequency order of entities in the dataset.

The Story Based News Dataset consists of non-synthetic news texts from real news sources, and the number of events within each story varies depending on the complexity and significance of the story. For instance, the Great Manavgat fire in Turkey in 2021 lasted for over ten days, resulting in numerous news articles covering developments during and after the fire, totaling more than 60 events. However, the Artemis 1 space flight mission, which is part of the Artemis Program, generated approximately five new events over a span of two years. In real datasets with an imbalanced number of cluster elements, it is crucial to select the clustering algorithm and parameter values carefully. The BIRCH and HDBSCAN clustering algorithms, which are known for their performance on large and unbalanced datasets, were employed in this study. The optimal parameter values were chosen using a Grid Search. However, the cluster sizes within the dataset impose limitations on the selection of the algorithms and parameters.

Although the proposed methodology shows promise in various languages, it also presents some potential limitations or challenges in its adaptation to different linguistic contexts. The Event Graph Based News Clustering approach uses a methodology to represent news texts, generating a dense vector based on graph structure and spatial organisation of nodes in a language-independent way. However, a fundamental step requires the construction of a basic event graph in which important elements extracted from news texts are included as nodes. These important elements include named entities. Transformer-based approaches used for the Named Entity Recognition (NER) task are usually trained independently for each language using large datasets due to the complex nature of the models. However, pre-trained models for low-resource languages such as Turkish are available with less extensive data compared to languages such as English, which have undergone extensive academic scrutiny. As a result, NER detection, an important aspect of the proposed methodology, has the potential to yield different performance results across languages. Furthermore, languages characterised by

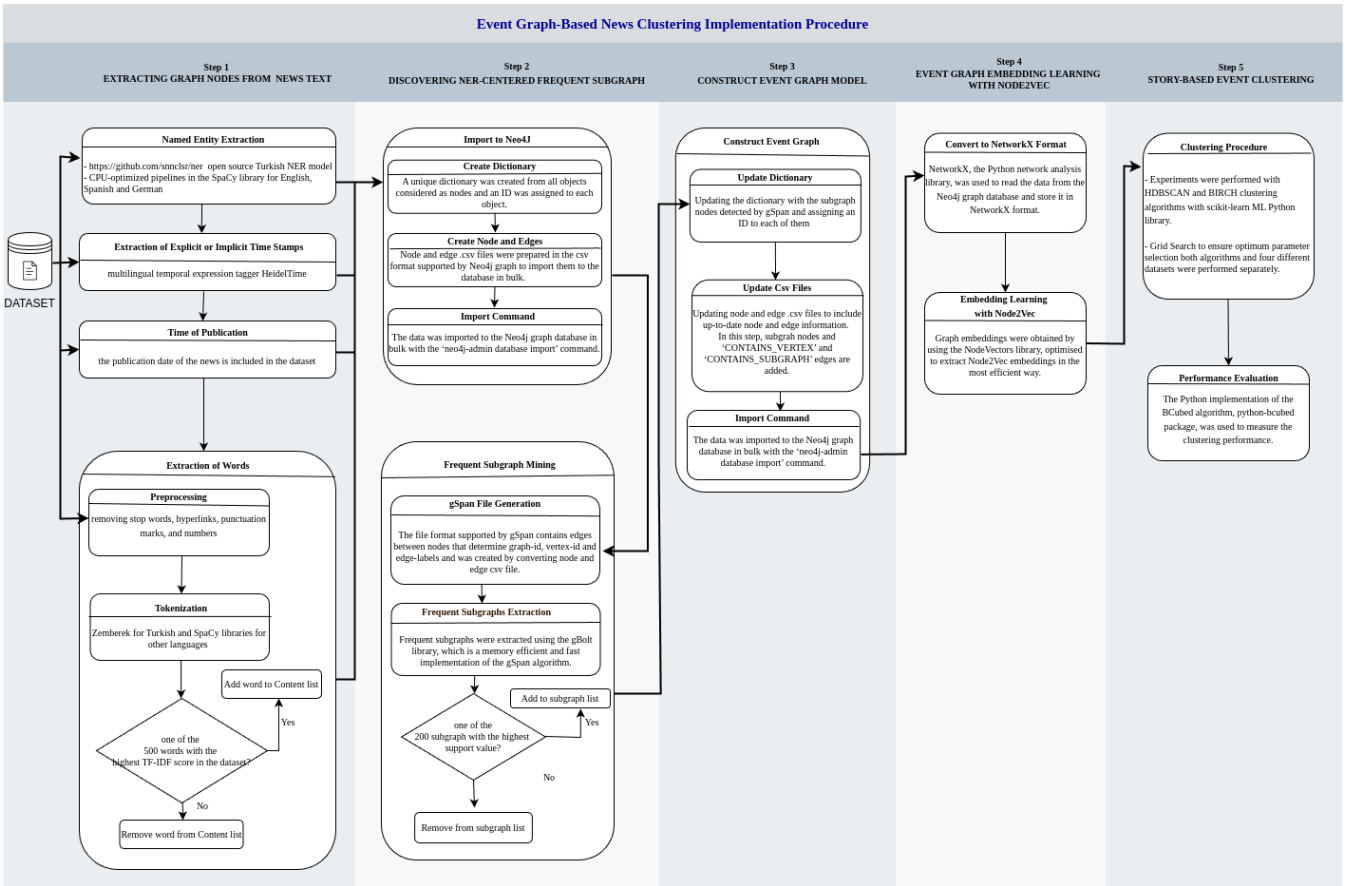

FIGURE 7. Event Graph-Based News Clustering Implementation Procedure

high morphological complexity, exemplified by the suffixes of Turkish, present numerous challenges, ranging from time tag detection to content tokenisation.

V. CONCLUSION

This paper presents a novel approach for detecting and modeling story clusters in a news text dataset. The main contribution is the construction of a density vector capable of representing not only the content of a news text, but also the characteristics of the story of which it is a part. To achieve this, a method was developed that learns the representation of event texts using an extended event graph model with named entity-centered frequent subgraphs. This method focuses on detecting patterns of frequently recurring named entities and timestamps in story clusters of events, and adding them as nodes to the event graph. The main goal is to increase the importance of the key elements that characterize a story in the event graphs constituting the story. In this manner, the similarity between event subsets of the same story at different time periods is increased, whereas the similarity between events from different stories is reduced. This feature vector, weighted with story-specific decomposable attributes, improves the performance of the traditional clustering algorithms. Node embeddings in the event network model were obtained using the skip-gram-based Node2Vec node embedding algorithm, a technique widely used in the literature. The embedding of each event is obtained by averaging the node embeddings of the nodes that they contain. Compared to other embedded text embedding models, event embeddings obtained by averaging Node2Vec node embeddings not only capture contextual relationships between news and event items, but also potentially include other links in the network. Overall, the proposed approach offers a new perspective for representing and clustering story clusters in news text datasets by leveraging the power of graph-based modeling and embedding techniques.

APPENDIX A
IMPLEMENTATION PROCEDURE

The methodology proposed in the study is illustrated in detail in Figure 7, which presents the implementation steps in the form of a flow diagram.

REFERENCES

[1] M. Stephens, *A History of News*. New York, USA: Oxford University Press, 2007.
[2] F. Hamborg, C. Breitinger and B. Gipp, "Giveme5w1h: A universal system for extracting main events from news articles,"2019, *arXiv:1909.02766*.
[3] R.E. Park, "News as a Form of Knowledge: A Chapter in the Soci-

ology," *American Journal of Sociology*, vol. 45, no. 5, pp. 669-686, 1940.[Online].https://www.jstor.org/stable/2770043.

[4] K. Xiao, Z. Qian and B. Qin, "A graphical decomposition and similarity measurement approach for topic detection from online news," *Information Sciences*, vol.570, pp. 262-277, 2021.

[5] J. Allan, J. Carbonell, H. Doddington, J. Yamron and Y. Yang, "Topic detection and tracking pilot study: Final report," in *Proc. Broadcast News Transcription Understand Workshop (DARPA)*, 1998, pp. 194–218.

[6] T. Staykovski, A. Barrón-Cedeño, G. Da San Martino and P. Nakov, "Dense vs. Sparse Representations for News Stream Clustering," in *Proc. Text2Story@ ECIR*, New York, USA, 2019.

[7] K. K. Saravanakumar, M. Ballesteros, M.K. Chandrasekaran and K. McKeown, "Event-driven news stream clustering using entity-aware contextual embeddings," in *Proc. 16th Conf. Eur. Ch*, Kiev, Ukraine, 2021.

[8] X. Hu, W. Ma, C. Chen, S. Wen, J. Zhang, Y. Xiang, and G. Fei, "Event detection in online social network: Methodologies, state-of-art, and evolution," *Computer Science Review*, vol. 46, p. 100500, 2022. DOI: 10.1016*.j.cosrev*.2022.100500, [Online].

[9] R.F. Cekinel and P. Karagoz, "Event prediction from news text using subgraph embedding and graph sequence mining," *World Wide Web*, vol. 25, no. 6, pp. 2403-2428, 2022. DOI: 10.1007*.s11280*-021-01002-1, [Online].

[10] P. Vossen, T. Caselli and Y. Kontzopoulou, "Storylines for structuring massive streams of news," in *Proc. First Workshop on Computing News Storylines*, ,pp. 40-49, 2015.

[11] B. Keith, M. Horning and T. Mitra, "Evaluating the inverted pyramid structure through automatic 5w1h extraction and summarization," *Computational Journalism C+ J*, 2020.

[12] N. Vanetik, M. Litvak and E. Levi, "Real-World Events Discovering with TWIST," *Natural Language Processing for Electronic Design Automation*, pp. 71-107, 2020.

[13] M. Linger and M. Hajaiej, "Batch Clustering for Multilingual News Streaming," in *Proc. Text2Story@ ECIR 2020*, Lisbon, Portugal, 2020.

[14] S. Miranda, A. Znotins, S.B. Cohen and G. Barzdins, "Multilingual Clustering of Streaming News," in *Proc. 2018 Conference on Empirical Methods in Natural Language Processing*,Brussels, Belgium, 2018.

[15] Y. Zhang, R. Jin and Z. H. Zhou, "Understanding bag-of-words model: A statistical framework," *Int. J. Mach. Learn. Cybern.*, vol. 1, pp. 43-52, 2010.

[16] J. Rupnik, A. Muhic, G. Leban, P. Skraba, B. Fortuna and M. Grobelnik, "News across languages-cross-lingual document similarity and event tracking," *Journal of Artificial Intelligence Research*, vol. 55, pp. 283-316, 2016.

[17] K. Xiao, Z. Qian and B. Qin, "A Survey of Data Representation for Multi-Modality Event Detection and Evolution," *Applied Sciences*, vol. 12, 2022.

[18] C. Li, M. Liu, J. Cai, Y. Yu and H. Wang, "Topic detection and tracking based on windowed DBSCAN and parallel KNN," *IEEE Access*, vol. 9, pp. 3858-3870, 2020.

[19] G. Xu, Y. Meng, Z. Chen, X. Qiu and C. Wang, "Research on topic detection and tracking for online news texts," *IEEE Access*, vol. 7, pp. 58407-58418, 2019.

[20] W. Liu, L. Jiang, Y. Wu, T. Tang and W. Li, "Topic detection and tracking based on event ontology," *IEEE Access*, vol. 8, pp. 98044-98056, 2020.

[21] G. Frisoni, G. Moro, G. Carlassare and A. Carbonaro, "Unsupervised event graph representation and similarity learning on biomedical literature," *Sensors*,vol. 22, 2021.

[22] Y. Lin, Z. Miao, M. Ni, H. Jiang, Wang C., J. Gao, L. Ji and G. Shi, "Online topic detection and tracking system and its application on stock market in china," in *Proc. Information Retrieval: 26th China Conference*,Xi'an, China, 2020.

[23] P. Rosso, "Tracking News Stories in Short Messages in the Era of Infodemic," in *Proc. Experimental IR Meets Multilinguality, Multimodality, and Interaction: 13th International Conference of the CLEF Association*, vol. 13390, p. 18, 2022.

[24] N. Mamo, J. Azzopardi, C. Layfield,"An automatic participant detection framework for event tracking on Twitter," *Algorithms*, vol. 14, no. 3, 2021.

[25] Y. Chen, L. Wu and M. Zaki, "Iterative deep graph learning for graph neural networks: Better and robust node embeddings," *Advances in neural information processing systems*,vol. 33, pp. 19314–19326,2020.

[26] I. A. Chikwendu, X. Zhang, I. O. Agyemang, I. Adjei-Mensah, U. C. Chima and C.J. Ejiyi, "A Comprehensive Survey on Deep Graph Representation Learning Methods," *Journal of Artificial Intelligence Research*,vol. 78, pp. 287–356,2023.

[27] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space,"2013,*arXiv:1301.3781.*

[28] Q. Du, N. Li, W. Liu, D. Sun, S. Yang and F. Yue, "A Topic Recognition Method of News Text Based on Word Embedding Enhancement," *Computational Intelligence and Neuroscience*, 2022.

[29] L. Hu, S. Yu, B. Wu, C. Shao and X. Li, "A neural model for joint event detection and prediction," *Neurocomputing*, vol. 407, pp. 376-384, 2020.

[30] J. Santos, A. Mendes and S. Miranda, "Simplifying Multilingual News Clustering Through Projection From a Shared Space,"2022,*arXiv:2204.13418.*

[31] I. Gusev and I. Smurov, "Russian News Clustering and Headline Selection Shared Task,"2021,*arXiv:2105.00981.*

[32] D. Hu, D. Feng and Y. Xie, "EGC: A novel event-oriented graph clustering framework for social media text," *Information Processing and Management*, vol. 59, no 6, 2022.

[33] B. N. dos Santos, R. G. Rossi, S. O. Rezende and R. M. Marcacini, "A two-stage regularization framework for heterogeneous event networks," *Pattern Recognition Letters*, vol. 138, 2020.

[34] J. P. R. Mattos and R. M. Marcacini, "Semi-supervised graph attention networks for event representation learning," in *Proc. IEEE International Conference on Data Mining (ICDM)*, Auckland, New Zealand, 2021.

[35] L. Mu, P. Jin, J. Zhao and E. Chen, "Detecting evolutionary stages of events on social media: A graph-kernel-based approach," *Future Generation Computer Systems*,vol. 123, pp. 219-232, 2021.

[36] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang and S.Y. Philip, "A comprehensive survey on graph neural networks," *IEEE transactions on neural networks and learning systems*, vol. 32,no. 1, pp. 4-24, 2020.

[37] A. Agibetov, "Neural graph embeddings as explicit low-rank matrix factorization for link prediction," *Pattern Recognition*, vol. 133, 2023.

[38] J. Qiu, Y. Dong, H. Ma, J. Li, C. Wang, K. Wang and J. Tang, "Netsmf: Large-scale network embedding as sparse matrix factorization," in *Proc. The World Wide Web Conference*,San Francisco, USA,2019.

[39] A. Grover and J. Leskovec, "node2vec: Scalable feature learning for networks," in *Proc. 22nd ACM SIGKDD International Conference On Knowledge Discovery And Data Mining*,California, USA, 2016.

[40] B. Perozzi, R. Al-Rfou and S. Skiena, "Deepwalk: Online learning of social representations," in *Proc. 20th ACM SIGKDD International Conference On Knowledge Discovery And Data Mining*,New York, USA, 2014.

[41] Hamilton, W. L., "Graph representation learning," *Synth. Lectures Artif. Intell. Mach. Learn.*,vol. 14, no. 3, pp. 1–159,2020.

[42] X. Liu, T. Murata, K.S. Kim, C. Kotarasu and C. Zhuang, "A general view for network embedding as matrix factorization," in *Proc. Twelfth ACM international conference on web search and data mining*, Melbourne, Australia, 2019.

[43] S.N. Mohammed and S. Gündüç, "Degree-based random walk approach for graph embedding," *Turkish Journal of Electrical Engineering and Computer Sciences*, vol. 30,no. 5, pp. 1868-1881, 2022.

[44] Y. Ma, L. Zong, Y. Yang and J. Su, "News2vec: News network embedding with subnode information," in *Proc. 2019 conference on empirical methods in Natural Language Processing and the 9th International Joint Conference on natural language processing (EMNLP-IJCNLP)*, Hong Kong, China, 2019.

[45] Q. Zeng, M. Li, T. Lai, H. Ji, M. Bansal and H. Tong, "Gene: Global event network embedding," in *Proc. 15th Workshop on Graph-Based Methods for Natural Language Processing (TextGraphs-15)*,Mexico City, Mexico, 2021.

[46] S. Deng, H. Rangwala and Y. Ning, "Learning dynamic context graphs for predicting social events," in *Proc. 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*,Anchorage, USA, 2019.

[47] M. H. Weng, S. Wu and M. Dyer, "Identification and Visualization of Key Topics in Scientific Publications with Transformer-Based Language Models and Document Clustering Methods," *Applied Sciences*, vol. 12,no. 21, pp. 11220, 2022.

[48] dbmdz Turkish bert model. Accessed: Oct. 19,2022. [Online]. Available: https://huggingface.co/dbmdz/bert-base-turkish-cased

[49] spaCy English Trained Pipelines. Accessed: Nov. 5,2022. [Online]. Available: https://spacy.io/models/en#en_core_web_lg

[50] spaCy German Trained Pipelines. Accessed: Nov. 05,2022. [Online]. Available: https://spacy.io/models/de#de_core_news_lg

[51] spaCy Spanish Trained Pipelines. Accessed: Nov. 11,2022. [Online]. Available: https://spacy.io/models/es#es_core_news_lg

[52] J. Strötgen and M. Gertz, "A Baseline Temporal Tagger for all Languages," in *Proc. 2015 Conference on Empirical Methods in Natural Language Processing*,Lisbon, Portugal, 2015.

[53] H. Camci and G. Eryiğit, "Türkçe Zamansal İfadelerin Yakalanması ve Tanımlanması,"*Bilişim Teknolojileri Dergisi*, vol. 14,no. 3, pp. 337-343, 2021.
[54] H. Genc and B. Yilmaz, "Text-Based Event Detection: Deciphering Date Information using Graph Embeddings," in *Proc. International Conference on Big Data Analytics and Knowledge Discovery*,Linz, UA, Austria, 2019.
[55] X. Yan and J. Han, "gspan: Graph-based substructure pattern mining," in *Proc. 2002 IEEE International Conference on Data Mining*,NW Washington DC, USA, 2002.
[56] A. Bagga and B. Baldwin, "Algorithms for scoring coreference chains," in *Proc. First International Conference on Language Resources And Evaluation Workshop On Linguistics Coreference*,Granada, Spain, 1998.
[57] J. Baley, "A comparison of extrinsic clustering evaluation metrics based on formal constraints"*Cahiers de Linguistique Asie Orientale*, vol. 52,no. 2, pp. 137-162, 2023.
[58] R. van Heusden, J. Kamps and M. Marx, "BCubed Revisited: Elements Like Me," in *Proc. ACM SIGIR International Conference on Theory of Information Retrieval*,Madrid, Spain, 2022.
[59] G. Stewart and M. Al-Khassaweneh, "An implementation of the HDBSCAN* clustering algorithm," *Applied Sciences*,vol. 12,no. 5, 2022.
[60] R. J. Campello, D. Moulavi and J. Sander, "Density-based clustering based on hierarchical density estimates," in *Proc. Pacific-Asia conference on knowledge discovery and data mining*,Berlin, Germany, 2013.
[61] D. Wang, Y. Huang and Z. Cai, "A two-phase clustering approach for traffic accident black spots identification: integrated GIS-based processing and HDBSCAN model," *International journal of injury control and safety promotion*,vol. 30,no. 2, 2023.
[62] T. Zhang, R. amakrishnan and M. Livny, "BIRCH: A new data clustering algorithm and its applications," *Data mining and knowledge discovery*,vol. 1, 1997.
[63] A. Lang and E.Schubert, "BETULA: Fast clustering of large data with improved BIRCH CF-Trees," *Information Systems*,vol. 108, 2022.
[64] B. Lorbeer, A. Kosareva, B. Deva, D. Softić, P. Ruppel and A. Küpper, "Variations on the clustering algorithm BIRCH," *Information Systems*,vol. 11, 2018.
[65] F. G. Cuevas, H. Schmid, "aMax-B-CUBED: A Supervised Metric for Addressing Completeness and Uncertainty in Cluster Evaluation,", 2023.
[66] N. Agarwal, G. Sikka and L. K. Awasthi, "Evaluation of web service clustering using Dirichlet Multinomial Mixture model based approach for Dimensionality Reduction in service representation," *Information Processing and Management*,vol. 57, 2020.
[67] P. Laban and M.A. Hearst, "newsLens: building and visualizing long-ranging news stories," in *Proceedings of the Events and Stories in the News Workshop*,Vancouver, Canada, 2017.

BASAK BULUZ KOMECOGLU was born in Istanbul, Türkiye in 1990. She received the B.S. degree in mathematics-computer and computer engineering by completing the double major program from Istanbul Aydin University, Istanbul, Türkiye in 2014. She completed her master's degree in computer engineering at Gebze Technical University, Kocaeli, Türkiye in 2018 with the thesis study "Graph Mining Approach in Modeling Academic Success". She has been continuing Ph.D. at Gebze Technical University in computer engineering since 2018.

From 2014 to 2020, she was a lecturer at Istanbul Aydın University Anadolu Bil Vocational School, Istanbul, Türkiye and also worked as the Program Head of the Computer Programming (English) Program for 5 years. Since 2020, she is a research assistant at Gebze Technical University, Institute of Information Technologies. Her research interests include machine learning, deep learning, text mining, data mining, natural language processing and graph mining and its applications.

Ms. Buluz Kömeçoğlu and her team were honoured the first place award in the national Teknofest 2021 Aerial Object Detection Challenge in Istanbul, Türkiye with their data-centric object detection technique.

BURCU YILMAZ received her Ph.D. degree from the Department of Computer Engineering at Gebze Institute of Technology, located in Kocaeli, Türkiye, in 2010. During her Ph.D. studies, she worked as a research assistant at Istanbul Kultur University in Istanbul, Türkiye. Since 2014, she has been serving as an Assistant Professor in the Institute of Information Technologies at Gebze Technical University in Kocaeli, Türkiye. She has publications in internationally recognized and indexed international journals and has presented papers at international conferences. Dr. Yilmaz is actively conducting research in data mining, machine learning, natural language processing, deep learning, graph mining, graph neural networks, and social network analysis.