RESEARCH PAPER
Contrib. Sci. & Tech Eng, 2026, 3(2)
Homepage: https://cste.journals.umz.ac.ir/
DOI: 10.22080/cste.2025.29964.1081

Content-Based Event Detection on Twitter: A Survey of NLP Techniques for Sub-Event Prediction and Evolution

Khalil Kolaee Darabi [1], Hamid Hassanpour [2]*, Amir Sheikhahmadi [3]

[1] *Ph.D. Candidate, Department of Computer Engineering, Sa.C., Islamic Azad University, Sanandaj, Iran.*

[2] *Professor, Faculty of Computer Engineering, Shahrood University of Technology, Iran.*

[3] *Associate Professor, Department of Computer Engineering, Sa.C., Islamic Azad University, Sanandaj, Iran.*

Article Info
Received 31 August 2025
Accepted 19 November 2025
Available online 23 May 2026

Keywords:
Event Detection;
Deep Learning;
Sub-event Prediction;
Social Media Analysis;
Natural Language Processing;
Content-based Methods.

Abstract:

The rapid growth of Twitter has transformed it into a critical real-time sensor for world events, often surfacing information about disasters, political upheavals, and public health crises ahead of traditional sources. While detecting major events is valuable, the ability to identify sub-events—fine-grained, evolving components—is crucial for deeper situational awareness. This survey provides a comprehensive review of NLP techniques for sub-event prediction and evolution on Twitter. We introduce a novel taxonomy that categorizes methods from traditional text-based and graph-based approaches to modern deep learning and transformer-based architectures, specifically evaluating their capacity to capture sub-event dynamics. Our analysis covers widely used benchmarks (e.g., CrisisNLP, CrisisBench, COVID-Twitter datasets) and evaluation protocols (e.g., precision, recall, F1, clustering metrics). The findings indicate that while significant advances have been made, the fusion of multimodal data, the application of large language models, and the adoption of privacy-preserving frameworks such as federated learning represent the most promising pathways to robust sub-event detection. By synthesizing methodological advances and evaluation practices, this paper underscores the central role of sub-event analysis in advancing research and its critical importance for real-time, high-stakes applications in disaster response, public health, and security.

***Corresponding Author:** h.hassanpour@shahroodut.ac.ir

Supplementary information: Supplementary information for this article is available at https://cste.journals.umz.ac.ir/

Please cite this paper as: Hassanpour, H., Kolaee, K., & Sheikhahmadi, A. (2026). Content-Based Event Detection on X (Twitter): A Survey of NLP Techniques for Sub-Event Prediction and Evolution. Contributions of Science and Technology for Engineering, 3(2), 11-28. doi:10.22080/cste.2025.29964.1081.

1. Introduction

Event detection on Twitter (now X) involves automatically identifying real-world events from ongoing streams of text data. An event is formally characterized as a cluster of temporally and semantically coherent tweets that reflect a specific occurrence [1, 2]. However, real-world events are rarely monolithic; they typically comprise multiple sub-events—smaller, evolving components representing specific phases, localized incidents, or thematic facets of the larger occurrence. The significance of detecting these sub-events extends far beyond academic interest, as it fundamentally transforms our capacity to achieve granular situational awareness.

This enables a nuanced response to crises, precise monitoring of public health progression, and a detailed understanding of societal dynamics. Twitter-based systems have demonstrated the ability not only to issue earthquake alerts but to track aftershocks and damage patterns [3], not only to detect disease outbreaks but to identify variant-specific waves and superspreader events days before official hospital reports [4], and to provide critical early warnings for the evolution of civil unrest and natural disasters [5, 6].

However, the computational challenges inherent in sub-event detection are multifaceted and demanding. The platform generates approximately 500 million tweets daily [2], each limited to 280 characters yet rich in linguistic complexity through hashtags, mentions, and emojis. This brevity fosters a unique lexicon of abbreviations, slang, and contextual references that consistently challenge traditional Natural Language Processing (NLP) techniques, especially when trying to discern the subtle linguistic cues that differentiate one sub-event from another [7, 8].

Furthermore, the high velocity of information flow necessitates real-time processing capabilities to track event evolution, while the platform's global nature introduces significant multilingual and cross-cultural complexities [3]

*ISSN **3060-6578***

[9]. Twitter's dual role as an information medium presents both opportunities and obstacles for sub-event tracking. Its real-time nature enables rapid information propagation, with significant events and their sub-components often trending within minutes of occurrence [10]. However, this immediacy also introduces substantial noise from spam, misinformation, and off-topic content, obscuring genuine sub-event signals [11, 12]. Similarly, while the social graph structure—follower relationships, retweets, and mentions—provides valuable signals for understanding diffusion patterns of sub-event information [10, 13], it also complicates disambiguation between organic viral content and artificially amplified messages.

A diverse and evolving set of methodologies has been developed to address these challenges. These approaches can be broadly categorized into several groups: text-based approaches, which analyze linguistic patterns to extract key terms and entities that signal specific sub-events [2, 10]; graph-based approaches, which leverage the social network structure to uncover patterns of information diffusion and community interactions related to event phases [5, 9]; time-series and machine learning methods, which apply statistical and learning algorithms to identify temporal patterns and predict the evolution of sub-events [1, 14]; and deep learning-based approaches, which utilize neural networks to model complex, nonlinear relationships that define sub-event sequences within large-scale data [15, 16].

The development and comparison of these methods rely heavily on standardized benchmark datasets—such as CrisisNLP, TREC, and domain-specific corpora (e.g., COVID-19 Twitter datasets)—and a suite of evaluation metrics including precision, recall, F1-score, and clustering measures, forming the essential framework for assessing sub-event segmentation accuracy and detection timeliness.

Recent technological advances have dramatically expanded the methodological landscape. The emergence of transformer-based language models (e.g., BERT and its variants) enables sophisticated contextual understanding of short-form text, vital for distinguishing similar sub-events [17-19], while progress in graph neural networks facilitates richer exploitation of Twitter's network structure for identifying evolving community roles during events [20]. Furthermore, multimodal approaches that integrate textual data with images and videos have gained traction for their potential to provide richer contextual signals for verifying and characterizing sub-events [21, 22].

Despite these significant advancements, a clear research gap remains. Existing methods predominantly focus on identifying major events without adequately addressing the complex dynamics of their constituent sub-events. Current approaches often struggle to capture the multi-scale, time-varying nature of sub-events effectively, and the potential of recent advances such as large language models (LLMs) and federated learning (FL) for this task remains underexplored.

This survey addresses this gap by providing a comprehensive review and classification of content-based event detection methods, with a specific focus on their capacity for sub-event prediction and evolution. We examine methodologies across the spectrum, from traditional keyword-based detection to state-of-the-art deep learning and LLM applications [15, 16] to emerging privacy-preserving FL techniques [14, 22]. Our analysis is structured around a clear taxonomy that categorizes existing approaches, facilitates comparison, and evaluates their strengths and limitations in capturing the intricate dynamics of sub-events. Particular attention is given to challenges and opportunities in modeling the progression and relationships between sub-events, highlighting this as the next frontier in advancing event detection research for real-time, high-stakes applications.

2. A Comprehensive Taxonomy of Event Detection Methods

A comprehensive taxonomy of event detection methods provides a systematic classification of techniques for identifying and describing important events across multiple data types, including text, images, and sensor data. This framework details the core principles, algorithmic strategies, and domain-specific adaptations of event detection methods, encompassing traditional rule-based systems, statistical models, and modern machine learning and deep learning approaches. By providing a clear overview, the taxonomy enables detailed comparisons, guides the selection of methods for specific tasks, and highlights common challenges and research opportunities in event detection. The taxonomy categorizes existing sub-event detection methods into six main groups: text-based, graph-based, time-series, machine-learning, deep-learning, and emerging techniques. The overall framework is illustrated in Figure 1, and the Conceptual Workflow of Sub-Event Detection Methods is shown in Figure 2.

2.1. Text-Based Approaches

Traditional text-based event detection methods form the foundation of Twitter event detection research, leveraging linguistic signals to identify unusual patterns indicative of emerging events and their constituent sub-events. These approaches assume that significant events and their finer-grained components manifest as sudden changes in term frequency distributions, topic compositions, or linguistic patterns within Twitter streams [2, 9].

For sub-event detection, early systems that rely on TF-IDF (Term Frequency-Inverse Document Frequency) provide a basic capability for identifying keywords associated with specific event phases. By monitoring temporal shifts in term frequencies [23], these systems can flag sudden spikes not only for main events but also for terminology signaling new sub-events (e.g., a shift from "fire" to "evacuation" within a disaster event).

A pivotal advancement is Kleinberg's burst detection algorithm, which models keyword frequency patterns as a state-based process [10]. Its value for sub-event detection lies in identifying the precise onset of distinct activity phases. However, its initial formulation had limitations. Multi-level burst detection extensions are crucial, enabling detection of both short-lived, high-intensity sub-events (e.g., explosions) and prolonged, lower-intensity sub-events

(e.g., recovery efforts) within the same overarching event [10], providing a multi-scale view of event evolution.

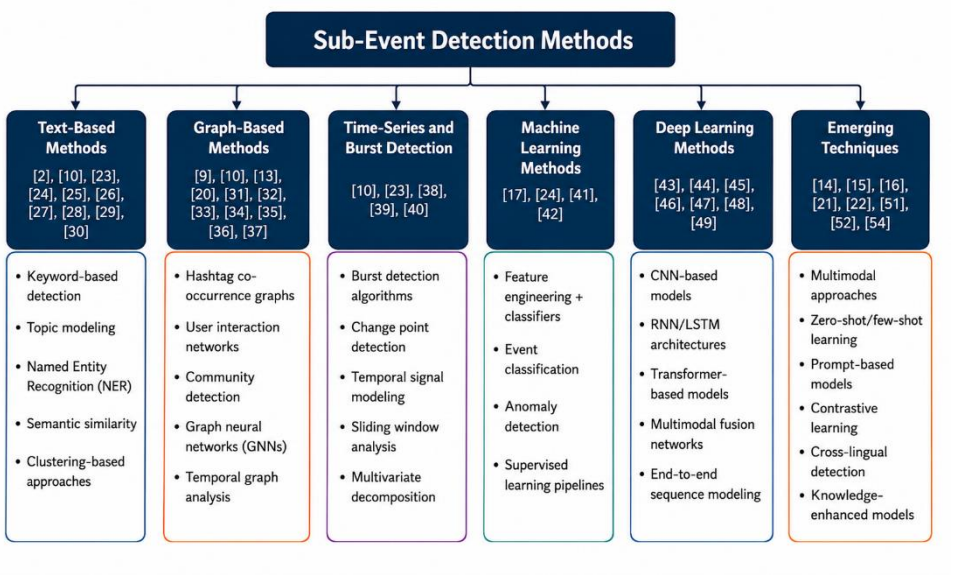

Figure 1. A taxonomy of content-based methods for sub-event detection on Twitter. The diagram categorizes approaches into six primary groups based on their core data processing principles: Text-Based, Graph-Based, Time-Series, Traditional Machine Learning, Deep Learning, and Emerging Techniques

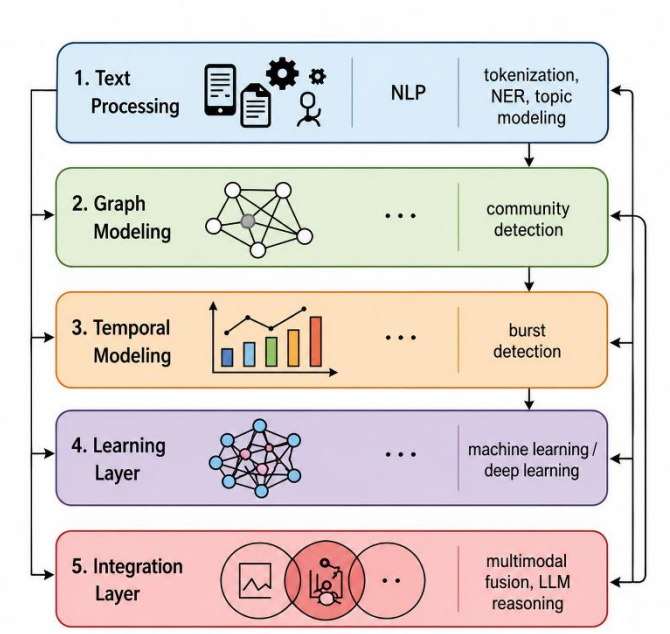

Figure 2. A conceptual workflow for a generalized sub-event detection system

Beyond keyword-centric methods, Latent Dirichlet Allocation (LDA) and its variants are cornerstone techniques for unsupervised discovery of thematic sub-events [2, 24]. By tracking the temporal evolution of topic distributions [25], these models identify the emergence, peak, and decline of specific sub-themes within larger events. This is enhanced by temporal adaptations like Dynamic Topic Models (DTM) and Continuous Time Dynamic Topic Models (cDTM), which explicitly model topic evolution over time [26, 27], making them directly applicable for mapping sub-event progression. Furthermore, online topic modeling techniques (e.g., Online LDA, Online TwitterLDA [28, 29]) are essential for incrementally updating event models as new sub-events unfold in real time, allowing systems to adapt to evolving narratives.

Named Entity Recognition (NER) serves a critical role in contextualizing and distinguishing sub-events [30]. Identifying specific entities (e.g., locations, persons, organizations) enables systems to progress from merely detecting that something is happening to understanding what is happening, where, and to whom. This is fundamental for isolating discrete incidents within larger events—for example, differentiating sub-events affecting

different locations or involving distinct key figures. Transformer-based NER models fine-tuned on social media text [30] Are key enabling technologies for this granular analysis.

2.2. Graph-Based Approaches

While text-based methods analyze what is being discussed, graph-based methods leverage Twitter's inherent network structure to detect events and their sub-events by analyzing how information propagates through social interactions and community dynamics [10, 13, 20]. These approaches recognize that significant events and their evolving phases often manifest through textual changes and measurable shifts in network behavior and information flow. Twitter's social graph provides rich structural signals for tracking event progression, with different network representations offering complementary insights into sub-event activity.

Community detection algorithms such as the Louvain method and Leiden algorithm are valuable for sub-event detection. They identify densely connected user groups whose communication patterns may indicate coordination around specific sub-events [10]. Sudden changes in community structure—such as new clusters or bridges between previously disconnected groups—often correlate with major events and the emergence of new sub-event phases that disrupt typical social boundaries. Similarly, centrality measures (degree, betweenness, PageRank) help identify influential users whose behavioral changes may signal emerging sub-events. Temporal monitoring of these metrics reveals how sub-events propagate through networks and which users act as key information sources or amplifiers for specific phases [31-33].

Retweet networks provide explicit traces of information diffusion, enabling the detection of viral content patterns characteristic of specific sub-events [10]. Event detection systems analyze these networks using metrics such as cascade size, propagation speed, and structural virality to distinguish organic sharing from event-driven information cascades linked to different sub-events. Mention networks similarly reveal communication patterns that often intensify during sub-events, with sudden increases in network density or novel interaction patterns serving as reliable indicators of new phases requiring coordinated discussion among previously unconnected users.

Dynamic network analysis techniques are particularly crucial for sub-event detection, as they track the temporal evolution of Twitter's social structure through graph similarity metrics. These approaches detect sub-events by identifying significant deviations from baseline network dynamics, capturing moments when real-world occurrences alter typical interaction patterns at the sub-event level. More recently, graph neural networks (GNNs) have emerged as cutting-edge tools for temporal network analysis [20], with variants like Dynamic Graph Neural Networks (DGNNs) and Temporal Graph Networks (TGNs) explicitly modeling temporal dependencies to learn evolving node and edge representations that capture local changes and global structural shifts indicative of sub-event progression and transitions.

Complementing these approaches, graph embedding techniques such as Node2Vec generate low-dimensional vector representations of network elements that preserve structural properties while enabling efficient machine-learning analysis [34, 35]. These embeddings—learned using methods such as temporal random walks—capture user behavior patterns, community membership, and local topology in a compact form. For sub-event detection, systems monitor changes in these embeddings over time, with sudden shifts in vector space often corresponding to behavioral changes associated with transitions between sub-events or participation in specific phases. Advanced variants incorporate dynamic embedding updates to better track evolving user roles during event propagation across sub-events [36, 37], offering powerful tools for identifying event occurrence and participant engagement through network geometry.

2.3. Time-Series and Burst Detection Methods

Complementing the structural view of graph-based methods, time-series analysis offers a purely temporal lens, providing powerful techniques to identify statistical patterns and anomalies in Twitter data streams that signal event occurrence and evolution [10, 23]. These methods are crucial for sub-event detection, providing the quantitative foundation for temporally segmenting continuous events into constituent phases by detecting statistical anomalies in metrics like keyword frequencies, posting rates, sentiment scores, and network activity.

The Kleinberg burst detection algorithm remains seminal for identifying discrete sub-event onsets, modeling keyword occurrences through a two-state Hidden Markov Model, and distinguishing baseline and elevated usage periods [38]. It identifies bursts by computing the most likely state sequence from observed frequencies, effectively capturing sudden surges in terminology specific to new sub-event phases. More sophisticated extensions employ multivariate time-series analysis, using Vector Autoregressive (VAR) models to capture dependencies between related terms or Principal Component Analysis (PCA) to identify dominant patterns in high-dimensional keyword spaces [39, 40]. These multivariate methods are essential for detecting coordinated shifts across multiple keywords characterizing new sub-events, beyond isolated term spikes.

Wavelet transforms enhance temporal analysis by decomposing signals across multiple timescales simultaneously [10], making them uniquely powerful for sub-event detection. The continuous wavelet transform reveals evolving periodic patterns, enabling detection of short-term spikes (abrupt sub-events such as explosions) and sustained trends (evolving sub-events such as recovery efforts) within the same data stream. This multi-resolution analysis is fundamental for understanding complete event lifecycles composed of multiple sub-events.

To address Twitter's inherent temporal patterns, seasonal decomposition techniques like STL (Seasonal-Trend decomposition using Loess) separate the time series into trend, seasonal, and residual components [23]. By analyzing residual irregular components, event detection systems gain

improved sensitivity to genuine anomalies while filtering predictable fluctuations. This is valuable for distinguishing routine activity spikes from meaningful sub-event signals, ensuring detected anomalies reflect actual deviations rather than periodic behaviors.

Change point detection algorithms are arguably the most direct statistical tools for sub-event detection, identifying statistically significant shifts in time-series properties [23]. Bayesian methods model data as piecewise stationary processes, with change points marking regime transitions. These often correspond precisely to boundaries between sub-events (e.g., a shift from a crisis event to a humanitarian response). For real-time monitoring, online variants like the CUSUM (Cumulative Sum) algorithm continuously test for distributional changes. They detect sub-event onsets through sudden shifts in mean activity or variance while processing streaming data, enabling immediate identification of new phases as they emerge. A comparative overview of sub-event detection approaches, including their strengths, weaknesses, and typical applications, is summarized in Table 1.

Table 1. A comparative analysis of sub-event detection methodologies, summarizing their core principles, key techniques, primary strengths, and inherent limitations

Method Category	Strengths	Weaknesses	Typical Use Cases	Key References
Text-Based	Simple, interpretable; effective for keyword bursts and topic shifts	Limited context understanding; sensitive to slang and abbreviations	Disaster phase tracking, protest theme evolution	[2, 10, 23, 24]
Graph-Based	Captures diffusion and community dynamics; robust to noise	Requires network data; complex modeling	Viral sub-event tracing, misinformation detection	[9, 10, 20, 31-33]
Time-Series	Precise temporal segmentation; anomaly detection	Sensitive to noise; lacks semantic depth	Sub-event onset detection, phase transitions	[10, 23, 38-40]
Machine Learning	Fast training; interpretable; handles structured features well	Requires feature engineering; limited generalization	Sub-event classification, early warning systems	[17, 24, 41, 42]
Deep Learning	Captures complex patterns; handles short text and multimodal inputs	Data-hungry; less interpretable	Multimodal sub-event detection, sentiment shifts	[43-49]
Emerging Techniques	Privacy-preserving; adaptable to novel sub-events	Computationally intensive; still maturing	Real-time crisis monitoring, federated health alerts	[14-16, 21, 22, 50-52]

2.4. Machine Learning and Deep Learning Approaches

Building upon the foundational signals identified by text, graph, and time-series methods, modern event detection systems increasingly leverage machine learning and deep learning techniques to uncover complex, non-linear patterns that traditional statistical methods often miss [17, 18, 41]. These approaches are particularly valuable for sub-event detection, as they can learn the nuanced patterns that distinguish between different phases and facets of a larger event.

Classical supervised methods, such as Support Vector Machines (SVMs), are effective for binary event classification. They perform well with high-dimensional features such as TF-IDF vectors, linguistic attributes, and metadata. A key advantage is their robust performance even with limited training data [17, 41]. For complex multi-class tasks—such as classifying specific sub-event types—ensemble methods, such as Random Forests, are often used. These provide interpretable models by combining decision trees and generating feature importance metrics. Gradient boosted variants (XGBoost, LightGBM) excel at handling heterogeneous features that combine textual, temporal, and network attributes [17, 42].

In unsupervised settings, clustering algorithms offer valuable alternatives for sub-event discovery without predefined labels. DBSCAN stands out for its ability to identify irregularly shaped clusters while handling noise. This is a critical feature for discovering emerging sub-events in messy Twitter data [9, 24]. While simpler K-means variants require careful parameter tuning, hierarchical clustering methods are especially relevant. They reveal the nested structure of events and sub-events through their dendrogram outputs, providing a multi-resolution view of event composition.

Deep learning has brought significant advances for modeling sub-event sequences and context. Convolutional Neural Networks (CNNs) extract local n-gram patterns via multi-scale filters operating on word embeddings. Text-CNN variants are optimized for short text classification [43, 44] and are effective at capturing key phrases indicative of sub-events. Recurrent architectures such as LSTMs and their bidirectional variants capture sequential dependencies in tweets over time [44–46], making them suitable for modeling sub-event evolution. GRUs offer similar performance at reduced computational cost. Attention mechanisms further enhance these models by focusing on relevant text segments, which helps characterize different sub-events.

The state-of-the-art resides in transformer architectures, particularly BERT and its variants. These models leverage self-attention to build rich contextual representations, accounting for word sense disambiguation and syntax [17, 18, 47]. This capability is crucial for distinguishing similar sub-events that share much of the same vocabulary but differ subtly in context. Domain-adapted models such as BERTweet and COVID-Twitter-BERT achieve superior performance through specialized pretraining on Twitter

corpora [48, 49]. RoBERTa's optimized training also enhances capabilities across multiple benchmarks [19, 47].

These transformers are often combined with ensemble strategies. These strategies aggregate multiple model predictions via voting or stacking [18], thereby creating robust detection systems capable of handling Twitter's diverse sub-event content.

The multimodal nature of modern Twitter content has spurred the development of integrated systems. These systems combine textual analysis with visual feature extraction using CNNs pretrained on large image datasets (ResNet, VGG) [21]. They employ various fusion strategies to merge information from different modalities, which helps verify and characterize sub-events more comprehensively.

Simultaneously, Graph Neural Networks (GNNs) have emerged as powerful tools for incorporating social network structure into sub-event detection. Graph Convolutional Networks learn user representations from neighborhood information. Scalable solutions like GraphSAGE enable the analysis of Twitter's massive networks [20]. Temporal extensions of these architectures (DynGCN, EvolveGCN) prove particularly valuable for tracking evolving events. They do this by modeling dynamic changes in network structure and user behavior over time [20], which enables the detection of how sub-events propagate and evolve through social networks.

2.5. Multimodal and Transformer-Based Approaches

Multimodal approaches leverage multiple data types available on social media—mainly textual, visual (images, videos), and sometimes audio—to enhance event detection beyond the limitations of unimodal methods. By integrating complementary information across modalities, these methods can better verify and characterize events and sub-events. Recent studies employ convolutional neural networks (CNNs), pretrained on large image datasets such as ResNet or VGG, to extract visual features, while transformers or CNNs handle text data. Interaction mechanisms such as attention enable the capture of deep semantic correlations between text and images, improving robustness to noisy, ambiguous, or misleading single-modal signals. For example, multimodal event memory networks have demonstrated improved accuracy in false information detection by combining text embeddings from BERT and image features from VGG-19 with attention-based feature fusion, outperforming previous models by about 7% in accuracy [53].

Transformer-based approaches, especially LLMs like BERT and its Twitter-specialized variants (e.g., BERTweet, COVID-Twitter-BERT), have revolutionized text-based event detection by modeling contextual word representations and disambiguating meanings in short, noisy social media posts. These models utilize self-attention mechanisms to capture relevant semantic and syntactic relationships, which is critical for distinguishing highly similar sub-events. Moreover, modern transformer architectures extend to multimodal fusion by jointly learning from text and visual inputs, leading to state-of-the-art results in social media event and misinformation detection. Generative multimodal models, including GPT-4o and LLaVA, have also been explored, showing promise in handling complex social media language phenomena such as leet speak and text elongation. However, supervised models currently lead in precision [50, 51].

Efforts to benchmark multimodal LLMs on social media reveal performance variability, highlighting ongoing challenges in understanding social context, detecting misinformation, and interpreting affective communication. Novel graph neural networks that incorporate transformer principles further enhance event detection by modeling network interactions and temporal dynamics among users and posts [52, 54].

In summary, each category within the taxonomy offers distinct advantages and faces specific limitations for sub-event detection. Text-based approaches are highly interpretable and foundational, but can struggle with semantic nuance without advanced models. Graph-based methods excel at capturing information diffusion and social dynamics but are less effective for content understanding in isolation. Time-series techniques provide robust, quantitative temporal segmentation but may lack contextual depth. Traditional machine learning offers a good balance of performance and interpretability for specific classification tasks, while deep learning methods, particularly transformers and multimodal systems, achieve state-of-the-art performance by capturing complex, contextual patterns across data types, albeit at the cost of computational complexity and reduced interpretability. This comparative landscape, summarized in Table 1, underscores that the most effective sub-event detection systems often hybridize these approaches, leveraging their complementary strengths to overcome their individual weaknesses.

3. Datasets and Evaluation Metrics

3.1. Common Datasets

Evaluating Twitter event detection systems requires high-quality, annotated datasets that capture diverse events and temporal patterns. Several benchmark datasets have become community standards, each with unique characteristics and challenges [2, 55]. Table 2 summarizes these widely used datasets and their characteristics.

3.2. CrisisNLP and Crisis-Related Datasets

CrisisNLP is among the most comprehensive collections of crisis-related Twitter data, aggregating labeled tweets from multiple natural disasters such as earthquakes, floods, hurricanes, and wildfires [55]. The dataset spans several years and diverse geographic regions, providing rich examples of disaster-related communication patterns. Annotations cover event relevance, information type classifications (e.g., casualties, infrastructure damage, resource needs), and humanitarian categories.

.

Table 2. Overview of benchmark datasets commonly used for evaluating sub-event detection on Twitter

Dataset Name	Size (Tweets)	Event Types Covered	Annotation Scheme	Key References
CrisisNLP	~50K–200K	Earthquakes, floods, wildfires	Relevance, humanitarian categories	[55]
CrisisBench	~300K+	Multi-crisis	Binary/multi-class labels, standardized protocols	[55]
TREC TS	~100K	Disasters, accidents, planned events	Time-stamped tweets, newswire ground truth	[2]
COVID-Twitter	Millions	Pandemic phases, misinformation	Sentiment, misinformation, health topic categories	[12, 48, 49, 56]
AIDR	Varies	Real-time crises	Rapid annotations, latency metrics	[51]
WNUT Shared Tasks	~10K–50K	COVID-19, noisy text	Informativeness, fine-grained event categorization	[48, 49, 56],
Traffic Events	~20K–100K	Accidents, congestion	Location, timestamp, severity	[57]
Political Events	~50K+	Elections, protests	Stance, demographic metadata	[58, 59]

Building on CrisisNLP, CrisisBench provides standardized evaluation protocols and additional annotations across multiple crises [50]. This benchmark enables fair comparison of detection methods through binary and multi-class classification tasks.

Additionally, the Artificial Intelligence for Disaster Response (AIDR) datasets provide real-time, rapidly annotated crisis event collections [51], enabling system evaluation under operational constraints with metrics for detection latency and temporal accuracy

3.3. TREC Temporal Summarization Datasets

The TREC Temporal Summarization track supplies time-stamped tweet streams aligned with newswire ground truth, enabling evaluation of both detection accuracy and temporal precision [2]. These datasets provide precise event timestamps derived from authoritative news sources, supporting latency-based metrics that measure the delay between actual event occurrences and system detections.

TREC datasets encompass diverse event types, including natural disasters, accidents, and planned events, and are carefully annotated with event boundaries and evolutionary phases. Including non-event periods allows evaluation of false favorable rates under realistic operational conditions.

3.4. COVID-19 and Pandemic-Related Datasets

The COVID-19 pandemic generated unprecedented volumes of health-related social media content, leading to the creation of multiple specialized datasets for pandemic event detection [12, 48, 49]. These collections include multilingual content, diverse geographic regions, and temporal coverage spanning different phases of the pandemic response.

COVID-Twitter datasets provide keyword-filtered and manually annotated subsets, enabling evaluation of different sampling strategies and annotation approaches. Specialized annotations include misinformation labels, sentiment classifications, and health topic categories.

The WNUT (Workshop on Noisy User-generated Text) shared tasks have produced several COVID-related datasets with standardized evaluation protocols, including functions for identifying informative tweets and fine-grained event categorization [48, 49, 56].

3.5. Specialized Domain Datasets

Traffic event datasets focus on transportation-related incidents, providing location-specific events with clear temporal boundaries [57]. These datasets often include geographic metadata and integration with official traffic incident reports for ground truth validation.

Sports event datasets capture planned events with predictable temporal patterns, enabling the evaluation of systems on different event characteristics compared to crisis or breaking news scenarios [24]. These datasets support analysis of how event detection approaches generalize across event types with different temporal and linguistic characteristics.

Political event datasets include elections, protests, and policy announcements, often incorporating geographic and demographic metadata that enable analysis of event-detection performance across different populations and regions [58, 59].

3.6. Synthetic and Simulation-Based Datasets

Synthetic datasets generated through data augmentation and simulation techniques address the scarcity of labeled examples for specific event types [5]. These approaches employ strategies such as paraphrasing, entity substitution, and temporal shifting to generate additional training examples while preserving essential event characteristics.

Simulation-based datasets model the information propagation process during events, generating synthetic Twitter streams that exhibit realistic temporal and network dynamics. These datasets enable controlled evaluation of system performance under varying network conditions and event evolution patterns.

3.7. Evaluation Metrics

3.7.1.Traditional Classification Metrics

Precision, recall, and F1-score are fundamental for evaluating event detection, especially in binary classification [2, 56]. Precision is the fraction of correct detections, recall is the fraction of actual events successfully detected, and F1 is their harmonic mean:

$$\text{Precision} = \frac{\text{TP}}{(\text{TP}+\text{FP})} \quad (1)$$

$$\text{Recall} = \frac{\text{TP}}{(\text{TP}+\text{FN})} \quad (2)$$

$$\text{F1} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (3)$$

where TP represents True Positives, the number of correctly detected events that were actually relevant; FP represents False Positives, the number of events detected by the system that were not relevant. FN represents False Negatives, the number of relevant events not detected by the system.

Macro-averaged and micro-averaged versions of these metrics provide different perspectives on system performance across multiple event types. Macro-averaging computes metrics separately for each event type before averaging, giving equal weight to rare and common event types:

$$\text{MacroPrecision} = \frac{1}{\text{N}} \sum_{\text{i}=1}^{\text{N}} \frac{\text{TP}_\text{i}}{(\text{TP}_\text{i} + \text{FP}_\text{i})} \quad (4)$$

$$\text{MacroRecall} = \frac{1}{\text{N}} \sum_{\text{i}=1}^{\text{N}} \frac{\text{TP}_\text{i}}{(\text{TP}_\text{i} + \text{FN}_\text{i})} \quad (5)$$

$$\text{MacroF1} = \frac{1}{N} \sum_{i=1}^{N} \frac{\text{Precision}_\text{i} \times \text{Recall}_\text{i}}{\text{Precision}_\text{i} + \text{Recall}_\text{i}} \quad (6)$$

where N is the total number of event types; TP_i, FP_i and, FN_i represent the true positives, false positives, and false negatives for event type i, respectively, $\text{Precision}_\text{i}$ and Recall_i represent the precision and recall for event type i.

Micro-averaging aggregates predictions across all event types before computing metrics, emphasizing performance on frequent event types:

$$\text{MicroPrecision} = \frac{\sum_{\text{i}=1}^{\text{N}} \text{TP}_\text{i}}{\sum_{\text{i}=1}^{\text{N}} \text{TP}_\text{i} + \text{FP}_\text{i}} \quad (7)$$

$$\text{MicroRecall} = \frac{\sum_{\text{i}=1}^{\text{N}} \text{TP}_\text{i}}{\sum_{\text{i}=1}^{\text{N}} \text{TP}_\text{i} + \text{FN}_\text{i}} \quad (8)$$

$$\text{MicroF1} = 2 \times \frac{\text{Micro}-\text{Precision} \times \text{Micro}-\text{Recall}}{\text{Micro}-\text{Precision} + \text{Micro}-\text{Recall}} \quad (9)$$

Multi-class extensions of precision and recall support evaluation of fine-grained event categorization tasks, where systems must detect events and classify them into specific types. Confusion matrices provide a detailed analysis of classification errors, revealing which event types are most frequently confused with each other. The confusion matrix is represented as follows:

$$\text{Confusion Matrix} = \begin{vmatrix} TP & FP \\ FN & TN \end{vmatrix} \quad (10)$$

Precision, recall, and the F1-score are core metrics for evaluating event detection systems. They balance detection accuracy and completeness. Macro-averaging treats all event types equally. In contrast, micro-averaging prioritizes frequent events. Together, they offer distinct insights into system performance. These metrics are often analyzed alongside confusion matrices. Confusion matrices highlight specific misclassifications between event types. This combined analysis provides a comprehensive view of a system's strengths and weaknesses. It is essential for guiding targeted improvements in detection and classification.

3.7.2. Clustering and Information Retrieval Metrics

Event clustering tasks require specialized evaluation metrics that assess the quality of tweet groupings rather than individual classification decisions [60]. Normalized Mutual Information (NMI) measures the mutual dependence between predicted and ground truth clusterings, with values ranging from 0 (independent clusterings) to 1 (identical clusterings):

$$NMI = \frac{I(U,V)}{\sqrt{H(U).H(V)}} \quad (11)$$

Here, $I(U,V)$ represents the mutual information between the predicted clustering U and the ground truth clustering V, while $H(U)$ and $H(V)$ are the entropies of the predicted and ground truth clusterings, respectively.

Adjusted Rand Index (ARI) provides an alternative clustering evaluation metric for chance agreement between clusterings. ARI values range from -1 to 1, with higher values indicating better clustering quality. The metric's adjustment for chance agreement makes it more suitable for comparing clusterings with different numbers of clusters. The formula for ARI is:

$$ARI = \frac{\sum_{ij}\binom{n_{ij}}{2} - \frac{\left[\sum_i \binom{n_{i+}}{2} \sum_j \binom{n_{+j}}{2}\right]}{\binom{N}{2}}}{\frac{\left[\sum_i \binom{n_{i+}}{2} + \sum_j \binom{n_{+j}}{2}\right]}{2} - \frac{\left[\sum_i \binom{n_{i+}}{2} \sum_j \binom{n_{+j}}{2}\right]}{\binom{N}{2}}} \quad (12)$$

In this formula, n_{ij} is the number of items in both the $i-th$ predicted cluster and the $j-th$ ground truth cluster, n_{i+} and n_{+j} represent the sizes of the predicted and ground truth clusters, respectively, and N is the total number of items. The binomial coefficients $\binom{n_{ij}}{2}$ compute pairwise combinations within clusters.

Purity and inverse purity metrics evaluate clustering quality from complementary perspectives. Purity measures the extent to which clusters contain tweets from a single event, while inverse purity assesses whether tweets from individual events are grouped into a single cluster. The formula for Purity is:

$$\text{Purity} = \frac{1}{\text{N}} \sum_{\text{i}=1}^{\text{k}} \max_{\text{j}} \left|\text{C}_\text{i} \cap \text{G}_\text{j}\right| \quad (13)$$

Here, C_i represents the $i-th$ predicted cluster, G_j represents the $j-th$ ground truth cluster, and N is the total number of tweets. The cardinality $|C_j \cap G_j|$ denotes the number of tweets in both the $i-th$ predicted cluster and the $j-th$ ground truth cluster.

The formula for Inverse Purity 14 is similar to Purity, but focuses on maximizing the overlap between ground truth clusters and predicted clusters for each ground truth cluster G_j.

$$\text{Inverse Purity} = \frac{1}{\text{N}} \sum_{\text{j}=1}^{\text{m}} \max_{i} \left|\text{C}_\text{i} \cap \text{G}_\text{j}\right| \quad (14)$$

The harmonic mean of purity and inverse purity 15 provides a balanced assessment similar to the F1-score for classification tasks. The formula for the harmonic mean is:

$$\text{Harmonic Mean} = 2 \times \frac{Purity \times Inverse\ Purity}{Purity + Inverse\ Purity} \quad (15)$$

Here, the harmonic mean combines the Purity and Inverse Purity values into a single metric that balances both perspectives.

3.7.3. Temporal and Latency Metrics

Real-time event detection systems require evaluation metrics that assess not only detection accuracy but also temporal precision and response latency [2, 9]. Detection latency is the time between the occurrence of an event and its detection, typically measured in minutes or hours, depending on the event type and application requirements. This can be formulated as:

$$\text{Detection Latency} = t_{detection} - t_{occurrence} \quad (16)$$

where $t_{detection}$ is the timestamp when the system detects the event, and $t_{occurrence}$ is the timestamp when the event actually occurs.

Temporal precision metrics evaluate the accuracy of estimated event timestamps, measuring the deviation between predicted and actual event onset times. These metrics are particularly important for applications that require precise temporal localization, such as disaster response and emergency management systems. This can be expressed as:

$$\text{Temporal Precision} = \left| t_{predicted} - t_{actual} \right| \quad (17)$$

where $t_{predicted}$ is the predicted timestamp of the event's onset, and t_{actual} is the actual timestamp of the event's onset.

Event duration estimation accuracy assesses systems' ability to predict the temporal extent of events, measuring the difference between expected and actual event durations. This metric becomes crucial for resource allocation and response planning applications that depend on accurate predictions of event timelines. This can be formulated as:

$$\text{Event Duration Accuracy} = \left| \Delta t_{predicted} - \Delta t_{actual} \right| \quad (18)$$

where $\Delta t_{predicted}$ is the predicted duration of the event, and Δt_{actual} is the actual duration of the event.

3.7.4. Coverage and Completeness Metrics

Event coverage metrics evaluate the comprehensiveness of detection systems across different event types, geographic regions, and temporal periods [2]. These metrics assess whether systems exhibit systematic biases toward particular event characteristics or demonstrate consistent performance across diverse conditions.

Geographic coverage analysis examines detection performance across different spatial regions, identifying potential biases toward urban areas with higher social media activity or specific linguistic communities. Temporal coverage metrics assess performance consistency across various periods, identifying potential degradation during high-activity periods or specific temporal patterns.

Population coverage metrics evaluate whether event detection systems provide equitable performance across different demographic groups and communities. These metrics address algorithmic bias concerns and ensure detection systems effectively serve diverse populations.

$$\text{Coverage Metric} = f\begin{pmatrix} Event\ Types, \\ Geographic\ Regions, \\ Temporal\ Periods, \\ Demographic\ Groups \end{pmatrix} \quad (19)$$

In this formula, $Event\ Types$ represents the different categories or kinds of events being detected by the system, such as natural disasters, political events, or social movements; $Geographic\ Regions$ refers to the spatial areas under consideration, which may include urban, rural, or specific linguistic or cultural regions that might influence detection performance; $Temporal\ Periods$ denotes the time frames over which the system's performance is evaluated, identifying if performance changes during high-activity periods, such as holidays or crises; $Demographic\ Groups$ represents the various population groups, categorized by age, gender, ethnicity, or socioeconomic status, to ensure the detection system serves them fairly and equitably.

3.7.5. Multimodal Evaluation Approaches

Multimodal event detection systems require specialized evaluation frameworks that assess performance across different modalities and their integration [21]. Visual-textual alignment metrics evaluate the consistency between textual descriptions and accompanying images, identifying cases where visual and textual information provide complementary or contradictory evidence. The formula for this metric is:

$$\text{VTA} = \frac{1}{\text{N}} \sum_{\text{i}=1}^{\text{N}} sim(T_i, V_i) \quad (20)$$

where VTA is the visual-textual alignment metric, measuring how well the textual and visual information align; T_i is the textual description of the $i-th$ event; V_i is the corresponding visual content for the $i-th$ event; $sim(T_i, V_i)$ is the similarity function between the text T_i and visual content V_i quantifying the degree of alignment; N is the total number of events being evaluated.

Cross-modal retrieval metrics assess systems' ability to identify relevant content across different modalities, such as finding images related to textual event descriptions or identifying tweets that describe visual content. The formula for this metric is:

$$\text{CMR} = \frac{1}{\text{M}} \sum_{\text{i}=1}^{\text{M}} rank(T_i, V_i) \quad (21)$$

where T_i is the textual event description for the $i-th$ retrieval task; V_i is the visual content related to the event in

the $i-th$ retrieval task; $rank(T_i, V_i)$ represents the rank of the retrieval of relevant content from the modality T_i to modality V_i, reflecting how well the system identifies the most relevant content; M is the total number of retrieval tasks being evaluated.

These metrics are particularly important for applications requiring comprehensive multimodal understanding. Modality contribution analysis evaluates the relative importance of different information sources for event-detection decisions, identifying cases in which specific modalities provide crucial information that others cannot capture. The formula for this analysis is:

$$MCA = \frac{w_T}{w_T + w_V} \times 100 \quad (22)$$

where w_T is the weight of the textual modality in the event detection system, quantifying its contribution to the overall detection process; w_V is the weight of the visual modality in the event detection system, quantifying its contribution to the overall detection process. This analysis guides system design decisions regarding modality weighting and fusion strategies.

3.7.6. Emerging Evaluation Metrics Beyond Classical Measures in Machine Learning

Recent advances in machine learning have revealed limitations in traditional evaluation metrics such as precision, recall, and F1-score, especially when applied to complex or large-scale problems. To address these challenges, several emerging metrics and evaluation frameworks have been introduced to provide a more comprehensive and context-sensitive assessment of model performance.

Probabilistic extensions of these classical metrics incorporate classifier confidence scores rather than relying on binary decision thresholds. Metrics such as confidence-Precision (cPrecision), confidence-Recall (cRecall), and confidence-F1 (cF1) capture prediction uncertainty and provide robust performance estimates in real-world applications [61].

Ranking-based metrics such as the Area Under the Precision-Recall Curve (AUPR), Area Under the ROC Curve (AUC), and Normalized Discounted Cumulative Gain (NDCG) have gained popularity for evaluating models on imbalanced datasets or for cases requiring fine-grained ranking. These metrics assess how well a model prioritizes true positive instances and discriminates among prediction confidences [62].

For unsupervised tasks such as clustering and event detection, traditional classification metrics are insufficient. Measures like Normalized Mutual Information (NMI), Adjusted Rand Index (ARI), and Purity quantify cluster quality and help evaluate alignment against labeled groupings [63]. Temporal evaluation metrics have also been proposed for real-time detection applications to measure detection latency, temporal precision, and duration accuracy, emphasizing the timeliness as well as accuracy of system outputs [63]. Coverage and completeness metrics assess model fairness across various dimensions such as event types, geographic regions, and demographic groups, helping to identify and mitigate bias [64].

In multimodal settings, where information is drawn from text, images, and other sources, composite metrics like Visual-Textual Alignment (VTA) and Cross-Modal Retrieval (CMR) evaluate how effectively multi-source data contribute to performance [63]. Finally, the integration of uncertainty quantification and interpretability metrics enables more reliable, nuanced evaluation by explicitly modeling the confidence and error risk associated with the metrics [65, 66]. Together, these emerging evaluation approaches provide a richer understanding of model behavior, enabling more reliable, fair, and effective deployment of modern AI systems.

3.8. Quantitative Comparison of Recent Models

To provide an analytical overview of methodological progress, Table 3 summarizes representative sub-event detection models from 2021–2025, highlighting datasets, feature modalities, model types, and their reported F1 or accuracy scores. Performance differences result from variations in dataset size, event domain, and evaluation metrics, but overall trends show that transformer-based and multimodal architectures consistently outperform earlier baselines.

Table 3. Performance comparison of representative sub-event detection models across methodological eras

Year	Model/ Study	Dataset	Modality	Approach	Reported F1 / Accuracy	Key Contribution
2021	Kleinberg + LDA Hybrid [5]	CrisisNLP	Text	Statistical + Topic	0.71 F1	Early hybrid of burst & topic modeling
2022	DBSCAN + GloVe [24]	TREC TS	Text	Unsupervised Clustering	0.76 F1	Noise-resistant unsupervised sub-event grouping
2023	GCN–BERT Fusion [20]	COVID-Twitter	Text + Graph	GNN + Transformer	0.84 F1	Structural + semantic fusion
2023	Multimodal Event Memory Net [50]	CrisisBench	Text + Image	CNN + BERT Fusion	0.88 Accuracy	Attention-based multimodal fusion
2024	LLaMA-EventDet [59]	Multi-domain	Text	LLM few-shot	0.90 F1	Prompt-based cross-event generalization
2024	Federated BERTweet [22]	Health Social Media	Text (Private)	FL + Transformer	0.87 Accuracy	Privacy-preserving multi-source learning
2025	Hyperbolic Graph Model [20]	Multi-Crisis	Text + Graph	HGNN	0.89 F1	Temporal community evolution tracking

4. Applications of Sub-Event Detection and Evolution

Within the broader scope of event detection, sub-event detection focuses on identifying finer-grained, often sequential components that make up a larger event. A sub-event, as defined earlier, is a significant occurrence that is temporally contained within and semantically part of a larger parent event, representing a specific phase, localized incident, or thematic aspect of the overall happening. Detecting and tracking these sub-events—such as aftershocks following an earthquake, variant-specific waves during a pandemic, or tactical shifts within a social movement—provides a dynamic, high-resolution view of event progression that is crucial for effective response and analysis.

This granular capability has fundamentally transformed disaster response and emergency management. Twitter-based systems now excel at tracking not only the main shock of an earthquake but also aftershocks and localized damage patterns through keyword clustering and spatial analysis [3]. In hydrological disasters, sub-event detection monitors flood progression street-by-street, using location-specific reports to identify the most critically affected areas [5]. For wildfires, analysis extends beyond ignition to track smoke-plume dispersion and detect new flare-ups, with user-reported air-quality changes providing real-time insights into evolving health risks [3, 6]. These systems leverage NLP and geographic clustering to model the progression of disasters through constituent phases.

In public health surveillance, sub-event evolution offers unparalleled detail. Twitter analysis moves beyond outbreak detection to identify variant-specific COVID-19 waves via emerging symptom clusters [4, 67]. Frameworks like SPEED detect superspreader events and localized flare-ups as critical sub-events within larger pandemics through network propagation patterns [4]. Additionally, mental health surveillance tracks the progression of depressive episodes or escalation of suicidal ideation via linguistic shifts, enabling timely interventions at specific crisis points [46, 68]. Public health behavior monitoring identifies sub-populations driving vaccine hesitancy as sub-events within broader campaigns, facilitating dynamic, targeted strategies [17, 47].

This paradigm is equally vital for political and social monitoring, deconstructing major events into defining sub-events. Protest detection systems analyze how movements evolve by identifying tactical shifts and splinter group formation through evolving hashtag networks [58, 69]. Election monitoring tracks sub-events such as the impact of individual or disinformation campaigns on public sentiment [58, 59]. Policy impact analysis is refined to monitor demographic-specific reactions as sub-events within broader public responses [59], while misinformation analysis tracks narrative pivots, treating each as a detectable sub-event [11, 70].

Finally, commercial applications leverage sub-event detection for a high-resolution view of market dynamics. A brand crisis is decomposed into a sequence of sub-events: initial defect reports, complaint spikes, and influencer reactions [18]. Market trend detection identifies micro-trends and geo-concentrated demand spikes within broader movements. Competitive intelligence focuses on detecting sub-events like supply chain disruptions via subtle shifts in employee social media activity—often preceding public announcements [18]. This approach provides a dynamic and granular measure of the business landscape.

5. Challenges and Open Problems in Sub-Event Detections

Detecting sub-events—the constituent phases and facets of larger events—introduces challenges beyond general event detection. These problems center on the need for finer granularity, more precise temporal segmentation, and the ability to disambiguate closely related occurrences within a noisy, fast-moving data stream.

5.1. Noise and Misinformation

The prevalence of noise and misinformation in Twitter data represents one of the most significant challenges facing sub-event detection systems [11, 12, 70]. For sub-events, this problem is exacerbated, as misinformation can specifically target and amplify a phase of an event (e.g., fabricating a detail of a disaster) or create an entirely fabricated sub-event within a real occurrence, making it extremely difficult to distinguish genuine, granular reports from false or misleading information [11, 12]. Sophisticated misinformation campaigns and bot networks [70] can artificially create the illusion of a sub-event's evolution or suppress reports of a real one, directly polluting the data needed to model an event's true progression. Developing robust, real-time frameworks for assessing information quality at this granular level, potentially through multi-dimensional source credibility modeling and consensus-based validation that accounts for information cascades, remains an open challenge [12].

5.2. Multimodal Fusion for Granularity

The increasing prevalence of multimedia content is both a necessity and a challenge for sub-event detection. Truly understanding a sub-event—such as the specific damage from an aftershock or the exact water level on a street—often requires integrating text with images and videos [21]. However, current multimodal systems struggle with the semantic alignment of different modalities at this fine-grained level; a text description of a specific sub-event may be paired with a generic or repurposed image, leading to misidentification [21]. Furthermore, analyzing user-generated visual content for sub-event evidence requires robust computer vision techniques to handle poor quality, unusual angles, and irrelevant backgrounds to verify specific, transient details. Effectively combining these inconsistent, noisy multimodal signals in real time to build a coherent picture of sub-event evolution is a significant unsolved problem.

5.3. Ethical and Privacy Concerns at Scale

The use of social media data for sub-event detection raises significant ethical and privacy concerns, intensified by the granular nature of the analysis [17, 22, 71]. Tracking the evolution of an event involves continuous, detailed monitoring of conversations, which can inadvertently expose sensitive personal information or group dynamics that users never intended to be analyzed by automated systems [14, 72]. Furthermore, the algorithmic biases inherent in these systems can be more damaging at the sub-event level; geographic bias may mean specific neighborhoods within a disaster zone are overlooked, and demographic bias could cause the concerns of a subgroup within a protest to be systematically suppressed [73]. There is a serious risk that sub-event detection capabilities, designed for beneficial purposes like crisis response, could be repurposed for detailed surveillance of specific populations or to suppress the emergence of dissenting sub-movements [14]. Ensuring fairness, transparency, and privacy through techniques like FL [14, 22, 71] is therefore paramount.

5.4. Scalability and Real-Time Processing for Evolution

The computational challenges of scale and velocity are magnified for sub-event detection [2, 9, 41]. Systems must process millions of tweets per minute and maintain complex stateful models of an event's ongoing evolution, tracking multiple, potentially overlapping sub-events in parallel with low latency. This requires distributed stream processing architectures [9, 10] and online learning algorithms [9] that can update models incrementally without expensive batch retraining, allowing the system to adapt as a sub-event unfolds. The core open problem is achieving this at scale: performing precise, granular, and multimodal analysis under extreme computational constraints and near-zero latency requirements to provide timely insights into an event's progression.

6. Future Directions for Sub-Event Detection

While event detection has matured as a field, the task of sub-event detection and evolution introduces distinct challenges and opportunities that warrant dedicated research. To move beyond coarse-grained event identification, future work should focus on developing methods to identify, characterize, and link the constituent phases of larger occurrences. The following directions outline promising pathways toward this goal, shifting attention from whether an event is happening to how it unfolds in detail [74-77].

6.1. Large Language Models for Granular Understanding

The emergence of LLMs such as GPT-4 and LLaMA presents significant opportunities to advance sub-event detection [15, 16, 78, 79]. Their strong contextual reasoning is particularly valuable for interpreting the abbreviated and ambiguous language often used to describe specific event phases on Twitter [15, 16]. Moreover, their few-shot and zero-shot learning capabilities enable rapid adaptation to novel or emerging sub-event types without extensive retraining, which is essential given the diversity of potential event facets [59].

With carefully designed prompts [79], LLMs could also reason about temporal dependencies and causal relationships between sub-events [78], supporting the automatic construction of semantic causal graphs that map how one sub-event (e.g., a building collapse) triggers another (e.g., a rescue operation). This would move beyond simple identification toward modeling the narrative structure of an event's evolution. In addition, LLMs could generate detailed, multi-perspective summaries for each event phase [80], giving analysts coherent narratives of sub-event progression.

To illustrate, consider a hurricane event. An LLM could be tasked with analyzing a stream of tweets to identify sub-events. It could distinguish between "preparations" (e.g., tweets about boarding up windows, evacuations), "landfall impact" (e.g., reports of flooding, power outages), and "post-storm recovery" (e.g., requests for aid, reports of road closures). Furthermore, it could infer a causal link: "The mandatory evacuation order (sub-event A) was followed by a reduction in reports of civilian injuries during landfall (sub-event B)."

However, several challenges remain. Real-time analysis of high-volume event streams poses substantial computational demands [16]; hallucinations may fabricate entire sub-event sequences [70, 79]; and biases could cause models to underrepresent sub-events affecting marginalized communities [78]. Addressing these challenges will be critical for deploying LLMs responsibly in sub-event detection.

6.2. Privacy-Preserving Collaboration on Sub-Event Data

The fine-grained nature of sub-event data amplifies privacy concerns, as it often contains sensitive information about affected individuals and communities. FL offers a promising direction for addressing these concerns [14, 22, 71, 81]. By enabling multiple entities—such as emergency services, NGOs, or social media platforms—to collaboratively train models without sharing raw data, FL allows detection systems to benefit from diverse datasets while preserving individual privacy [82, 83].

Cross-silo FL between organizations [71] could produce powerful models trained on sub-event data from different jurisdictions and crisis types [84, 85]. To ensure robust privacy protections, techniques such as differential privacy [71, 81, 86] and secure multi-party computation [87, 88] will be essential. Together, these approaches can safeguard sensitive information while enabling collaborative learning.

A practical application could involve a federated system for public health surveillance. Imagine a model for detecting sub-events of a disease outbreak (e.g., "suspected cases," "testing site overload," "vaccination drives"). Hospitals in different countries could collaboratively train this model using their local, privacy-sensitive data on positive cases and Twitter chatter. The model learns to recognize global patterns in sub-event evolution without any hospital patient

data ever leaving its secure server, thus complying with regulations such as HIPAA (the U.S. Health Insurance Portability and Accountability Act) or GDPR (the European Union's General Data Protection Regulation).

Nonetheless, significant challenges remain. The heterogeneity of sub-event data across sources poses difficulties for FL algorithms [81, 89], since the definition and characteristics of a "rescue operation" may differ significantly between datasets. In addition, the communication overhead required for distributed training [90] and the vulnerability of federated systems to adversarial attacks [89, 91] must be addressed to ensure both efficiency and security in real-world applications.

6.3. Cross-Platform and Multimodal Integration for Richer Context

Future sub-event detection systems must move beyond single-platform analysis to capture a complete and more accurate picture of unfolding events. Cross-platform integration—combining data from sources such as Twitter, Facebook, Reddit, and TikTok [3]—is critical, as different sub-events may surface more prominently on other platforms (e.g., emergency requests on Twitter vs. community organizing on Facebook) [92].

A key technical challenge is data standardization: unifying disparate formats and APIs into consistent representations of sub-events. Hybrid systems that fuse social media signals with traditional news feeds, government bulletins, and official data sources [93] will also be essential for ground-truthing, helping to validate sub-events and filter noise.

The ultimate vision is multimodal systems that integrate text, images, video, audio, and sensor data to characterize sub-events in detail [21]. For instance, confirming a "flooding" sub-event could involve synchronizing a tweeted photo of rising water with audio of sirens and a textual plea for help. This requires advanced cross-modal reasoning to infer sub-event characteristics from partial evidence, producing a richer and more verifiable event model.

6.4. Emerging Technologies for Novel Paradigms

Longer-term research should investigate how emerging computing paradigms might fundamentally reshape sub-event detection. Quantum computing, for example, could offer new approaches to solving the complex optimization problems involved in tracking millions of potential sub-event signals across high-dimensional data streams [94].

Neuromorphic computing and spiking neural networks [95] They are especially promising due to their event-driven and energy-efficient architectures. These characteristics align naturally with the sparse, temporal nature of sub-event data, potentially enabling real-time analysis at scales not achievable with conventional hardware.

As digital ecosystems evolve, systems must also prepare for immersive media from AR/VR environments [96]. Such platforms will generate rich spatial and contextual information about sub-events, requiring new techniques for spatial event detection within virtual and augmented worlds.

Finally, blockchain technology [11, 71] offers the possibility of immutable, tamper-resistant ledgers for recording the provenance and verification status of sub-event reports. This capability could provide trusted audit trails from initial detection through resolution, strengthening defenses against misinformation at the sub-event level.

6.5. Standardization and Benchmarking for Sub-Event Detection

The advancement of sub-event detection depends on the development of standardized evaluation frameworks tailored to its unique challenges [97]. Existing benchmarks for general event detection are inadequate for measuring key aspects such as temporal segmentation accuracy, causal linking precision, or the recall of subtle sub-events. Therefore, Future work must establish dedicated datasets with annotated sub-event sequences and reproducible evaluation protocols that capture detection accuracy and the progression and evolution of events over time.

Evaluation should also be multi-dimensional, assessing whether models can fairly detect sub-events across diverse demographics, geographies, and event types. As the field matures, professional standards and certification processes will become essential for deploying sub-event detection systems in high-stakes settings. These efforts will help ensure that such systems are accurate but also ethical, robust, and trustworthy in practice.

6.6. Conclusion

Content-based event detection on Twitter has seen remarkable advancements, transitioning from basic keyword monitoring to complex AI-driven systems capable of real-time analysis of rich, multimodal data streams. This survey has outlined a comprehensive taxonomy and discussed key challenges in event detection, particularly in understanding sub-events and their evolution.

Looking ahead, future research should be guided by several actionable recommendations to bridge current gaps and build more robust systems. First, we recommend developing specialized benchmarks that challenge models to perform nuanced temporal reasoning, explicitly requiring them to identify sub-event sequences and causal relationships within event narratives. Second, the field should prioritize creating standardized protocols and open-source frameworks for federated learning in a multi-modal context, enabling privacy-preserving collaboration across institutions without centralizing sensitive user data. Third, researchers must adopt and extend rigorous auditing toolkits to proactively measure and mitigate biases—such as geographic, demographic, and linguistic biases—in training data and model outputs for event detection systems. Finally, we urge a stronger focus on interdisciplinary collaboration with domain experts in disaster response and public health to co-design evaluation metrics that truly reflect real-world operational needs, moving beyond purely academic performance scores.

By addressing these interconnected research avenues, the field can move toward creating comprehensive, trustworthy, and practical event detection frameworks that translate cutting-edge technological innovation into meaningful, positive real-world impact.

7. References

[1] Atefeh, F., & Khreich, W. (2015). A survey of techniques for event detection in Twitter. Computational Intelligence, 31(1), 133–164. doi:10.1111/coin.12017.

[2] Li, Q., Chao, Y., Li, D., Lu, Y., & Zhang, C. (2022). Event Detection from Social Media Stream: Methods, Datasets and Opportunities. Proceedings - 2022 IEEE International Conference on Big Data, Big Data 2022, 3509–3516. doi:10.1109/BigData55660.2022.10020411.

[3] Mredula, M. S., Dey, N., Rahman, M. S., Mahmud, I., & Cho, Y. Z. (2022). A Review on the Trends in Event Detection by Analyzing Social Media Platforms' Data. Sensors, 22(12), 4531. doi:10.3390/s22124531.

[4] Parekh, T., Mac, A., Yu, J., Dong, Y., Shahriar, S., Liu, B., Yang, E., Huang, K. H., Wang, W., Peng, N., & Chang, K. W. (2024). Event Detection from Social Media for Epidemic Prediction. Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2024, 1, 5758–5783. doi:10.18653/v1/2024.naacl-long.322.

[5] Belcastro, L., Marozzo, F., Talia, D., Trunfio, P., Branda, F., Palpanas, T., & Imran, M. (2021). Using social media for sub-event detection during disasters. Journal of Big Data, 8(1). doi:10.1186/s40537-021-00467-1.

[6] Rajaby Faghihi, H., Alhafni, B., Zhang, K., Ran, S., Tetreault, J., & Jaimes, A. (2022). CrisisLTLSum: A Benchmark for Local Crisis Event Timeline Extraction and Summarization. Findings of the Association for Computational Linguistics: EMNLP 2022, 5455–5477. doi:10.18653/v1/2022.findings-emnlp.400.

[7] Senthilkumar, K. K., Bhatt, C., Renuka Jyothi, S., Patil, S. B., Ravivarman, G., & Mahajan, S. (2024). Twitter Sarcasm Detection using Natural Language Processing and Deep Learning Techniques. 2024 Global Conference on Communications and Information Technologies, GCCIT 2024. doi:10.1109/GCCIT63234.2024.10862514.

[8] Narasamma, V. L., & Sreedevi, M. (2021). Twitter based Data Analysis in Natural Language Processing using a Novel Catboost Recurrent Neural Framework. International Journal of Advanced Computer Science and Applications, 12(5), 440–447. doi:10.14569/IJACSA.2021.0120555.

[9] Kolajo, T., Daramola, O., & Adebiyi, A. A. (2022). Real-time event detection in social media streams through semantic analysis of noisy terms. Journal of Big Data, 9(1). doi:10.1186/s40537-022-00642-y.

[10] Fedoryszak, M., Frederick, B., Rajaram, V., & Zhong, C. (2019). Real-time Event Detection on Social Data Streams. Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2774–2782. doi:10.1145/3292500.3330689.

[11] Thilak Raj, M., Nivetha, R., Nithish Kumar, S., & Varshini, C. (2024). Cyber Sleuth: Harnessing NLP and Blockchain for Twitter Based Fake News Detection. 2nd International Conference on Artificial Intelligence and Machine Learning Applications: Healthcare and Internet of Things, AIMLA 2024. doi:10.1109/AIMLA59606.2024.10531313.

[12] Dahlan, F., & Suyanto, S. (2023). Data Augmentations to Improve BERT-based Detection of Covid-19 Fake News on Twitter. 2023 International Conference on Computer Science, Information Technology and Engineering (ICCoSITE), 140–145. doi:10.1109/iccosite57641.2023.10127796.

[13] Cui, J. (2025). Enhancing Security Event Detection on Twitter with Graph-based Tweet Embedding. Network and Distributed System Security (NDSS) Symposium Symposium. doi:10.14722/aiscc.2024.23002.

[14] -, A. S., -, D. K., & -, A. G. (2024). AI-Enhanced Cyberbullying Detection in Encrypted Social Media: A Privacy-Preserving Federated Learning Approach. International Journal on Science and Technology, 15(2). doi:10.71097/ijsat.v15.i2.4011.

[15] Cai, Z., Kung, P.-N., Suvarna, A., Ma, M., Bansal, H., Chang, B., . . . Peng, N. (2024, August). Improving Event Definition Following For Zero-Shot Event Detection. In L.-W. Ku, A. Martins, & V. Srikumar (Ed.), Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 2842–2863). Bangkok: Association for Computational Linguistics. doi:10.18653/v1/2024.acl-long.157

[16] Pouran Ben Veyseh, A., Lai, V., Dernoncourt, F., & Nguyen, T. H. (2021). Unleash GPT-2 Power for Event Detection. Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, 1, 6271–6282. doi:10.18653/v1/2021.acl-long.490.

[17] Balci, E., & Sarac, E. (2024). Automated Depression Detection from Tweets: A Comparison of NLP Techniques. 8th International Artificial Intelligence and Data Processing Symposium, IDAP 2024. doi:10.1109/IDAP64064.2024.10711029.

[18] ALAMSYAH, N., Saparudin, & Prima Kurniati, A. (2024). Event Detection Optimization Through Stacking Ensemble and BERT Fine-Tuning for Dynamic Pricing of Airline Tickets. IEEE Access, 12, 145254–145269. doi:10.1109/ACCESS.2024.3466270.

[19] Akintoye, O., Wei, N., & Liu, Q. (2024). Suicide Detection in Tweets Using LSTM and Transformers. Proceedings - 2024 4th Asia Conference on Information Engineering, ACIE 2024, 22–27. doi:10.1109/ACIE61839.2024.00011.

[20] Qiu, Z., Wu, J., Yang, J., Su, X., & Aggarwal, C. (2025). Heterogeneous Social Event Detection via Hyperbolic Graph Representations. IEEE Transactions on Big Data, 11(1), 115–129. doi:10.1109/TBDATA.2024.3381017.

[21] El-Niss, A., Alzu'Bi, A., & Abuarqoub, A. (2023). Multimodal Fusion for Disaster Event Classification on Social Media: A Deep Federated Learning Approach.

Proceedings of the 7th International Conference on Future Networks and Distributed Systems, 758–763. doi:10.1145/3644713.3644840.

[22] Vasconcelos, A. B., Drummond, L. M. d. A., Brum, R. C., & Paes, A. (2023). Exploring Federated Learning to Trace Depression in Social Media with Language Models. 2023 International Symposium on Computer Architecture and High Performance Computing Workshops (SBAC-PADW), 24–30. doi:10.1109/sbac-padw60351.2023.00014.

[23] Healy, P., Hunt, G., Kilroy, S., Lynn, T., Morrison, J. P., & Venkatagiri, S. (2015). Evaluation of peak detection algorithms for social media event detection. 2015 10th International Workshop on Semantic and Social Media Adaptation and Personalization (SMAP), 1–9. doi:10.1109/smap.2015.7370090.

[24] Malik, M., Aslam, W., Aslam, Z., Alharbi, A., Alouffi, B., & Rauf, H. T. (2022). A Performance Comparison of Unsupervised Techniques for Event Detection from Oscar Tweets. Computational Intelligence and Neuroscience, 2022. doi:10.1155/2022/5980043.

[25] Zhang, B., Yang, Y., Niu, F., Fu, X., Dai, G., & Huang, H. (2025, November). SPARK: Simulating the Co-evolution of Stance and Topic Dynamics in Online Discourse with LLM-based Agents. In C. Christodoulopoulos, T. Chakraborty, C. Rose, & V. Peng (Ed.), Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (pp. 23061–23073). Suzhou: Association for Computational Linguistics. doi:10.18653/v1/2025.emnlp-main.1176

[26] Sasaki, K., Yoshikawa, T., & Furuhashi, T. (2014). Online topic model for Twitter considering dynamics of user interests and topic trends. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 1977–1985. doi:10.3115/v1/d14-1212.

[27] Lim, K. W., & Buntine, W. (2014). Twitter Opinion Topic Model. Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management, 1319–1328. doi:10.1145/2661829.2662005.

[28] Lossio-Ventura, J. A., Gonzales, S., Morzan, J., Alatrista-Salas, H., Hernandez-Boussard, T., & Bian, J. (2021). Evaluation of clustering and topic modeling methods over health-related tweets and emails. Artificial Intelligence in Medicine, 117, 102096. doi:10.1016/j.artmed.2021.102096.

[29] Lossio-Ventura, J. A., Morzan, J., Alatrista-Salas, H., Hernandez-Boussard, T., & Bian, J. (2019). Clustering and topic modeling over tweets: A comparison over a health dataset. 2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 1544–1547. doi:10.1109/bibm47256.2019.8983167.

[30] Ganeshkumar, P., BR, A. K., Padmanabhan, S., & others. (2022). Social Media Personal Event Notifier Using NLP and Deep Learning. 2022 International Conference on Power, Energy, Control and Transmission Systems (ICPECTS), (pp. 1–5). DOI: 10.1109/ICPECTS56089.2022.10047710

[31] Desai, M., Mehta, R. G., & Rana, D. P. (2024). Anatomising the impact of ResearchGate followers and followings on influence identification. Journal of Information Science, 50(3), 607–624. doi:10.1177/01655515221100716.

[32] Singh, R.R. (2022). Centrality Measures: A Tool to Identify Key Actors in Social Networks. Principles of Social Networking. Smart Innovation, Systems and Technologies, vol 246. Springer, Singapore. doi:10.1007/978-981-16-3398-0_1.

[33] Romero, D. M., Meeder, B., & Kleinberg, J. (2011). Differences in the mechanics of information diffusion across topics: Idioms, political hashtags, and complex contagion on twitter. Proceedings of the 20th International Conference on World Wide Web, WWW 2011, 695–704. doi:10.1145/1963405.1963503.

[34] Lin, Y., Yang, D., Hou, J., Yan, C., Kim, M., Laurienti, P. J., & Wu, G. (2021). Learning dynamic graph embeddings for accurate detection of cognitive state changes in functional brain networks. NeuroImage, 230, 117791. doi:10.1016/j.neuroimage.2021.117791.

[35] Qiu, Z., Ma, C., Wu, J., & Yang, J. (2024). An Efficient Automatic Meta-Path Selection for Social Event Detection via Hyperbolic Space. WWW 2024 - Proceedings of the ACM Web Conference, 2519–2529. doi:10.1145/3589334.3645526.

[36] Huang, W., Zong, Y., Shi, Z., & Liu, P. (2023). MESCAL: Malicious Login Detection Based on Heterogeneous Graph Embedding with Supervised Contrastive Learning. 2023 IEEE Symposium on Computers and Communications (ISCC), 1274–1279. doi:10.1109/iscc58397.2023.10218074.

[37] Cekinel, R. F., & Karagoz, P. (2022). Event prediction from news text using subgraph embedding and graph sequence mining. World Wide Web, 25(6), 2403–2428. doi:10.1007/s11280-021-01002-1.

[38] Kleinberg, J. (2003). Bursty and Hierarchical Structure in Streams. Data Mining and Knowledge Discovery, 7(4), 373–397. doi:10.1023/a:1024940629314.

[39] Lütkepohl, H. (2005). New introduction to multiple time series analysis. Springer, Berlin, Germany. doi:10.1007/978-3-540-27752-1.

[40] Jolliffe, I. (2025). Principal Component Analysis. In: Lovric, M. (eds) International Encyclopedia of Statistical Science. Springer, Berlin, Germany. doi:10.1007/978-3-662-69359-9_483

[41] Seetha, A., Chouhan, S. S., Pilli, E. S., & Raychoudhury, V. (2024). DiEvD: Disruptive Event Detection from Dynamic Datastreams Using Continual Machine Learning: A Case Study with Twitter. IEEE Transactions on Emerging Topics in Computing, 12(3), 727–738. doi:10.1109/TETC.2023.3272973.

[42] Rashmi, C., Soumya, B., Harika, A., & Harathi, D. (2024). Live Event Detection for People's Safety Using Nlp and Deep Learning. Turkish Journal of Computer and

Mathematics Education (Turcomat), 15(3), 434–441. doi:10.61841/turcomat.v15i3.14956.

[43] Purnomo, A., Naufal, A. A., Yudha, E. P., & Arifin, A. Z. (2020). Tweet Classification Using Deep Learning Architecture for Concert Event Detection. Jurnal Ilmu Komputer Dan Informasi, 13(2), 57–63. doi:10.21609/jiki.v13i2.815.

[44] Fang, Y., Gao, J., Liu, Z., & Huang, C. (2020). Detecting cyber threat event from twitter using IDCNN and BiLSTM. Applied Sciences (Switzerland), 10(17), 5922. doi:10.3390/app10175922.

[45] Sen, A., Rajakumaran, G., Mahdal, M., Usharani, S., Rajasekharan, V., Vincent, R., & Sugavanan, K. (2024). Live Event Detection for People's Safety Using NLP and Deep Learning. IEEE Access, 12, 6455-6472. doi:10.1109/ACCESS.2023.3349097

[46] Ghaswala, M., Gomes, C., Kumar, S. M., & Sweta, S. (2024). Depression Detection from social media Text Using NLP and Deep Learning Techniques. 2024 First International Conference for Women in Computing (InCoWoCo), 1–7. doi:10.1109/incowoco64194.2024.10863118.

[47] Adesokan, A., Madria, S., & Nguyen, L. (2023). HatEmoTweet: low-level emotion classifications and spatiotemporal trends of hate and offensive COVID-19 tweets. Social Network Analysis and Mining, 13(1). doi:10.1007/s13278-023-01132-6.

[48] Maveli, N. (2020). EdinburghNLP at WNUT-2020 Task 2: Leveraging Transformers with Generalized Augmentation for Identifying Informativeness in COVID-19 Tweets. Proceedings of the Sixth Workshop on Noisy User-Generated Text (W-NUT 2020, 455–461. doi:10.18653/v1/2020.wnut-1.67.

[49] Tran, K., Phan, H., Nguyen, K., & Thuy Nguyen, N. L. (2020). UIT-HSE at WNUT-2020 Task 2: Exploiting CT-BERT for Identifying COVID-19 Information on the Twitter Social Network. Proceedings of the Sixth Workshop on Noisy User-Generated Text (W-NUT 2020, 383–387. doi:10.18653/v1/2020.wnut-1.53.

[50] Dey, A., Bothera, A., Sarikonda, S., Aryan, R., Podishetty, S. K., Havalgi, A., Srivastava, S., & Singh, G. (2026). Multimodal Event Detection: Current Approaches and Defining the New Playground Through LLMs and VLMs. Natural Language Processing and Information Systems. NLDB 2025. Lecture Notes in Computer Science, vol 15836. Springer, Cham, Switzerland. doi:10.1007/978-3-031-97141-9_31.

[51] Kashif, M., Zohair, M., & Ali, S. (2023). Lexical squad@ multimodal hate speech event detection 2023: Multimodal hate speech detection using fused ensemble approach. Proceedings of the 6th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text, 7 September, 2023, Varna, Bulgaria.

[52] Jin, Y., Choi, M., Verma, G., Wang, J., & Kumar, S. (2024). MM-SOC: Benchmarking Multimodal Large Language Models in Social Media Platforms. Findings of the Association for Computational Linguistics ACL 2024, 6192–6210. doi:10.18653/v1/2024.findings-acl.370.

[53] Xu, J., Zhao, H., Liu, W., & Ding, X. (2023). Research on False Information Detection Based on Multimodal Event Memory Network. 2023 3rd International Conference on Consumer Electronics and Computer Engineering (ICCECE), 566–570. doi:10.1109/iccece58074.2023.10135191.

[54] Esackimuthu, S., & Balasundaram, P. (202). Verbavisor@ multimodal hate speech event detection 2023: Hate speech detection using transformer model. Proceedings of the 6th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text, 7 September, 2023, Varna, Bulgaria.

[55] Alam, F., Sajjad, H., Imran, M., & Ofli, F. (2021). CrisisBench: Benchmarking Crisis-related Social Media Datasets for Humanitarian Information Processing. Proceedings of the International AAAI Conference on Web and Social Media, 15, 923–932. doi:10.1609/icwsm.v15i1.18115.

[56] Varachkina, H., Ziehe, S., Dönicke, T., & Pannach, F. (2020). #GCDH at WNUT-2020 Task 2: BERT-Based Models for the Detection of Informativeness in English COVID-19 Related Tweets. Proceedings of the Sixth Workshop on Noisy User-Generated Text (W-NUT 2020, 462–465. doi:10.18653/v1/2020.wnut-1.68.

[57] Neruda, G. A., & Winarko, E. (2021). Traffic Event Detection from Twitter Using a Combination of CNN and BERT. 2021 International Conference on Advanced Computer Science and Information Systems, ICACSIS 2021. doi:10.1109/ICACSIS53237.2021.9631334.

[58] Sech, J., DeLucia, A., Buczak, A. L., & Dredze, M. (2020, November). Civil Unrest on Twitter (CUT): A Dataset of Tweets to Support Research on Civil Unrest. In W. Xu, A. Ritter, T. Baldwin, & A. Rahimi (Ed.), Proceedings of the Sixth Workshop on Noisy User-generated Text (W-NUT 2020) (pp. 215–221). Online: Association for Computational Linguistics. doi:10.18653/v1/2020.wnut-1.28

[59] Suppa, M., Skala, D., Jass, D., Sucik, S., Svec, A., & Hraska, P. (2024). Bryndza at ClimateActivism 2024: Stance, Target and Hate Event Detection via Retrieval-Augmented GPT-4 and LLaMA. Proceedings of the 7th Workshop on Challenges and Applications of Automated Extraction of Socio-Political Events from Text (CASE 2024), 166–177. doi:10.18653/v1/2024.case-1.23.

[60] Owusu-Adjei, M., Ben Hayfron-Acquah, J., Frimpong, T., & Abdul-Salaam, G. (2023). Imbalanced class distribution and performance evaluation metrics: A systematic review of prediction accuracy for determining model performance in healthcare systems. PLOS Digital Health, 2(11), e0000290. doi:10.1371/journal.pdig.0000290.

[61] Munshi, M., Gupta, R., Jadav, N. K., Polkowski, Z., Tanwar, S., Alqahtani, F., & Said, W. (2024). Quantum machine learning-based framework to detect heart failures in

Healthcare 4.0. Software - Practice and Experience, 54(2), 168–185. doi:10.1002/spe.3264.

[62] Yacouby, R., & Axman, D. (2020). Probabilistic Extension of Precision, Recall, and F1 Score for More Thorough Evaluation of Classification Models. Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems, 79–91. doi:10.18653/v1/2020.eval4nlp-1.9.

[63] Fox, A., Swarup, S., & Adiga, A. (2025, April). A Unifying Information-theoretic Perspective on Evaluating Generative Models. Proceedings of the AAAI Conference on Artificial Intelligence, 39, 16630–16638. doi:10.1609/aaai.v39i16.33827

[64] Jiao, X., Wan, S., Liu, Q., Bi, Y., Lee, Y. L., Xu, E., Hao, D., & Zhou, T. (2024). Comparing discriminating abilities of evaluation metrics in link prediction. Journal of Physics: Complexity, 5(2), 25014. doi:10.1088/2632-072X/ad46be.

[65] Richardson, E., Trevizani, R., Greenbaum, J. A., Carter, H., Nielsen, M., & Peters, B. (2024). The receiver operating characteristic curve accurately assesses imbalanced datasets. Patterns, 5, 100994. doi:https://doi.org/10.1016/j.patter.2024.100994

[66] Perrella, S., Proietti, L., Huguet Cabot, P.-L., Barba, E., & Navigli, R. (2024). Beyond Correlation: Interpretable Evaluation of Machine Translation Metrics. Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, 20689–20714. doi:10.18653/v1/2024.emnlp-main.1152.

[67] Farruque, N., Goebel, R., Sivapalan, S., & Zaïane, O. R. (2024). Depression symptoms modelling from social media text: an LLM driven semi-supervised learning approach. Language Resources and Evaluation, 58(3), 1013–1041. doi:10.1007/s10579-024-09720-4.

[68] Ilham, F., & Maharani, W. (2022). Analyze Detection Depression In Social Media Twitter Using Bidirectional Encoder Representations from Transformers. Journal of Information System Research (JOSH), 3(4), 476–482. doi:10.47065/josh.v3i4.1885.

[69] Kostakos, P., Nykanen, M., Martinviita, M., Pandya, A., & Oussalah, M. (2018). Meta-Terrorism: Identifying Linguistic Patterns in Public Discourse After an Attack. 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), 1079–1083. doi:10.1109/asonam.2018.8508647.

[70] Williams, A. R., Burke-Moore, L., Chan, R. S.-Y., Enock, F. E., Nanni, F., Sippy, T., . . . Bright, J. (2025, March 17). Large language models can consistently generate high-quality content for election disinformation operations. PLOS ONE, 20, e0317421. doi:10.1371/journal.pone.0317421

[71] Salim, S., Turnbull, B., & Moustafa, N. (2024). A Blockchain-Enabled Explainable Federated Learning for Securing Internet-of-Things-Based Social Media 3.0 Networks. IEEE Transactions on Computational Social Systems, 11(4), 4681–4697. doi:10.1109/TCSS.2021.3134463.

[72] Sen, J., Waghela, H., & Rakshit, S. (2025). Privacy in Federated Learning. Data Privacy - Techniques, Applications, and Standards. IntechOpen Limited, London, United Kingdom. doi:10.5772/intechopen.1006677.

[73] Qiu, H., Dou, Z.-Y., Wang, T., Celikyilmaz, A., & Peng, N. (2023). Gender Biases in Automatic Evaluation Metrics for Image Captioning. Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, 8358–8375. doi:10.18653/v1/2023.emnlp-main.520.

[74] Chowdhury, S. R., Basu, S., & Maulik, U. (2022). A survey on event and subevent detection from microblog data towards crisis management. International Journal of Data Science and Analytics, 14(4), 319–349. doi:10.1007/s41060-022-00335-y.

[75] Nolasco, D., & Oliveira, J. (2019). Subevents detection through topic modeling in social media posts. Future Generation Computer Systems, 93, 290–303. doi:10.1016/j.future.2018.09.008.

[76] Lu, G., Mu, Y., Gu, J., Kouassi, F. A. P., Lu, C., Wang, R., & Chen, A. (2021). A hashtag-based sub-event detection framework for social media. Computers and Electrical Engineering, 94, 107317. doi:10.1016/j.compeleceng.2021.107317.

[77] Xiao, K., Qian, Z., & Qin, B. (2022). A Survey of Data Representation for Multi-Modality Event Detection and Evolution. Applied Sciences (Switzerland), 12(4), 2204. doi:10.3390/app12042204.

[78] Koupaee, M., Bai, X., Chen, M., Durrett, G., Chambers, N., & Balasubramanian, N. (2025, July). Causal Graph based Event Reasoning using Semantic Relation Experts. In W. Che, J. Nabende, E. Shutova, & M. T. Pilehvar (Ed.), Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 26169–26199). Vienna: Association for Computational Linguistics. doi:10.18653/v1/2025.acl-long.1269

[79] Chen, R., Qin, C., Jiang, W., & Choi, D. (2024, March). Is a Large Language Model a Good Annotator for Event Extraction? Proceedings of the AAAI Conference on Artificial Intelligence, 38, 17772–17780. doi:10.1609/aaai.v38i16.29730

[80] Dusart, A., Pinel-Sauvagnat, K., & Hubert, G. (2023). TSSuBERT: How to Sum Up Multiple Years of Reading in a Few Tweets. ACM Transactions on Information Systems, 41(4). doi:10.1145/3581786.

[81] Wang, Y., Su, Z., Pan, Y., Luan, T. H., Li, R., & Yu, S. (2024). Social-Aware Clustered Federated Learning With Customized Privacy Preservation. IEEE/ACM Transactions on Networking, 32(5), 3654–3668. doi:10.1109/TNET.2024.3379439.

[82] Pare, T., Haque, M. J., & Bhaladhare, P. R. (2024). Federated Learning Approach for Social Media Sentiment Analysis: Analyzing Public Opinion. 2024 8th International Conference on Computing, Communication, Control and Automation, ICCUBEA 2024. doi:10.1109/ICCUBEA61740.2024.10774886.

[83] Tarun Pare. (2024). Crowdsourced Intelligence: A Federated Learning Approach to Analyzing Public Opin-ion on Social Media. Journal of Electrical Systems, 20(10s), 7909–7920. doi:10.52783/jes.7006.

[84] Mistry, D., Plabon, J. D., Diba, B. S., Mukta, M. S. H., & Mridha, M. F. (2024). Federated Learning-Based Architecture for Personalized Next Emoji Prediction for Social Media Comments. IEEE Access, 12, 140339–140358. doi:10.1109/ACCESS.2024.3448470.

[85] Zhang, S. (2024). Using Federated Learning Technology to Build a Social Media User Model for China's Multicultural Integration. 2024 International Conference on Language Technology and Digital Humanities (LTDH), 81–86. doi:10.1109/ltdh64262.2024.00025.

[86] Huang, W., Tiropanis, T., & Konstantinidis, G. (2023). A Dual-Layer Privacy-Preserving Federated Learning Framework. Web Information Systems Engineering – WISE 2023: 24th International Conference, Melbourne, VIC, Australia, October 25–27, 2023, Proceedings (pp. 245–259). Berlin: Springer-Verlag. doi:10.1007/978-981-99-7254-8_19

[87] Yang, X., Liu, Z., Tang, X., Lu, R., & Liu, B. (2024). An Efficient and Multi-Private Key Secure Aggregation Scheme for Federated Learning. IEEE Transactions on Services Computing, 17(5), 1998–2011. doi:10.1109/TSC.2024.3451165.

[88] Li, Y., Xu, G., Meng, X., Du, W., & Ren, X. (2024). LF3PFL: A Practical Privacy-Preserving Federated Learning Algorithm Based on Local Federalization Scheme. Entropy, 26(5), 353. doi:10.3390/e26050353.

[89] Hayashitani, M., Mori, J., & Teranishi, I. (2025). Survey of Privacy Threats and Countermeasures in Federated Learning. 2025 3rd International Conference on Federated Learning Technologies and Applications (FLTA), (pp. 78-85). doi:10.1109/FLTA67013.2025.11336767

[90] Abd El-Kareem Abd El-Moaty Saleh, H., Fernández Vilas, A., Fernández-Veiga, M., El-Sonbaty, Y., & El-Bendary, N. (2022). Using Decentralized Aggregation for Federated Learning with Differential Privacy. Proceedings of the 19th ACM International Symposium on Performance Evaluation of Wireless Ad Hoc, Sensor, &Amp; Ubiquitous Networks on 19th ACM International Symposium on Performance Evaluation of Wireless Ad Hoc, Sensor, & Ubiquitous Networks, 33–39. doi:10.1145/3551663.3558682.

[91] Wang, L., Zhu, T., Zhou, W., & Yu, P. S. (2025). Linkage on security, privacy and fairness in federated learning: New balances and new perspectives. Neural Networks, 192, 107874. doi:10.1016/j.neunet.2025.107874.

[92] Hodorog, A., Petri, I., & Rezgui, Y. (2022). Machine learning and Natural Language Processing of social media data for event detection in smart cities. Sustainable Cities and Society, 85, 104026. doi:10.1016/j.scs.2022.104026.

[93] Yan, F., Zhang, M., Wei, B., Ren, K., & Jiang, W. (2024). SARD: Fake news detection based on CLIP contrastive learning and multimodal semantic alignment. Journal of King Saud University - Computer and Information Sciences, 36, 102160. doi:https://doi.org/10.1016/j.jksuci.2024.102160

[94] Derakhshan, M., & Mohammadi, F. (2025). Leveraging GPT-4o Efficiency for Detecting Rework Anomaly in Business Processes. 2025 3rd International Conference on Foundation and Large Language Models (FLLM), (pp. 939-945). doi:10.1109/FLLM67465.2025.11391132

[95] Yu, Z., Qu, Q., Chen, X., & Wang, C. (2025). Can Large Language Models Grasp Event Signals? Exploring Pure Zero-Shot Event-based Recognition. ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 1–5. doi:10.1109/icassp49660.2025.10887714.

[96] Liu, S., Li, J., Zhao, G., Zhang, Y., Meng, X., Yu, F. R., Ji, X., & Li, M. (2025). EventGPT: Event Stream Understanding with Multimodal Large Language Models. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR, 29139–29149. doi:10.1109/cvpr52734.2025.02713.

[97] Liu, Y. L., Blodgett, S. L., Cheung, J. C. K., Liao, Q. V., Olteanu, A., & Xiao, Z. (2024). ECBD: Evidence-Centered Benchmark Design for NLP. Proceedings of the Annual Meeting of the Association for Computational Linguistics, 1, 16349–16365. doi:10.18653/v1/2024.acl-long.861.