

From What-If Scenarios to Event Associations: A Novel Approach to Social Media Event Analysis

Aigerim Mussina¹^a, Sanzhar Aubakirov¹^b, Paulo Trigo²^c and Madina Mansurova¹^d

¹*Al-Farabi Kazakh National University, Almaty, Kazakhstan*

²*ISEL - Instituto Superior de Engenharia de Lisboa, Lisbon, Portugal*

Keywords: Events Association, Counterfactual Analysis, Association Rules, What-If Analysis.


Abstract: This paper introduces a novel approach to event prediction in social media by applying association rules to generate counterfactual what-if scenarios. Using the Events2012 dataset as a foundation, we developed the EventsAssociation2012 dataset to systematically identify patterns within event sequences and assess the predictive power of what-if scenarios. Employing a Large Language Model (LLM) to generate event embeddings, similarity scores, and conditional probabilities, we mapped real-world scenarios to intra-event and inter-event associations, thereby creating a robust framework for understanding the interconnected nature of social media discussions. Our methodology leverages association rule mining to model causal relationships between events, enabling predictions of plausible future outcomes based on hypothetical scenarios. The results demonstrate the potential for applying what-if scenarios to new event datasets, revealing challenges and opportunities for refining this approach. The study further discusses areas for improvement, such as expanding the identification of intra-event scenarios, exploring multi-event associations, and enhancing topic embedding techniques. Overall, this work advances counterfactual analysis in event prediction, providing a more accurate and comprehensive method for modeling event associations in the dynamic landscape of social media.


1 INTRODUCTION


In recent years, Online Social Networks (OSNs) have become central to information exchange, allowing millions of users worldwide to share their experiences, opinions, and perceptions in real-time. Platforms such as Twitter, Facebook, and Telegram offer a vast repository of user-generated content, making them invaluable sources for capturing collective social dynamics. Studying user messages on OSNs has provided insights into societal trends, public opinions, and information propagation. One emerging research area is the exploration of associations between significant events discussed on these platforms. Identifying and analyzing the most discussed events can uncover underlying patterns and causal relationships, contributing to a deeper understanding of social phenomena (Tangcharoensathien et al., 2020; Valdez et al., 2020).


Events within OSNs, as derived from public discussions, represent the main points of collective attention. These events could range from political events, natural disasters, and cultural movements to technological innovations (Kaliyar et al., 2021; Daud et al., 2020). They are often interconnected, forming chains of discussions that could reflect sequences of causal effects in real-world scenarios. For example, a political debate may trigger widespread online discourse, which in turn can influence public opinion, shape policy decisions, and lead to further events in a domino effect (Islam et al., 2020). Understanding these associations provides a way to decode public sentiment and an opportunity to forecast future developments based on current discussions.

Event association detection within OSNs is a relatively new research area, with a limited number of studies exploring the nuances of how events discussed on social media platforms are interconnected. Most existing research focuses on detecting events based on the intensity and frequency of discussions, such as using keyword-based extraction, topic modeling, or sentiment analysis (Ali et al., 2021). Few studies have explored the more profound dimension of con-

^a <https://orcid.org/0000-0002-7043-0810>

^b <https://orcid.org/0000-0002-8416-527X>

^c <https://orcid.org/0000-0001-5850-615X>

^d <https://orcid.org/0000-0002-9680-2758>

necting events, in which one event may potentially trigger or affect another. Some studies in this domain often rely on time-series analysis or basic correlation techniques, which might not fully capture the complex web of event associations as they unfold in real-world contexts (Daud et al., 2020; Bian et al., 2020). For example, link prediction (Daud et al., 2020) provides insights into associations but does not always reflect causal relationships. Methods like rumor detection using graph convolutional networks (Bian et al., 2020) explore information propagation but must establish event-to-event causality. Our work addresses this gap by introducing a more sophisticated approach to identifying these associations through counterfactual analysis.

Counterfactual analysis traditionally finds its applications in business and healthcare. For example, in business, counterfactuals help assess the impact of strategic decisions (Eabrasu, 2008), while in healthcare, they evaluate the potential outcomes of various treatment options (Shalit et al., 2017). One work suggested using counterfactual models to assess the impact of Twitter misinformation on future events (Zhang et al., 2022). Their work focuses on capturing the temporal dynamics of information dissemination and its potential influence on public discourse. By employing a neural temporal point process model, they estimated the causal effects of misinformation propagation on social networks, demonstrating the value of counterfactual reasoning in understanding the broader consequences of false information. The counterfactual analysis was used to explore the impact of social media campaigns on user behavior (Yu et al., 2022). Researchers developed a causal impact model to assess how the diffusion of social media content influences user actions, such as participation in social movements or purchasing behavior. Their work highlights how counterfactual reasoning can offer a deeper understanding of the causal mechanisms behind information spread on OSN. Another research applied counterfactual reasoning to detecting rumors on social networks, offering insights into how events could unfold differently with changes in key events (Zhang et al., 2023). Their work introduced diverse counterfactual evidence to model the spread of rumors, facilitating the exploration of alternative scenarios in which different events influence the rumor's propagation. This research underscores the potential of counterfactual analysis in understanding the dynamics of information spread between events on social media.

While these works have contributed significantly to understanding event associations and the effects of social media content, gaps remain in the application

of counterfactual analysis specifically for event association detection. Previous studies have focused on individual aspects like misinformation, user behavior, narrative structure, and rumor detection. However, a comprehensive exploration of how social media-derived events can generate “what-if” scenarios using association rules to forecast future events is still in its infancy. Our research aims to address this gap by leveraging counterfactual analysis to investigate the causal interconnections between events within social networks.

In our previous work, we introduced the concept of counterfactual “what-if” scenarios for understanding event associations (Mussina et al., 2023). However, this earlier work lacked a formal evaluation of these scenarios. The present study seeks to build upon that foundation by thoroughly evaluating the “what-if” scenarios and demonstrating their practical application in detecting event associations on OSNs. Our central hypothesis is: “Social media-derived event detections can generate what-if scenarios using association rules from event topics, which can then be applied to assess their applicability for future events.” This novel application of counterfactual reasoning to event associations opens up new possibilities for understanding the flow of information and influence within digital ecosystems.

2 MATERIALS AND METHODS

In this section, we introduce our study’s formal definitions and methodologies, detailing how we created a dataset for event association detection, generated “what-if” scenarios and evaluated them, and applied scenarios.

2.1 Definitions

At first we need to describe additional data used in event detection.

Definition 1. A topic of interest, *ToI*, is defined as a dictionary of topics, where each topic has a value of thematic coefficient. A topic, *t*, is defined as an N-gram, a sequence of N words, related to the interest of study. This relation of topic is represented as thematic coefficient of topic which is defined by the next formula:

$$M_t = \log \frac{N_t^{target}}{N_t^{common}} > 0 \quad (1)$$

, where N_t^{target} is a frequency of topic *t* in the target corpora, N_t^{common} is a frequency of topic *t* in the common corpora, and M_t is a thematic coefficient. □

In this work we examined events of different categories. ToI generation is based on the idea that words specific to a particular event category might appear in texts of other categories, but they are most frequently found within their own category (Mussina et al., 2022). During ToI generation event tweets of each category was used as a target corpora while others constructed common corpora. If thematic coefficient was greater than 0, this topic was added to ToI dictionary.

Definition 2. An event $E(w, T)$ is a subset of topics from a predefined set of Topics of Interest (ToI), observed within a time window w . Formally,

$$E(w, T) = \{t_1, t_2, \dots, t_n\} \subseteq T \quad (2)$$

, where:

- T is a ToI dictionary related to the category
- w is the time-window during which the event occurs
- t_x represents individual topics from T , and $\{t_1, t_2, \dots, t_n\}$ is the subset of topics discussed within w

□

An additional characteristic of an event is its newsworthiness, a value calculated during the event detection process. This value indicates the significance of the topics within the event, helping to prioritize or identify key events within the online discussions.

Definition 3. Topic space is a subset of an event's topic-set that causes one event to become another. The counterfactual analysis aims to explore cause-and-effect relations by searching for statements such as “if A occurred, then B is also likely to occur”. Considering A and B as events and “Definition 1”, we may rephrase such a statement as “if, within a time-window, w , topics $\{t_{A1}, t_{A2}, \dots, t_{Ar}\}$ occurred, then topics $\{t_{B1}, t_{B2}, \dots, t_{Bs}\}$ are also likely to occur in w ”, where t_{Ex} means that topic x was addressed in event E . Following the idea of what-if perspective, we suppose that A's topic-set, $\{t_{A1}, t_{A2}, \dots, t_{Ar}\}$, includes “topic space” that has been intervened such that discussion goes to B's topic-set, $\{t_{B1}, t_{B2}, \dots, t_{Bs}\}$. □

The following definitions are formulated based on the market basket analysis approach, utilizing its fundamental concepts, namely: basket, item, itemset, left-hand side (LHS), right-hand side (RHS), support, and confidence. The itemset $I = t_1, t_2, \dots, t_k$ is a set of items with the length k , where each t_x is a topic (item). The association rule consists of subitemsets

in the form $LHS \Rightarrow RHS$, where $LHS, RHS \subset I$ and $LHS \cap RHS = \emptyset$. Support and confidence are special metrics of association rules. The support is the joint probability of LHS and RHS that items in LHS and RHS occur together. The confidence is a conditional probability of the form $P(RHS|LHS)$.

Definition 4. A what-if scenario is generated from an itemset I_N as an association rule, $LHS \Rightarrow RHS$, that takes the form: $Base_L \cup WI_M \Rightarrow RHS_R$, where $Base_L \cup WI_M = LHS$ and $Base_L \cap WI_M = \emptyset$, also $LHS \subset I_N$ and $RHS \subset I_N$; each L, M , and R subscript is the size of the respective subitemset. □

This way, the counterfactual scenario perspective can be read as “if WI_M occurs together with the $Base_L$ then RHS_R is also likely to occur.”

Definition 5. An intra-event scenario is taken from a what-if scenario where $support(WI_M) = support(Base_L) = support(RHS_R)$. □

Definition 6. A sub-event is defined within an intra-event scenario, where multiple events are clustered based on shared topics. Within such a cluster, there exists a center event, which is the event with the highest newsworthiness, indicating its relative importance within the cluster. The other events in the cluster, which are associated with but less significant than the center event, are referred to as sub-events. These sub-events represent smaller occurrences that contribute to the overall context of the center event, highlighting the hierarchical nature of event relationships within social media discussions. □

Definition 7. A one-rule-based inter-event scenario is taken from a what-if scenario where:

- $support(WI_M)$ is the minimum from all subitemsets of size M ,
- $support(Base_L)$ is the maximum from all subitemsets of size L ,
- $support(RHS_R)$ is the minimum from all subitemsets of size R .

□

Definition 8. A two-rules-transitivity-based inter-event scenario is taken from two association rules of the form: $Base_L \Rightarrow WI_M$ and $WI_M \Rightarrow RHS_R$, where $M > L$, $M > R$ and $Base_L, WI_M \subset I_{N1}$, $WI_M, RHS_R \subset I_{N2}$, I_{N1} and I_{N2} are different itemsets. This scenario generation is based on the association rule's confidence antimonotone property. □

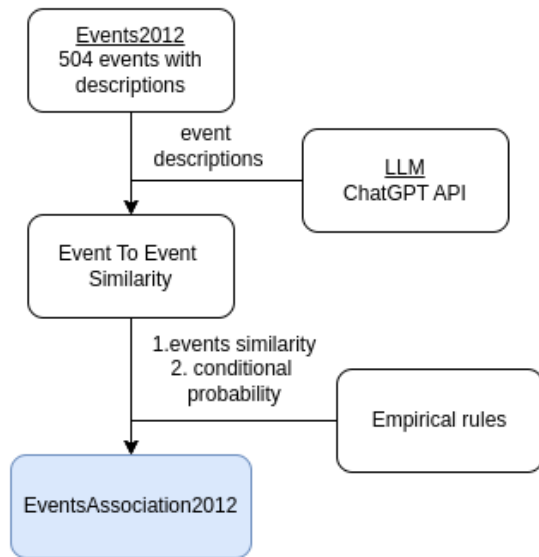


Figure 1: EventsAssociation2012 generation schema.

2.2 EventAssociation Dataset Generation

There is a gap in available datasets with event associations suitable for evaluating what-if scenarios, so we decided to create one, see Figure 1. As the foundation for our new dataset, we used the labeled events from the Events2012 (SEDTWik) dataset (McMinn et al., 2013), which has 504 events. The SEDTWik dataset comprises events in id, description, category. Each description is a one-sentence summary of the topic discussed, and the events are categorized into eight distinct types, including Sport, Politics, Business, and Disaster.

We used a Large Language Model (LLM), to identify associations between these events and detect similarities. LLMs can understand context, semantics, and the subtle nuances in natural language, making them well-suited for comparing event descriptions. We prompted the LLM to compare events within each category and provide the following:

- Similarity – a percentage indicating the events’ similarity based on their textual descriptions.
- Conditional Probability – the likelihood (in percentage) that one event stems from another and vice versa.
- Explanation – a textual description elaborating on the relationship between the events.

The LLM performs these tasks by leveraging its vast training on diverse datasets containing patterns in language, relationships, and causality. When asked to compare events, the LLM analyzes the semantic

content of the event descriptions to calculate similarity and infer potential causal connections. It estimates the conditional probabilities by considering linguistic cues and contextual information that suggest the likelihood of one event leading to another. The LLM generates explanations using its contextual understanding to provide a coherent rationale for the similarity and probability assessments.

After obtaining 24,719 event comparisons through this process, we developed definitions for intra-event and inter-event associations.

Definition 9. Intra-event Association: This describes a single event along with its sub-events. For an association to qualify as intra-event, it must have high similarity, with both conditional probabilities $P(A|B) > 80\%$ and $P(B|A) > 80\%$ and the absolute difference between these probabilities $|P(A|B) - P(B|A)| < 1\%$. The requirement for high conditional probability ensures that even if events are similar within a category, they genuinely describe the same event rather than two unrelated occurrences. When events exhibit high similarity but low conditional probabilities, they do not indicate causality. Therefore, intra-event scenarios represent one event and its sub-events. \square

Definition 10. Inter-event Association: This type of association occurs between two distinct events. It is characterized by high similarity between the events and a significant difference in conditional probability, formulated as $P(A|B) > sim$ or $P(B|A) > sim$, where sim is a similarity between events, and $P(A|B) - P(B|A) > n$, where $n = 20\%$. The variation in conditional probability indicates the direction and possible causal link between two events. \square

Associations were also validated on time-windows. One event could not cause another event in the past. By using an LLM to extract these associations, we can efficiently process and evaluate the complex relationships between events, providing a rich dataset for what-if scenario analysis. Our dataset is called EventAssociation2012.

2.3 Counterfactual What-If Scenarios Evaluation

The purpose of this evaluation is to identify if the generated what-if scenarios correspond to the event associations within the EventsAssociation2012 dataset. To achieve this, we employ the following multi-step process:

- Conversion of Text Data to Vectors
First, we convert the text data from real event descriptions and detected event topics into numeri-

cal vectors using text embeddings. For this purpose, we use the text-embedding-3-small model from the OpenAI client. This model processes each input text to produce a dense vector representation in a high-dimensional space. These vectors capture the semantic essence of the text, allowing for similarity comparisons. The text-embedding-3-small model uses a neural network trained on diverse textual data, creating embeddings that reflect both the syntactic and contextual features of the input text. This embedding process generates vectors that we then use to measure similarity through cosine distance.

- Mapping Detected Events to Real Events

In this step, we map each detected event from Events2012 dataset to its corresponding real event in the same dataset. Let: $R = r_1, r_2, \dots, r_n$ be the set of vectors representing the real event descriptions from Events2012. Let $D = d_1, d_2, \dots, d_m$ be the set of vectors representing the topics of the detected events. We determine the mapping by comparing each detected event vector d_i to all real event vectors r_j , see “Equation 3”. The detected event is considered a match to a real event if it has the maximum similarity, measured using the cosine distance between the vectors. The cosine similarity gives a value between -1 and 1, where 1 indicates that the vectors are identical. This approach helps identify which real event most closely aligns with each detected event based on the discussion topics.

$$event_{d_i} = \max_{j=1}^n \frac{d_i * r_j}{|d_i| * |r_j|}, \quad (3)$$

- Matching What-if Scenarios with Detected Events

A what-if scenario consists of three components: *Base*, *WI* (What-If), and *RHS* (Right-Hand Side). For this evaluation, we combine the *Base* and *WI* into a single component referred to as the *LHS* (Left-Hand Side). The process for matching each side of the what-if scenario to the detected events is as follows: Let $LHS = lhs_1, lhs_2, \dots, lhs_n$ be the set of vectors representing the *LHS* (*Base* + *WI*) of the what-if scenarios. Let $RHS = rhs_1, rhs_2, \dots, rhs_n$ be the set of vectors representing the *RHS* of the what-if scenarios. Let $D = d_1, d_2, \dots, d_m$ be the set of vectors representing the detected event topics. For each what-if scenario, we compare the vectors of the *LHS* and *RHS* with the vectors of the detected events. A scenario is considered to be mapped to a detected event if the *LHS* and *RHS* vectors show the highest cosine similarity with the vectors in D . Let the event with the highest similarity from

LHS from nth what-if scenario with D be named $Event_{LHS}$, see “Equation 4”, and the event with the highest similarity from *RHS* from nth what-if scenario with D be named $Event_{RHS}$, see “Equation 5”. This process enables us to identify which detected events best match the hypothetical conditions outlined in the what-if scenarios.

$$Event_{LHS}(lhs_i) = \max_{j=1}^m \frac{lhs_i * d_j}{|lhs_i| * |d_j|} \quad (4)$$

$$Event_{RHS}(rhs_i) = \max_{j=1}^m \frac{rhs_i * d_j}{|rhs_i| * |d_j|} \quad (5)$$

- Calculating Accuracy of Matched Event Associations

The final step is to evaluate how many of the mapped what-if scenarios correspond to the associations in the EventsAssociation2012 dataset. This involves checking if the matched events for both the *LHS* and *RHS* of each what-if scenario align with an entry in the EventsAssociation2012 dataset. Accuracy is calculated as the ratio of matched scenarios to the total number of evaluated scenarios. Let $N_{matched}$ represent the number of what-if scenarios that successfully match an entry in EventsAssociation2012, and N_{total} represent the total number of evaluated scenarios. The accuracy is given by “Equation 6”.

$$Accuracy = \frac{N_{matched}}{N_{total}} * 100\% \quad (6)$$

This accuracy metric provides an indication of how effectively the what-if scenario generation and detection process mirrors real-world event associations.

We will call scenarios that match EventsAssociation2012 real-world what-if scenarios. In the next section, we will apply these real-world scenarios to a newly detected events dataset.

2.4 Event Association Detection via Real-World What-If Scenarios on a New Dataset

Our hypothesis is that “Social media-derived event detections can generate what-if scenarios using association rules from event topics, which can then be applied to assess their applicability for future events.” To test this hypothesis, we apply real-world what-if scenarios to a new dataset of detected events, following the outlined algorithm.

- Matching Events with Scenario Parts

Each what-if scenario has two main components:

the Left-Hand Side (*LHS*) and the Right-Hand Side (*RHS*). The *LHS* represents the initial event, while the *RHS* represents the subsequent event. To match these scenarios to the new set of detected events, we first use embeddings of the detected event topics and the scenario topic sets. We calculate the cosine similarity between the topic embeddings of the detected events and the topic sets of the scenario components. The detected events with the maximum similarity are selected as the matched events in the scenario. This process results in a mapping where a detected event, $eventA_{matched}$, corresponds to the *LHS* of the scenario, and another detected event, $eventB_{matched}$, corresponds to the *RHS*.

- Identifying the WI Part in the Matched Event
With the matched pair $eventA_{matched} \Rightarrow eventB_{matched}$ established according to the scenario's *LHS* \Rightarrow *RHS* relationship, the next step is to identify the What-If (*WI*) part within $eventA_{matched}$. The *WI* part represents a hypothetical or counterfactual condition within the initial event that could lead to the subsequent event. To identify the *WI* part, we compare the embeddings of various combinations of detected event topics in $eventA_{matched}$ with the embedding of the *WI* component of the scenario. The number of topics in each combination is based on the length of the *WI* in the original what-if scenario. The combination of topics that exhibits the highest similarity to the *WI* embedding is defined as the topic space for the counterfactual *WI* part. This topic space represents the set of conditions within $eventA_{matched}$ that could potentially trigger $eventB_{matched}$, thereby validating the applicability of the what-if scenario to newly detected events.

By following this algorithm, we can effectively apply real what-if scenarios to newly detected events, enabling us to explore the causal and associative dynamics of social media-derived events. This process allows for the practical evaluation of our hypothesis, demonstrating whether event-topic associations detected in the past can be used to predict and assess potential future events.

3 RESULTS

In this section, we present the findings of our study, encompassing three key aspects: the development of the EventsAssociation2012 dataset, the evaluation of what-if scenarios using this dataset, and the application of these scenarios to detect event associations on a new set of social media-derived events. The results

from each stage contribute to a comprehensive understanding of how event-topic associations in social media can be modeled, evaluated, and used for future event prediction.

3.1 Dataset for Event Association Detection

First, we describe the characteristics of the EventsAssociation2012 dataset, which was constructed by applying an LLM to the Events2012 dataset to generate event-to-event similarity scores, conditional probabilities, and textual explanations. This dataset, which contains both intra-event and inter-event associations, serves as the foundation for evaluating our what-if scenarios.

A pairwise comparison of 504 events was conducted within each respective category of the Events2012 dataset. This approach ensured that events in the "Sport" category, for example, were only compared to other events within the same category. The Events2012 dataset consists of eight categories: "Armed Conflicts & Attacks," "Arts, Culture & Entertainment," "Business & Economy," "Disasters & Accidents," "Law, Politics & Scandals," "Science & Technology," and "Sports." The result of EventsAssociation2012 is presented in Table 1.

Table 1: EventsAssociation2012 information.

	Event pairings	intra-event associations	inter-event associations
Number of associations	24,719	14	16

The example of inter-event association is presented in Table 2.

3.2 What-if Scenarios Evaluation

Next, we detail the evaluation of what-if scenarios, which involved matching these scenarios with the entries in EventsAssociation2012. We outline the criteria used for successful matching and assess the accuracy of these scenarios, providing insight into their effectiveness in capturing real-world event associations.

Since what-if scenarios are constructed from detected events, we needed to first match real-events to detected events. According to the steps described in Section 3.3, we detected events mapped to real events from Events2012 dataset. For example, the real event description is "Lebron and the Heat get-

Table 2: Inter-event association example.

$Event_A$: “Hurricane Sandy in the Bahamas.”	$Event_B$: “Tweets for Praying for people affected by the hurricane sandy.”
Similarity: 65%; Similarity reason: “Both describe reactions to Hurricane Sandy.”	
$P(A B)$: 30% $P(A B)$ reason: “People in Bahamas might be among those prayed for.”	$P(B A)$: 70% $P(B A)$ reason: “Prayers likely include people affected in multiple regions, including Bahamas.”
Category: Disasters & Accidents	Association type: inter-event association

ting their NBA championship rings” and corresponding detected event’s topic set is ‘heat’, ‘ring’, ‘ring ceremoni’, ‘miami heat’, ‘miami’, ‘championship’. Then, scenario components were mapped to detected events, and evaluated counterfactual what-if scenarios were received, see Table 3.

Table 3: What-if scenario example.

What-If scenario	aftermath, hurrican, news → damag, superstorm
$Event_{LHS}$	“Hurricane Sandy makes landfall near Atlantic City, New Jersey, with widespread flooding and at least 29 deaths in the Northeastern United States” discussed during 29.10.2012 - 31.10.2012.
$Event_{RHS}$	“Superstorm Sandy hits the east coast of the USA” discussed during 02.11.2012 - 02.11.2012.

During this experiment, we concentrated on detecting associations between two different events. For that purpose, we have created scenarios by Definitions 6 and 7. The results of the evaluation are presented in Table 4.

The analysis revealed that one-rule-based inter-event scenarios, with an accuracy of 26%, yield relatively better results than two-rules-based inter-event scenarios, with an accuracy of 3%. However, the result is a small number of real-world what-if scenarios. Out of 6,672 scenarios, only 8 were identified as real-world what-if scenarios from a possible 30. We also can see that inter-event scenarios could identify the intra-event associations. This outcome indicates a limitation in the current approach for what-if scenario identification, suggesting that improvements are necessary. Strategies for enhancing this process will be

Table 4: What-if scenarios evaluation results.

	One-rule-based inter-event scenarios	Two-rules-based inter-event scenarios
Parameters	$L = 2, M = 1, R = 2$	$L = 3, M = 2, R = 2$
Number of scenarios	6672	2284
Number of intra-event associations	3 out of 14	0 out of 14
Number of inter-event associations	5 out of 16	1 out of 16

discussed in the following section. Despite this limitation, these 8 identified what-if scenarios can still be applied to the new dataset to uncover potential event associations.

3.3 Event Associations on a New Dataset

Lastly, we apply the validated real-world what-if scenarios to a new dataset of detected events to explore the potential of using social media-derived event detections to forecast future associations. By matching the scenarios’ LHS (initial event) and RHS (subsequent event) with the newly detected events, we assess how well these scenarios can identify event associations, thereby validating our hypothesis about the predictive capabilities of what-if scenarios in the context of social media discussions.

Events were detected from Telegram messages between January 1, 2024, and May 31, 2024. A total of 389 events were identified, all belonging to the “Disasters” category. This category was selected for analysis because all real-world what-if scenarios fall under this specific category. The event detection was performed using the same methodology evaluated in (Mussina et al., 2022).

It is important to note that there are 8 real-world what-if scenarios, but this number represents unique event pairs. Generating what-if scenarios can result in multiple variations of topic sets in both the LHS and RHS components, even when they correspond to the same pair of events. In total, 128 scenarios were generated, describing these 8 unique event pairs.

From these 128 scenarios, we received 47 associations between events, of which 28 were unique, see Figure 2. Here the same text-embedding-3-small is used to generate vectors for detected events, D . When using this model with non-English languages, it can

still generate embeddings that capture some semantic information, but the quality and representational accuracy might be lower compared to English.

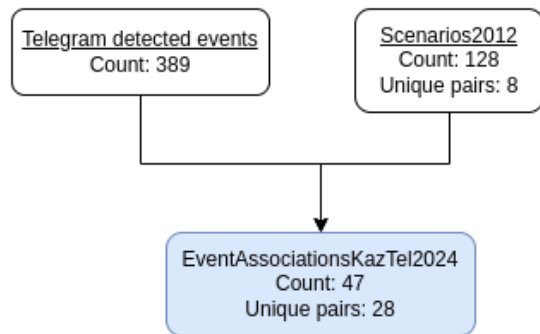


Figure 2: EventAssociationsKazTel2024 generation schema.

Next, after we found associations, we tried to find the *WI* part in $Event_{LHS}$ to explain why this event could lead to $Event_{RHS}$. As described in subsection 2.4 we found the *WI* part for each association. This resulting dataset was named EventAssociationsKazTel2024. Some of the event association with the highlighted *WI* part is presented in Table 5. Crawled data from Telegram is mostly written in Russian. Words are translated from Russian to English.

It can be seen, that when the topic “Department of Emergency Situations” appears in the event topic set, which is about the disaster itself, the associated event concentrates more on rescue operations.

4 DISCUSSION

This section discusses the results obtained and outlines the potential improvements for future research.

Firstly, further testing is necessary to identify intra-event scenarios as outlined in “Definition 4”. Currently, the focus has primarily been on detecting inter-event scenarios, based on the assumption that this approach would yield a larger number of scenarios. Expanding the focus to intra-event scenarios will provide a more comprehensive understanding of event associations.

Additionally, while associations between two events were successfully identified, future work may explore associations involving three events by treating the *WI* (What-If) component as a distinct event. This would require extending the size of the *WI* part in the scenarios. For this task, two-rules-based inter-event scenarios may offer a more suitable framework. However, this approach necessitates a larger set of real-world what-if scenarios derived from two-rules-based

Table 5: What-if scenarios evaluation results.

	LHS	WI	RHS
1	floor, epicenter, cut, Department of Emergency Situations, depth, register earthquake	Department of Emergency Situations	elimination, get off, rescuer, descent, Department of Emergency Situations, slope, residential building, search work
2	fire, fireman, observe, eliminate, cylinder, be installed, victim to suffer, ignition, private residential building, salon	fire, victim to suffer	disaster, operational, emergency, emergency situation, training, response
3	disaster, fire, occur, burn, fire, meter, district, operational, Ministry of Emergency Situations	disaster	need, today, situation, operational, monitoring, operational, medical assistance

inter-event associations. During initial experiments, it was not feasible to conduct all tests with every possible variation in the sizes of *L*, *WI*, and *RHS* in the scenario itemsets due to RAM limitations. To address this, future tests can be split into batches or run on a more powerful computing environment.

In the current study, the similarity between the detected event topic sets and scenario components was calculated using the cosine similarity of sentence embeddings. Since embeddings are influenced by the order of words, an alphabetical arrangement was used for consistency. However, future work could explore using all possible combinations of word order or implement a method for embedding calculation that considers sets of topics without regard to word sequence.

Furthermore, the relationship between support and confidence in one-rule-based inter-event scenarios can be represented in a matrix format, as shown in Table 6. This matrix could assist in identifying scenarios of various types, such as rare, popular, or common scenarios. This study primarily focused on generating rare scenarios; however, exploring differ-

ent scenario types in future research could provide valuable insights into the dynamics of event associations.

Table 6: One-rule-based inter-event scenario types matrix.

Scenario type	Support (WI_M)	Support ($Base_L$)	Support (RHS_R)
Popular	min	max	max
Rare	min	max	min
Common	max	max	max

5 CONCLUSIONS

This work presents a novel approach to event prediction by applying association rules to generate counterfactual what-if scenarios. Hypothetical scenarios are leveraged through association rule mining, allowing the methodology to systematically identify key patterns within event sequences and thereby facilitate the prediction of future events.

The study also introduces the EventsAssociation2012 dataset, which serves as the foundation for evaluating the accuracy and applicability of what-if scenarios. Through a detailed analysis using a Large Language Model (LLM) to generate event-to-event similarities and conditional probabilities, this work establishes criteria for matching scenarios with real-world events. The evaluation results demonstrate the potential of this approach for identifying both two-event and multi-event associations, providing a robust framework for understanding the interconnected nature of social media discussions.

Searching for causal relationships can be achieved by integrating association rules into counterfactual analysis. This study advances the modeling of causal relationships within event associations, offering a more precise and comprehensive method for predicting plausible alternative outcomes based on observed data. Additionally, the work highlights several areas for future improvement, including the identification of intra-event scenarios, exploring associations among three events, refining the What-If component, and implementing more advanced embedding techniques, which are key steps toward further strengthening the predictive capabilities of the proposed methodology.

ACKNOWLEDGEMENTS

This research has been supported by the Science Committee of the Ministry of Education and Sci-

ence of the Republic of Kazakhstan (Grant No. BR24993001) "Creation of a large language model (LLM) to maintain the implementation of Kazakh language and increase the technological progress".

REFERENCES

- Ali, F., Ali, A., Imran, M., et al. (2021). Traffic accident detection and condition analysis based on social networking data. *Accident Analysis & Prevention*, 151:105973.
- Bian, T., Xiao, X., Xu, T., et al. (2020). Rumor detection on social media with bi-directional graph convolutional networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 549–556.
- Daud, N., Ab Hamid, S., Saadoon, M., and Sahran, F. (2020). Applications of link prediction in social networks: A review. *Journal of Network and Computer Applications*, 166:102716.
- Eabrasu, M. (2008). A what if? fine-tuning the expectations of business simulation technology through the lens of philosophical counterfactual analysis. *Organization*, 30(4):694–711.
- Islam, M., Liu, S., Wang, X., and Xu, G. (2020). Deep learning for misinformation detection on online social networks: a survey and new perspectives. *Social Network Analysis and Mining*, 10(1):82.
- Kaliyar, R., Goswami, A., and Narang, P. (2021). Fakebert: Fake news detection in social media with a bert-based deep learning approach. *Multimedia Tools and Applications*, 80(8):11765–11788.
- McMinn, A. J., Moshfeghi, Y., and Jose, J. M. (2013). Building a large-scale corpus for evaluating event detection on twitter. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*, pages 409–418.
- Mussina, A., Aubakirov, S., and Trigo, P. (2022). Parametrized event analysis from social networks. *Scientific Journal of Astana IT University*, 10(10).
- Mussina, A., Trigo, P., and Aubakirov, S. (2023). Scenario generation with transitive rules for counterfactual event analysis. In *Proceedings of the 15th International Conference on Agents and Artificial Intelligence*, V.3, pages 1047–1051.
- Shalit, U., Johansson, F., and Sontag, D. (2017). Estimating individual treatment effect: Generalization bounds and algorithms. In *Proceedings of the 34th International Conference on Machine Learning*, pages 3076–3085.
- Tangcharoensathien, V., Calleja, N., Nguyen, T., et al. (2020). Framework for managing the covid-19 infodemic: methods and results of an online, crowd-sourced who technical consultation. *Journal of Medical Internet Research*, 22(6):e19659.
- Valdez, D., Ten Thij, M., Bathina, K., et al. (2020). Social media insights into us mental health during the covid-19 pandemic: Longitudinal analysis of twitter data. *Journal of Medical Internet Research*, 22(12):e21418.

- Yu, X., Mashhadi, A., Boy, J., Nielsen, R. C., and Hong, L. (2022). Causal impact model to evaluate the diffusion effect of social media campaigns. In *EUSSET*.
- Zhang, K., Yu, J., Shi, H., Liang, J., and Zhang, X. Y. (2023). Rumor detection with diverse counterfactual evidence. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 3321–3331.
- Zhang, Y., Cao, D., and Liu, Y. (2022). Counterfactual neural temporal point process for estimating causal influence of misinformation on social media. *Advances in Neural Information Processing Systems*, 35:10643–10655.

