Vol. 8, No. 3, June 2026

Journal of Information Systems and Informatics

Published By Asosiasi Doktor Sistem Informasi Indonesia

Context-Aware Disaster Cause Mining from Indonesian Online News Using GA-Optimized Apriori: A Forest and Land Fire Case Study

Qonitah Alia Puteri[1], Amalia Utamima[2]

[1,2] Department of Information Systems, Sepuluh Nopember Institute of Technology, Surabaya, Indonesia

Received:
October 11, 2025
Revised:
May 17, 2026
Accepted:
June 1, 2026
Published:
June 25, 2026

Corresponding Author:
Author Name*:
Qonitah Alia Puteri
Email*:
qonitaaliaputri@gmail.com

DOI:
10.63158/journalisi.v8i3.1656

Abstract. Online disaster news contains reported cause information, but the narratives are unstructured and difficult to use for systematic disaster risk analysis. This study develops a context-aware disaster cause mining framework to extract and analyze reported cause-context association patterns from Indonesian online news. The framework integrates text-based cause extraction, contextual enrichment using population density and meteorological variables, GA-optimized Apriori-based Association Rule Mining, and merged rule interpretation. Disaster news records were transformed into transactions containing disaster type, reported causes, population density category, three-day rainfall category, and maximum temperature category. From 742 final transaction records, the Apriori process generated 200 initial association rules. After filtering rules with reported causes and contextual attributes in the antecedent and disaster type in the consequent, 97 target rules were retained. The empirical analysis focused on forest and land fire as a case study, producing 24 rules and 5 merged rule patterns. The strongest merged pattern was related to land burning, with a merged support of 0.1173. The findings show that the framework can organize disaster narratives into interpretable reported cause-context association patterns for disaster risk analytics. However, the results should not be interpreted as verified causal evidence.

Keywords: Disaster Cause Mining, Association Rule Mining, Genetic Algorithm, Indonesian Online News, Forest and Land Fire

1. INTRODUCTION

Indonesia is highly exposed to various disaster risks due to its geographical, climatic, environmental, and socio-demographic conditions. Disaster events such as floods, landslides, forest and land fires, droughts, and water pollution frequently occur across different regions of the country [1], [2]. These events are associated not only with natural processes, but also with human activities, land-use change, environmental degradation, infrastructure limitations, and demographic pressure. Therefore, analyzing the relationship between disaster events, reported causes, and contextual conditions is important for supporting disaster risk reduction and sustainability-oriented disaster risk analytics.

Disaster-related information is widely available in online news articles published by national and local media. These articles often provide narrative descriptions of disaster events, affected locations, environmental conditions, and reported possible causes that may not be fully captured in structured disaster databases [3], [4]. For example, forest and land fire reports may mention land burning, dry weather, wind conditions, or local environmental characteristics as part of the event narrative. However, because online news articles are unstructured, the reported cause information contained in these narratives cannot be directly used for systematic analysis without further extraction and standardization [5]. In this study, the term reported causes refers to cause-related information described in news narratives, not verified causal mechanisms.

Previous studies have contributed to disaster analysis through structured environmental data analysis, disaster text mining, forest and land fire vulnerability prediction, and rule-based pattern discovery. Structured environmental studies can identify spatial, temporal, and hazard-related patterns, while text-based studies show that disaster reports and online news can provide valuable narrative information about disaster events. Other studies on forest and land fire have examined the role of human, environmental, meteorological, and land-use factors in fire vulnerability. However, these studies generally focus on separate analytical components, such as environmental pattern detection, textual information extraction, vulnerability prediction, or rule mining. A comparison of representative studies is presented in Table 1.

Table 1. Comparison of Related Work in Disaster Analysis

Study	Data & Method	Main Limitation
Zheng et al. [6]	Data: Gridded SPEI datasets from 1901 to 2015. Method: Improved association rule mining involving time-lagged rule generation and DBSCAN clustering to identify associated geographic zones.	The method identifies statistical patterns and probabilities but does not fully explain the underlying physical or climatic mechanisms behind the discovered associations.
Shidik et al. [3]	Data: Indonesian disaster-related online news and disaster information sources, including government websites and trusted news portals. The study developed an Indonesian disaster corpus with disaster-specific entity labels. Method: Named Entity Recognition using BiLSTM with random oversampling and optimizer comparison to identify disaster-related entities from Indonesian textual data.	The study focuses on disaster entity recognition, such as disaster type, location, scale, supplies, and casualties, but does not extract reported disaster causes or analyze cause-context association patterns. Therefore, the extracted textual information is not integrated with environmental or demographic context for rule-based disaster pattern interpretation.
Karurung et al. [7]	Data: Forest fire data in Indonesia from 2015 to 2022, combined with human, environmental, meteorological, and land-use/land-cover factors for Kalimantan, Sumatra, and Papua. Method: Forest fire vulnerability prediction using Random Forest and XGBoost, supported by feature	The study provides structured spatial prediction of forest fire vulnerability using remote sensing and environmental variables, but it does not analyze reported causes from online news narratives. As a result, it does not connect unstructured disaster

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

Study	Data & Method	Main Limitation
	importance, partial dependence plots, and vulnerability mapping.	narratives with cause-context association rules.
This research	Data: Indonesian online disaster news articles combined with extracted reported causes, population density category, three-day rainfall category, and maximum temperature category. Method: Text-based cause extraction and GA-optimized Apriori-based Association Rule Mining, followed by merged rule pattern interpretation.	This study focuses on forest and land fire as an empirical case study. The discovered patterns are exploratory associations derived from reported news narratives, not direct causal evidence. Broader multi-disaster validation using official disaster records or expert assessment is reserved for future work.

As shown in Table 1, previous studies have provided important contributions to disaster analysis, but several limitations remain. Structured environmental studies can identify statistical patterns from climate or hazard datasets, but they provide limited explanation of reported causes described in disaster narratives. Text-based studies can capture disaster-related information from reports or news articles, but they do not always transform reported causes into structured variables that can be analyzed together with environmental and demographic context. Meanwhile, Association Rule Mining can generate interpretable co-occurrence patterns among categorical attributes, but its results are sensitive to parameter selection, especially minimum support, confidence, and lift [8], [9]. Strict parameter settings may remove less frequent but meaningful disaster patterns, while loose parameter settings may generate excessive and less interpretable rules. Although Genetic Algorithm can be used to search for more suitable ARM parameter combinations [10], [11], existing disaster studies generally treat disaster narratives, contextual indicators, and optimized rule mining as separate analytical components. Therefore, an integrated approach is needed to connect reported cause extraction from disaster news, contextual enrichment, GA-optimized Apriori rule mining, and merged rule interpretation within a unified disaster cause mining framework.

Based on this gap, this study aims to develop a context-aware disaster cause mining framework for extracting and analyzing reported cause-context association patterns from Indonesian online news. Specifically, this study aims to transform unstructured disaster news into structured transaction data, integrate reported causes with environmental and demographic contextual attributes, optimize Apriori parameter selection using Genetic Algorithm, and interpret association patterns through raw and merged rule analysis. The framework uses disaster type, reported causes, population density category, three-day rainfall category, and maximum temperature category as the main transaction attributes. In this study, the discovered rules are interpreted as reported cause-context association patterns, not as direct causal evidence.

Forest and land fire was used as the empirical case study because it is closely related to both human-related reported causes and environmental conditions in Indonesian disaster narratives. This disaster type often appears together with reported land burning or land-clearing activities, while rainfall, temperature, and dry weather provide important surrounding context. In this study, forest and land fire was selected for detailed interpretation after the target-rule filtering and rule-quality evaluation stages because it produced the most suitable rule set based on rule volume, support count, confidence, lift, and interpretability. Therefore, the case study was not selected based on manual preference, but was determined from the post-filtering rule evaluation results. This case provides a suitable setting to demonstrate how the proposed framework links reported causes extracted from news narratives with contextual attributes.

The main contribution of this study is an integrated context-aware disaster cause mining framework that combines news-based reported cause extraction, contextual enrichment, GA-optimized Apriori-based Association Rule Mining, and merged rule pattern interpretation. The framework transforms Indonesian online disaster news into structured transaction data and supports the discovery of interpretable reported cause-context association patterns. Through the forest and land fire case study, this research demonstrates how unstructured disaster narratives and contextual data can be organized to support exploratory and sustainability-oriented disaster risk analytics.

2. METHODS

The proposed method was designed as a context-aware disaster cause mining framework that integrates unstructured disaster news with environmental and demographic context. As shown in Figure 1, the framework consists of three main stages: multi-source data input, data preparation and feature construction, and rule mining and interpretation. In the first stage, Indonesian online news articles are combined with meteorological indicators and population density data. In the second stage, the news text is preprocessed and disaster causes are extracted, while contextual attributes are integrated to construct the final transaction dataset. In the final stage, GA-optimized Apriori-based Association Rule Mining is applied to generate association rules, followed by post-processing to merge similar rules and support context-aware interpretation.

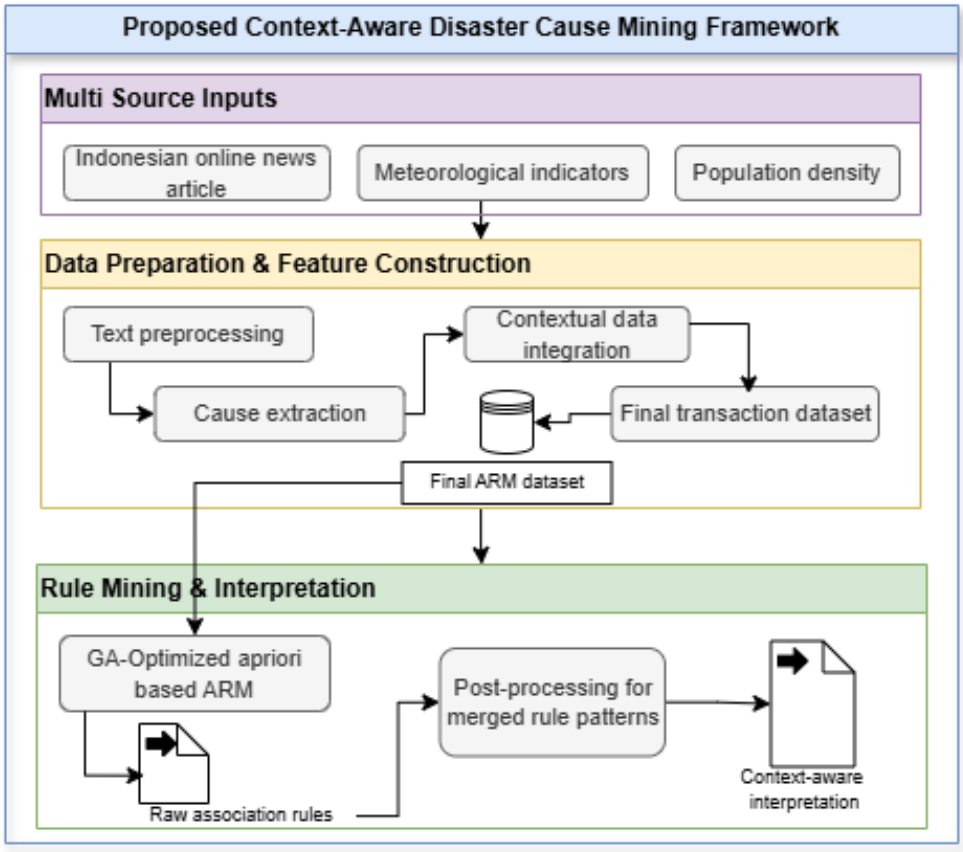

Figure 1. Proposed Context-Aware Disaster Cause Mining Framework

The detailed procedures of each stage are explained in the following subsections.

2.1. Research Framework

This research was conducted through a context-aware disaster cause mining framework, as shown in Figure 1. The framework consists of three main stages: multi-source input, data preparation and feature construction, and rule mining with interpretation. The first stage uses Indonesian online disaster news as the main textual source, supported by meteorological indicators and population density data. The second stage transforms the unstructured news data into structured analytical variables through text preprocessing, cause extraction, contextual data integration, and final transaction construction. The third stage applies GA-optimized Apriori-based Association Rule Mining to generate disaster cause patterns, followed by rule filtering and merged rule pattern construction.

The framework was designed to identify interpretable cause-context associations rather than to predict disaster occurrence. Therefore, the analysis focuses on how reported causes and contextual conditions appear together with specific disaster types. In this study, forest and land fire was used as the empirical case after the target-rule filtering and rule-quality evaluation stages because it produced the most suitable rule set based on rule volume, support count, confidence, lift, and interpretability. Therefore, the case study was determined from the post-filtering rule evaluation results rather than manual preference.

2.2. Data Sources and Final Variables

The data used in this study were obtained from Indonesian online news articles related to disaster events. Each article contains information about disaster type, location, date, title, and narrative description. The disaster categories in the overall dataset include flood, landslide, forest and land fire, drought, and water pollution. Online news was used because disaster reports often describe not only the event itself, but also reported possible causes, environmental conditions, and local context that may not be fully captured in structured disaster databases [12].

The online news articles were collected from Indonesian national and local news portals, including Detik, Kompas, and CNN Indonesia, covering disaster reports published between 2019 and 2026. The data collection used disaster-related keywords, including flood, landslide, forest and land fire, drought, and water pollution. Articles were included when they reported disaster events in Indonesia and contained usable information on disaster

type, location, date, and event narrative. Articles were excluded when they were not related to Indonesian disaster events, had inaccessible or incomplete content, duplicated the same report, or did not contain sufficient information for cause extraction and contextual matching. Potential duplicate reports were reduced based on similarities in disaster type, location, event date, and narrative content before the final transaction dataset was formed. This step was applied to reduce repeated reports referring to the same disaster event. Records with ambiguous locations, incomplete dates, inaccessible content, or insufficient information for cause extraction and contextual matching were not retained in the final transaction dataset.

To enrich the news-based disaster records, this study incorporated meteorological and demographic contextual data. Meteorological context was represented by three-day rainfall category and maximum temperature category, while demographic context was represented by population density category. These contextual variables were used to support the interpretation of reported cause-context association patterns. The data sources used in this study are summarized in Table 2.

Table 2. Data Sources Used in the Study

Data Source	Data Used	Role in Analysis	Output Variable
Indonesian online news articles	Disaster type, location, date, title, and narrative text	Main source for disaster event identification and cause extraction	disaster_type, cause_1, cause_2, cause_3
Meteorological indicators	Three-day rainfall category	Short-term rainfall context associated with the disaster record	rain_category
Meteorological indicators	Maximum temperature category	Temperature context associated with the disaster record	temp_max_category

Data Source	Data Used	Role in Analysis	Output Variable
Population density data	Density category of disaster location	Demographic context of the affected area	density_category
Final integrated dataset	Standardized disaster, cause, and contextual variables	Input dataset for Association Rule Mining	Final ARM transaction dataset

The dataset preparation process was conducted in several stages before the final transaction dataset was formed. The initial dataset consisted of 2,139 collected news articles. After filtering and deduplication, 1,580 articles were retained. The dataset was then processed through full-text checking, date parsing, non-Indonesian article removal, merging, deduplication, cause extraction, and contextual data integration. The summary of data preparation is presented in Table 3.

Table 3. Summary of Data Preparation Stages

Data Preparation Stage	Records
Initial collected news articles	2139
After filtering and deduplication	1580
After removing non-Indonesian records	1579
After final merging and deduplication	1572
Records with explicit reported causes	782
Final transaction dataset with complete contextual variables	742

The initial dataset consisted of 2,139 collected online news articles, after initial filtering and duplicate reduction, 1,580 articles were retained. One non-Indonesian record was removed, resulting in 1,579 Indonesian disaster news records. After final merging and deduplication, 1,572 records remained. The cause extraction process identified 782 records with explicit reported causes, consisting of 519 single-cause records and 263 multi-cause records, while 790 records did not contain explicit cause information and were not used for rule mining. After integrating contextual variables and retaining

records with usable disaster type, reported cause, population density, rainfall, and temperature information, the final transaction dataset used in this journal analysis consisted of 742 records.

This study used the final cleaned and integrated disaster dataset as the basis for Association Rule Mining. The initial cleaning and hazard integration stages were not repeated because the analysis focused on the final transaction dataset prepared for rule mining. The final variables used in the transaction dataset were disaster_type, cause_1, cause_2, cause_3, density_category, rain_category, and temp_max_category. The rain_category variable represents only the three-day rainfall category, while the seven-day rainfall category was excluded to avoid redundant rainfall indicators and maintain a consistent feature set for the final analysis.

2.3. Text-Based Cause Extraction

Text-based cause extraction was conducted to transform unstructured disaster news narratives into standardized reported-cause variables [13]. Before extraction, the news text was processed through normalization, sentence splitting, irrelevant-content removal, and standardization of disaster-related terms [12]. This stage was necessary because online news articles often contain varied writing styles, repeated information, and narrative descriptions that are not directly suitable for structured analysis.

The extraction process followed a three-layer approach, consisting of keyword and regular expression matching, contextual cue rules, and semantic rescue using multilingual Sentence-BERT. The first layer was used to detect explicit cause expressions based on a predefined dictionary and regex patterns. The second layer identified cause-related sentences using contextual cue terms such as *akibat*, *karena*, *dipicu*, *disebabkan*, *lantaran*, *pemicu*, and *penyebab* [14]. The third layer used multilingual Sentence-BERT as a semantic rescue mechanism to support implicit cause identification when the first two layers produced weak or incomplete results. The semantic rescue layer used the sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2 model, which was not fine-tuned and was used only to measure semantic similarity between candidate sentences and predefined cause-label prototypes.

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

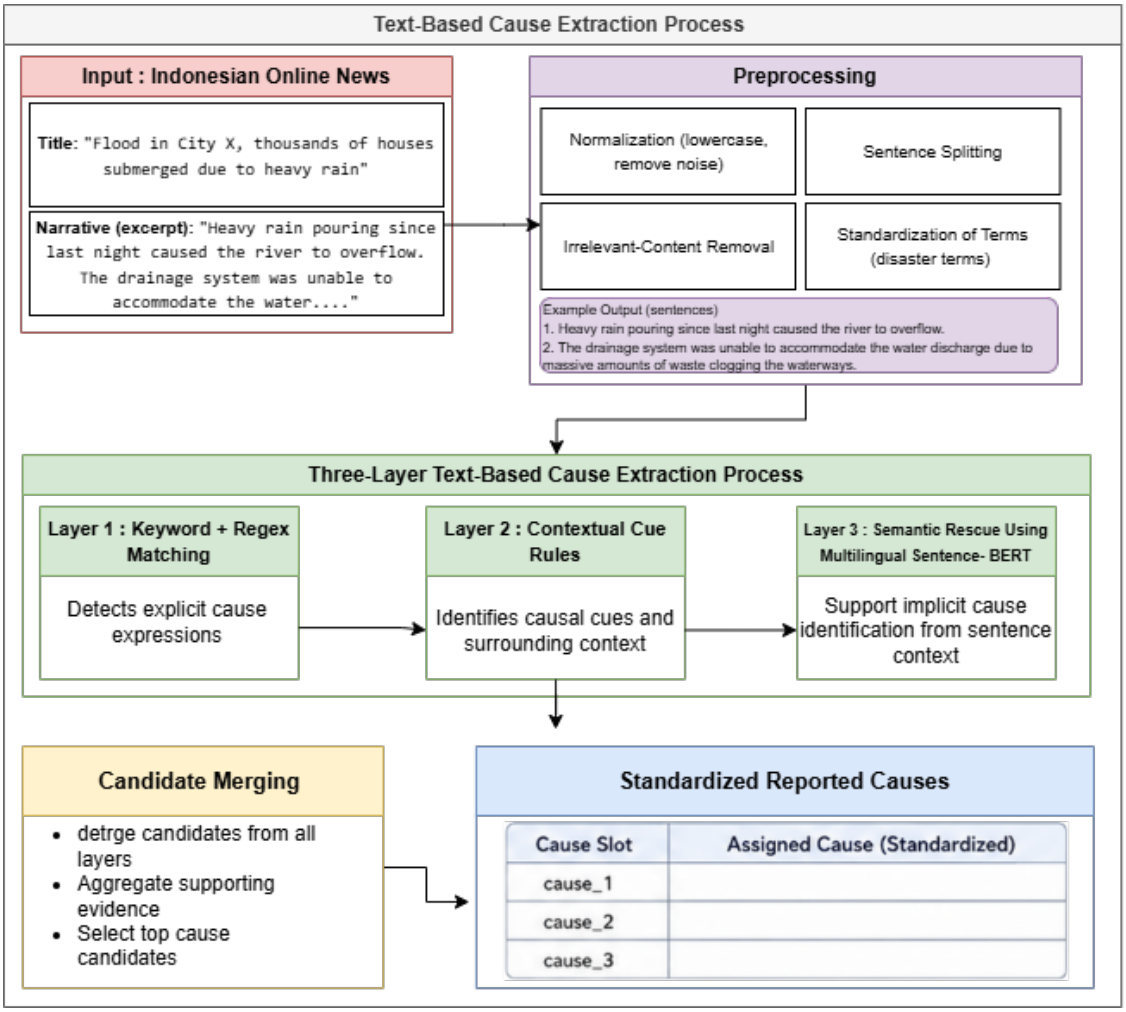

Figure 2. Text Based Cause Extraction Process

After candidate causes were identified, the results from the three extraction layers were merged, ranked, and manually checked by the authors against the original news narrative and the predefined cause-label schema. This manual checking was conducted to verify whether the extracted label was consistent with the textual evidence and the disaster category. The standardized reported-cause labels covered meteorological causes, infrastructure-related causes, land-use and environmental causes, anthropogenic causes, and pollution-related causes. The full standardized cause-label schema is provided in table 4. The final reported causes were stored in three fields: cause_1, cause_2, and cause_3, allowing one article to contain more than one reported cause. In the Association Rule Mining stage, these three cause fields were treated equally, meaning that their order did not indicate priority or causal strength.

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

Table 4. Standardized Reported-Cause Label Schema

Disaster Type	Standardized Cause Labels
Flood	heavy_rain, poor_drainage, waste_clogging, river_overflow, levee_failure, upstream_water_flow, coastal_tide, river_sedimentation, flow_obstruction
Forest_land_fire	land_burning, human_activity, prolonged_dry_season, hot_windy_weather, el_nino
landslide	heavy_rain, steep_slope, unstable_soil, deforestation
water_pollution	industrial_waste, domestic_waste, mining_activity, oil_spill, waste_contamination
drought	Prolonged_dry_season, reduced_water_discharger, el_nino, hot_windy_weather

2.4. Contextual Data Integration

After the reported causes were extracted from news narratives, each disaster record was enriched with contextual attributes. The contextual variables used in this study consisted of population density category, three-day rainfall category, and maximum temperature category. Population density was used to represent demographic context, while rainfall and maximum temperature were used to represent short-term meteorological conditions associated with each disaster record [15].

Population density data were obtained from the 2024 provincial population statistics dataset, which includes population size, annual population growth rate, population distribution percentage, population density, and sex ratio by province [16]. In this study, only the population density variable was used. Population density was included as a demographic contextual variable because it is commonly used in disaster vulnerability and exposure analysis, and previous studies have shown that population density can be associated with disaster damage at the community level [17]. The density value was matched to each disaster record based on the normalized province name. The population density category was not determined using a national administrative threshold. Instead, the category was generated using the tertile method based on the distribution of population density values in the research dataset[18]. Therefore, the categories low, medium, and high represent relative density levels within the dataset used in this study.

Meteorological variables were obtained from the Open-Meteo Historical Weather API using the latitude, longitude, and event date of each disaster record [19], [20] Latitude and longitude were obtained from the normalized disaster location during the location standardization process. When exact coordinates were unavailable, the normalized city or province-level location was used as the spatial reference for retrieving meteorological data. For rainfall, daily rainfall data were retrieved and accumulated over a three-day time window ending on the event date. The resulting accumulated value was used to form the rain_category variable. For maximum temperature, the daily maximum temperature at 2 meters was retrieved for the event date and stored as the raw maximum temperature value before being categorized into low, medium, and high. Similar to population density, the categories for three-day rainfall and maximum temperature were generated using tertile-based categorization from the distribution of valid values in the final research dataset. The tertile thresholds were calculated only from non-missing values, as shown in Table 5. In the final analysis, only the three-day rainfall category was used as the rainfall indicator, while the seven-day rainfall variable was excluded to avoid redundant rainfall indicators and maintain a consistent feature set.

Table 5. Contextual Variable Category Thresholds

Variable	Low	Medium	High
Population density	≤ 793.0 persons/km²	> 793.0 – ≤ 1359.0 persons/km²	> 1359.0 persons/km²
Three-day rainfall	≤ 6.9 mm	> 6.9 – ≤ 21.0 mm	> 21.0 mm
Maximum temperature	≤ 29.4 °C	> 29.4 – ≤ 31.6 °C	> 31.6 °C

Records with missing or unmatched contextual information, ambiguous locations, incomplete event dates, or unavailable coordinates were not included in the final transaction dataset because they could not be reliably matched with demographic or meteorological data. After cause extraction, 782 records contained explicit reported causes. After integrating contextual variables and retaining only records with usable disaster type, reported cause, population density, rainfall, and maximum temperature information, the final transaction dataset consisted of 742 records. The final variables used in the transaction dataset are shown in Table 6.

Table 6. Final Variables Used in the Association Rule Mining Dataset

Variable	Role in Analysis	Description
disaster_type	Consequent	Disaster category associated with the event record
cause_1	Antecedent	First extracted reported cause from the news narrative
cause_2	Antecedent	Second extracted reported cause, if available
cause_3	Antecedent	Third extracted reported cause, if available
density_category	Antecedent	Population density category of the disaster location
rain_category	Antecedent	Three-day rainfall category used as the final rainfall indicator
temp_max_category	Antecedent	Maximum temperature category associated with the disaster record

The final dataset was then transformed into a transaction format for Association Rule Mining. Each transaction represents one disaster record and contains reported-cause items, contextual items, and one disaster type item. The fields cause_1, cause_2, and cause_3 were treated equally to reduce positional bias. Therefore, the order of the cause fields did not indicate priority or causal strength. In this study, the antecedent items consist of reported causes and contextual attributes, while the consequent item consists only of the disaster type. This structure allows the generated rules to be interpreted as reported cause-context association patterns rather than direct causal evidence.

2.5. GA-Optimized Apriori Association Rule Mining

Association Rule Mining was used to identify interpretable relationships among reported causes, contextual attributes, and disaster types. The Apriori algorithm was selected because the final dataset was represented as categorical transaction data, making it suitable for frequent itemset generation and association rule extraction [21]. Rule quality was measured using support, confidence, and lift. Support represents the frequency of an itemset in the dataset, confidence represents the reliability of a rule, and lift indicates whether the association is stronger than random co-occurrence [22].

Because Apriori results are sensitive to parameter settings, a Genetic Algorithm was used to optimize the Apriori parameters. Previous studies have shown that evolutionary computation can support Association Rule Mining by searching for better rule or parameter combinations, including parameter settings used in frequent itemset and association rule generation [10], [11]. The optimized parameters included minimum support, minimum confidence, minimum lift, and maximum itemset length. This optimization was intended to reduce the subjectivity of manual parameter selection and to search for a parameter configuration that balances rule quantity, statistical strength, and interpretability. Each individual in the GA population represented one candidate Apriori parameter configuration. The configuration used in this study is shown in Table 7.

Table 7. Genetic Algorithm Configuration

Parameter	Value
Population size	18
Number of generations	12
Elite size	4
Tournament size	4
Mutation rate	0.25
Crossover application	Applied to each non-elite offspring
Parameter mixing probability	0.50
Averaging probability for numeric parameters	0.50
Random seed	42

The fitness function considered several rule-quality indicators, including the number of valid target rules, number of strong rules, average confidence, average lift, average support count, and the presence of contextual attributes. The fitness score was calculated as a weighted combination of these indicators, where rule count, strong rule count, confidence, lift, support count, and contextual coverage were used to evaluate the quality of each candidate parameter set. The fitness score was calculated using the following weighted formula as shown in Equation 1.

$$\begin{aligned} fitness = {} & 0.20\,(rule\ count\ score) + 0.25(strong\ rule\ score) \\ & + 0.20\,(confidence\ score) + 0.20(lift\ score) \\ & + 0.10\,(support\ count\ score) \\ & + 0.05\,(context\ coverage\ score) \end{aligned} \quad (1)$$

If the number of target rules was less than five, the fitness score was multiplied by 0.5 as a penalty to avoid selecting parameter sets that produced too few interpretable rules. Each component was normalized before being combined into the final fitness score. The rule count score measured the number of valid target rules, the strong rule score measured the number of rules satisfying the minimum strength criteria, the confidence and lift scores represented average rule reliability and association strength, the support count score represented transaction coverage, and the context coverage score measured the presence of contextual attributes in the generated rules. This fitness design was used to ensure that the selected parameter set did not only produce many rules, but also generated rules with sufficient strength, coverage, and contextual relevance. The GA process applied tournament selection, elitism, crossover, and mutation to search for the parameter set with the highest fitness value. The pseudo-code of the GA-optimized Apriori process is presented in Figure 3.

```
Input  : Transaction dataset T; parameter ranges for minimum support,
         minimum confidence, minimum lift, and maximum itemset length
Output : Optimized Apriori parameter set P*

1. Initialize a population of candidate Apriori parameter sets
2. For each generation do
3.     For each candidate parameter set P in the population do
4.         Generate frequent itemsets using Apriori with P
5.         Generate association rules using the selected thresholds
6.         Filter rules based on the required rule structure
7.         Evaluate the fitness of P using rule-quality indicators
8.     End for
9.     Select high-performing candidates using tournament selection
10.    Preserve elite candidates for the next generation
11.    Apply crossover to generate new candidate parameter sets
12.    Apply mutation to maintain search diversity
13. End for
14. Return the parameter set P* with the highest fitness value
```

Figure 3. GA-Optimized Apriori Parameter Selection Pseudocode

The analysis was implemented in Python using pandas and NumPy for data processing, re for regular expression matching, sentence-transformers for Sentence-BERT representation, mlxtend for Apriori and association rule extraction, and NetworkX/matplotlib for network visualization and the selected parameter set was then used to rerun Apriori and generate the final association rules. Since this study focuses on reported cause-context interpretation, the generated rules were filtered so that the antecedent contained only reported causes and contextual attributes, while the consequent contained only the disaster type. This rule structure allows the generated rules to be interpreted as reported cause-context association patterns related to disaster events, rather than as direct causal evidence.

2.6. Rule Filtering and Merged Rule Patterns

After the GA-optimized Apriori process generated association rules, the rules were filtered according to the analytical objective of this study. Only rules with reported causes and contextual attributes in the antecedent and disaster type in the consequent were retained. Rules were excluded if the antecedent contained a disaster item, if the consequent contained more than one item, or if the antecedent did not contain both reported-cause and contextual attributes. This filtering was applied so that the retained rules could be interpreted as reported cause-context association patterns [23].

For the forest and land fire case study, the filtered raw rules were further summarized into merged rule patterns. Raw rules were merged when they had the same disaster consequent and similar main reported-cause structure in the antecedent. Contextual attributes, such as population density, rainfall category, and maximum temperature category, were summarized as contextual variations within each merged pattern. A rule variant refers to one raw association rule included in a merged pattern. Therefore, the number of rule variants shows how many raw rules were grouped under the same main pattern.

Merged support was calculated based on the coverage of each merged pattern in the final transaction dataset. In other words, merged support was obtained by dividing the number of unique transactions covered by a merged pattern by the total number of transactions in the final dataset. This approach was used to avoid double counting transactions that may satisfy more than one raw rule within the same merged pattern.

3. RESULTS AND DISCUSSION

3.1. Final Transaction Dataset

The final dataset used in this study consists of disaster records that had been transformed into transaction data for Association Rule Mining. Each transaction contains standardized reported causes, contextual attributes, and one disaster type. The contextual attributes include population density category, three-day rainfall category, and maximum temperature category [7]. In the final analysis, only the three-day rainfall category was used as the rainfall indicator, while the seven-day rainfall category was excluded to maintain feature consistency and avoid redundant rainfall information.

The final transaction dataset contains 742 records across five disaster categories. As shown in Table 8, flood has the highest number of records with 276 transactions, followed by water pollution with 160 records and forest and land fire with 136 records. Landslide and drought contain 99 and 71 records, respectively. Although the dataset covers multiple disaster types, the detailed rule interpretation in this study focuses on forest and land fire as the empirical case study.

Table 8. Disaster Category Distribution in the Final Transaction Dataset

Information	Number of Records
Flood	276
Water pollution	160
Forest and land fire	136
Landslide	99
Drought	71
Total	742

Although the dataset covers multiple disaster types, the detailed rule interpretation in this study focuses on forest and land fire as the empirical case study. This category was selected in the rule interpretation stage because it produced valid and interpretable cause-context association rules after the GA-optimized Apriori filtering process. To illustrate how unstructured disaster narratives were transformed into structured reported-cause variables, examples of text-based cause extraction results are presented in Table 9. These examples show how reported causes identified from news narratives

were standardized into cause_1, cause_2, and cause_3. The excerpts are shown only for readability, while the actual extraction process used the available title, excerpt, and full article narrative.

Table 9. Examples of Text-Based Cause Extraction Results

News Narrative Excerpt	Disaster Type	cause_1	cause_2	cause_3
"Mayor Inspects Margonda Floods, Orders Trash Cleanup and Additional Drainage…"	flood	heavy_rain	waste_clogging	levee_failure
"Wildfires in PALI-Muba Optimally Extinguished, Area Remains Smoky.."	Forest land fire	prolonged_dry_season	land_burning	-
"Pohuwato Tap Water Turns Muddy Brown Due to Illegal Gold Mining…"	Water pollution	mining_activity	waste_contamination	-

The extracted reported causes were then used as cause items in the Association Rule Mining process. In the transaction structure, reported causes and contextual attributes were placed as antecedent items, while disaster type was placed as the consequent item. A dash indicates that no additional reported cause was identified for that cause slot. This structure allows the generated rules to be interpreted as reported cause-context association patterns rather than direct causal evidence.

3.2. GA-Optimized Apriori Parameter Selection and Rule Generation

The Apriori-based Association Rule Mining process was optimized using a Genetic Algorithm. The optimization process selected the parameter configuration that balanced rule quantity, statistical strength, and interpretability. The optimized parameters consisted of minimum support, minimum confidence, minimum lift, and maximum itemset length. The selected GA-optimized Apriori parameters are shown in Table 10.

Table 10. Selected GA-Optimized Apriori Parameters

Parameter	Value
Minimum support	0.0122
Minimum confidence	0.6168
Minimum lift	3.8428
Maximum itemset length	4

Using the selected parameters, the Apriori process generated 752 frequent itemsets and 200 initial association rules. The initial rules were then filtered according to the analytical structure of this study, where the antecedent must contain reported causes and contextual attributes, while the consequent must contain only one disaster type. After filtering, 97 target rules were retained. From these target rules, 24 rules were associated with forest and land fire and were used for detailed case-study interpretation. These 24 rules were then summarized into 5 merged rule patterns to reduce repetitive interpretation. The rule generation and filtering process is summarized in Table 11.

Table 11. Summary of Rule Generation and Filtering Process

Stage	Description	Output
Frequent itemset generation	Item combinations generated by Apriori using the selected GA-optimized parameters	752
Initial association rule generation	Association rules generated from the frequent itemsets before analytical filtering	200
Target rule filtering	Rules retained after applying the required rule structure	97
Case-study rule selection	Target rules with forest and land fire as the consequent	24
Merged rule pattern construction	Similar forest and land fire rules grouped into merged patterns	5

To evaluate the contribution of GA-based parameter selection, the GA-optimized Apriori result was compared with standard Apriori using manually selected thresholds, as shown in Table 12. The standard Apriori configuration used a minimum support of 0.0100,

minimum confidence of 0.5000, minimum lift of 1.0000, and maximum itemset length of 4. This comparison was used to examine whether GA optimization produced a more selective and interpretable rule set than manually selected Apriori parameters.

Table 12. Comparison between Standard Apriori and GA-Optimized Apriori

Indicator	Standard Apriori	GA-Optimized Apriori
Min support	0.0100	0.0122
Min confidence	0.5000	0.6168
Min lift	1.0000	3.8428
Max itemset length	4	4
Target rules	262	97
Forest_land_fire rules	36	24
Avg. confidence	0.9338	0.9619
Avg. lift	3.942	5.378
Fitness	0.8911	0.9428

Table 12 shows that standard Apriori generated a larger number of target rules, with 262 target rules and 36 forest and land fire rules. However, the GA-optimized Apriori produced a more selective rule set, with 97 target rules and 24 forest and land fire rules, while achieving higher average confidence, average lift, and fitness value. This indicates that GA optimization helped reduce excessive rule generation and selected a parameter configuration that better balanced rule quantity, association strength, and interpretability. The GA-optimized parameter set was then used as the basis for the subsequent forest and land fire rule analysis. Forest and land fire was selected as the empirical case study after the target-rule filtering and rule-quality evaluation stages because it produced the most suitable rule set based on rule volume, support count, confidence, lift, and interpretability. Therefore, the case study was determined from the post-filtering rule evaluation results rather than manual preference.

3.3. Raw Association Rules for Forest and Land Fire

From the 97 target rules generated by the GA-optimized Apriori process, 24 rules were associated with forest and land fire as the consequent. These rules were selected for detailed interpretation because they satisfied the required rule structure, where the antecedent contains reported causes and contextual attributes, while the consequent

contains only the disaster type. Table 13 presents the complete list of forest and land fire association rules. Since all rules in this table have the same consequent, namely disaster:forest_land_fire, the consequent column is not repeated to improve readability.

Table 13. Association Rules for Forest and Land Fire

No	Antecedent	Support	Support Count	Confidence	Lift
1.	cause:land_burning | density:low	0.1051	78	1.0	5.4559
2.	cause:land_burning | density:low | tempmax:high	0.0580	43	1.0	5.4559
3.	cause:land_burning | density:low | rain:low	0.0418	31	1.0	5.4559
4.	cause:land_burning | tempmax:high	0.0647	48	1.0	5.4559
5.	cause:land_burning | rain:low | tempmax:high	0.0377	28	1.0	5.4559
6.	cause:land_burning | density:low | tempmax:medium	0.0363	27	1.0	5.4559
7.	cause:land_burning | density:low | rain:high	0.0323	24	1.0	5.4559
8.	cause:land_burning | density:low | rain:medium	0.0310	23	1.0	5.4559
9.	cause:land_burning | rain:high | tempmax:medium	0.0243	18	1.0	5.4559
10.	cause:land_burning | rain:medium | tempmax:high	0.0216	16	1.0000	5.4559
11.	cause:land_burning | rain:low	0.0472	35	1.0000	5.4559
12.	cause:land_burning | tempmax:medium	0.0404	30	1.0000	5.4559
13.	cause:prolonged_dry_season | density:low | tempmax:high	0.0270	20	0.8696	4.7442
14.	cause:land_burning | rain:high	0.0364	27	1.0000	5.4559

No	Antecedent	Support	Support Count	Confidence	Lift
15.	cause:hot_windy_weather | density:low | tempmax:high	0.0229	17	0.8947	4.8816
16.	cause:land_burning | rain:medium	0.0337	25	1.0000	5.4559
17.	cause:prolonged_dry_season | density:low | rain:medium	0.0175	13	0.8667	4.7284
18.	cause:land_burning | cause:prolonged_dry_season | density:low	0.0189	14	1.0000	5.4559
19.	cause:hot_windy_weather | cause:land_burning | density:low	0.0175	13	1.0000	5.4559
20.	cause:prolonged_dry_season | density:low	0.0404	30	0.7317	3.9921
21.	cause:hot_windy_weather | cause:land_burning | tempmax:high	0.0135	10	1.0000	5.4559
22.	cause:hot_windy_weather | density:low	0.0297	22	0.7586	4.1389
23.	cause:hot_windy_weather | density:low | rain:low	0.0149	11	0.7333	4.0010
24.	cause:prolonged_dry_season | rain:medium	0.0202	15	0.7894	4.3073

The raw association rules show that land_burning is the most dominant reported cause associated with forest and land fire in the generated rule set. The highest-support rule combines land_burning and density:low, with a support count of 78, confidence of 1.0000, and lift of 5.4559. This means that all transactions containing this antecedent pattern in the final dataset were associated with forest and land fire. However, high confidence values, including confidence of 1.0000, should be interpreted cautiously because they may be influenced by the relatively small forest and land fire subset and by the strict target-rule filtering criteria. Therefore, the rule should be interpreted as a reported

cause-context association pattern derived from online news narratives, not as direct causal evidence.

3.4. Merged Rule Pattern Analysis

Although the raw association rules provide detailed rule-level results, several rules contain similar meanings and differ only in their contextual attributes. Therefore, the 24 forest and land fire raw rules were summarized into 5 merged rule patterns. This process was conducted to reduce repetitive interpretation and to provide a clearer view of the dominant reported cause-context patterns associated with forest and land fire.

The merged rule patterns were constructed by grouping raw rules that had the same disaster consequent and similar main reported-cause structure in the antecedent. Contextual attributes, such as population density, rainfall category, and maximum temperature category, were then summarized as contextual variations within each merged pattern. A rule variant refers to one raw association rule included in a merged pattern. Therefore, the number of rule variants shows how many raw rules were grouped under the same merged pattern.

Merged support was calculated based on the number of unique transactions covered by each merged pattern. In other words, merged support was obtained by dividing the number of unique transactions covered by a merged pattern by the total number of transactions in the final dataset. This calculation differs from individual raw rule support because it avoids double counting transactions that may satisfy more than one raw rule within the same merged pattern. To show how the five merged patterns were derived from the 24 raw rules, Table 14 also provides the raw rule IDs included in each merged pattern.

Table 14. Merged Rule Patterns for Forest and Land Fire

Pattern	Main Reported Cause	Contextual Variations	Rules IDs	Rule Variants	Sup Count	Merged Support
P1	land_burning	density=low; rain=high/low/medium;	R1–R12, R14, R16	14	87	0.1173

Pattern	Main Reported Cause	Contextual Variations	Rules IDs	Rule Variants	Sup Count	Merged Support
		tempmax=high/medium				
P2	hot_windy_weather, land_burning	density=low; tempmax=high	R19, R21	2	14	0.0189
P3	land_burning, prolonged_dry_season	density=low	R18	1	14	0.0189
P4	hot_windy_weather	density=low; rain=low; tempmax=high	R15, R22, R23	3	36	0.0485
P5	prolonged_dry_season	density=low; rain=medium; tempmax=high	R13, R17, R20, R24	4	35	0.0472
Total	-	-		24	-	-

Table 14 shows that land_burning is the most dominant merged pattern, appearing in 14 rule variants with a support count of 87 and a merged support of 0.1173. This pattern is consistently associated with forest_land_fire and appears together with several contextual variations, including low population density, different three-day rainfall categories, and medium to high maximum temperature categories. This finding indicates that land burning is the central reported cause pattern in the forest and land fire records analyzed in this study.

Other merged patterns provide additional context. The pattern involving hot_windy_weather shows that weather-related conditions also appear in forest and land fire rules, especially when combined with low density and high maximum temperature. The pattern involving prolonged_dry_season indicates that dry seasonal conditions are also relevant, although they appear in fewer rule variants than land burning. The repeated

appearance of density=low does not indicate that low population density causes forest and land fire. Rather, it indicates that low-density locations frequently co-occurred with reported land burning, dry weather, or prolonged dry-season conditions in the analyzed news-based transaction dataset. Overall, the merged patterns provide a compact and interpretable summary of reported cause-context associations, not direct causal evidence.

3.5. Network-Based Interpretation

To support the interpretation of the merged rule patterns, the association results were visualized using a network graph. The network graph represents the relationship between the selected disaster type, main reported causes, and contextual attributes. In this visualization, forest and land fire is positioned as the disaster outcome, while the connected nodes represent the cause and contextual attributes that appear in the association rules.

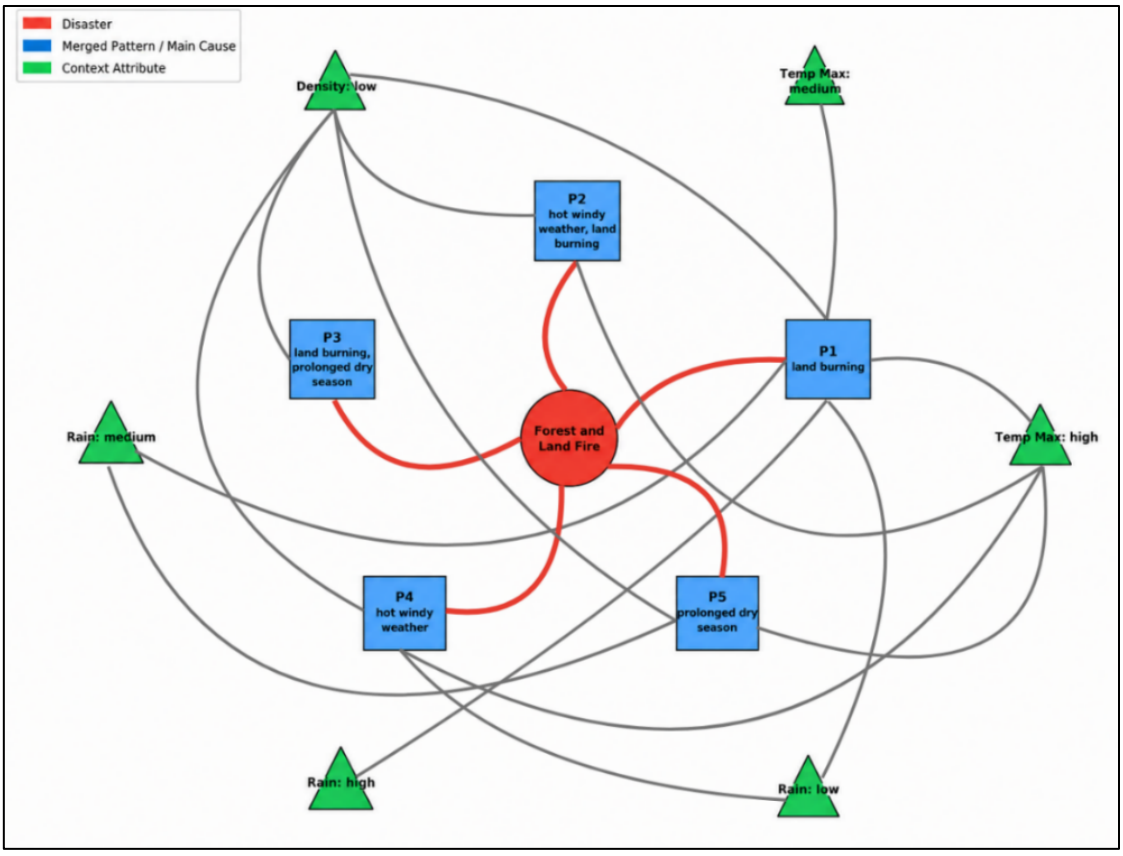

Figure 4. Network Graph of Forest and Land Fire Association Patterns

The network graph consists of three types of nodes. The red circular node represents the selected disaster category, namely forest and land fire. The blue square nodes represent the five merged rule patterns, P1 to P5, which summarize the main reported-cause structures. The green triangular nodes represent contextual attributes, including population density, rainfall category, and maximum temperature category.

In this visualization, the red edges connect forest and land fire with the merged pattern nodes, indicating that each pattern is associated with the selected disaster category. The gray edges connect each merged pattern with its contextual attributes. Therefore, the graph should be read as a structural visualization of the merged rule patterns rather than as a causal diagram. The edge connections indicate the presence of association relationships in the merged patterns, while the numerical values of rule variants, support count, and merged support are provided in Table 14. The network graph is unweighted; edge thickness is used only for visual emphasis and does not represent support, confidence, frequency, or causal strength.

Figure 4 shows that P1, which represents land_burning, is the most prominent merged pattern connected to forest and land fire. This is consistent with Table 14, where P1 has the largest number of rule variants and the highest merged support. The graph also shows that density=low is connected to all merged patterns, indicating that low population density frequently appears as a contextual attribute in the forest and land fire rules. Meanwhile, rainfall and maximum temperature categories provide additional meteorological context across several patterns.

The network also shows that hot_windy_weather and prolonged_dry_season contribute to additional merged patterns. Although these patterns appear in fewer rule variants than land_burning, they provide supporting environmental context for interpreting forest and land fire association patterns. Overall, the network visualization complements the merged pattern table by showing how reported causes and contextual attributes are connected within the forest and land fire rule structure.

3.6. Discussion and Limitations

The results show that land_burning is the most dominant reported cause pattern associated with forest and land fire in the analyzed news-based dataset. This finding is

consistent with previous Indonesian forest and land fire studies, which show that forest fire vulnerability is related to human activity, meteorological conditions, and environmental context [7]. However, the dominance of land_burning in this study should be interpreted as a reported narrative pattern, not as direct evidence that land burning caused all forest and land fire events. Online news may emphasize visible or socially salient causes, such as land burning, while less visible environmental, administrative, or land-management factors may be underreported.

The interpretation of confidence and lift also requires caution. Forest and land fire represents 136 of 742 transactions, or approximately 18.3% of the final dataset. Therefore, high confidence values, including rules with confidence of 1.0000, may be influenced by the relatively small forest and land fire subset and by the strict rule-filtering structure used in this study. Lift values indicate that the observed associations are stronger than random co-occurrence relative to the base rate of forest and land fire, but they should not be interpreted as predictive performance or causal strength. Similarly, the frequent appearance of density=low should be understood as a contextual association, possibly reflecting that forest and land fire reports often occur in less densely populated, rural, plantation, or peripheral areas, rather than as evidence that low population density causes forest and land fire.

This study also has limitations related to the use of online news data. News reports may contain media-reporting bias, regional coverage bias, incomplete cause statements, and possible event-duplication bias. Although duplicate reports were reduced during data preparation, some reports may still describe similar disaster situations from different media perspectives. External validation using official disaster records or expert assessment was not fully conducted in this stage. Therefore, future work should validate the framework using official BNPB/BMKG datasets, expert judgment, larger news corpora, event-level deduplication, additional disaster types, and richer contextual variables.

4. CONCLUSION

This study developed a context-aware disaster cause mining framework to transform Indonesian online disaster news into structured reported-cause and contextual transaction data for Association Rule Mining. The framework integrates text-based cause

extraction, contextual enrichment using population density and meteorological variables, GA-optimized Apriori, and merged rule pattern interpretation. From 742 final transaction records, the GA-optimized Apriori process retained 97 target rules, while the forest and land fire case study produced 24 valid association rules and 5 merged rule patterns. The strongest merged pattern was related to land_burning, with 14 rule variants, a support count of 87, and a merged support of 0.1173, indicating that land burning was the most dominant reported cause pattern in the analyzed news dataset. These findings show that the proposed framework can help disaster analysts and policymakers organize unstructured disaster narratives into interpretable reported cause-context association patterns for disaster risk analytics and decision-support preparation. However, the results should be understood as exploratory associations based on reported news narratives, not as verified causal evidence. Future work should validate the framework using official disaster records, expert assessment, larger news datasets, additional disaster types, event-level deduplication, and richer contextual variables.

REFERENCES

[1] M. R. Lessy, J. Lassa, and K. K. Zander, "Understanding Multi-Hazard Interactions and Impacts on Small-Island Communities: Insights from the Active Volcano Island of Ternate, Indonesia," *Sustainability (Switzerland)*, vol. 16, no. 16, Aug. 2024, doi: 10.3390/su16166894.

[2] D. Muthuvel and X. Qin, "Spatial concurrence risk of extreme precipitations in Southeast Asia under climate change using temporally dynamic complex networks," *J. Hydrol. (Amst)*., vol. 664, p. 134526, Jan. 2026, doi: 10.1016/j.jhydrol.2025.134526.

[3] G. F. Shidik *et al*., "Indonesian disaster named entity recognition from multi source information using bidirectional LSTM (BiLSTM)," *Journal of Open Innovation: Technology, Market, and Complexity*, vol. 10, no. 3, Sep. 2024, doi: 10.1016/j.joitmc.2024.100358.

[4] N. Istiqomah and F. Novika, "Extracting Post-Disaster Health Impact Information from News Reports Using Named Entity Recognition," *Journal of Informatics Information System Software Engineering and Applications (INISTA)*, vol. 7, no. 2, pp. 163–172, May 2025, doi: 10.20895/inista.v7i2.1814.

[5] Z. Wang, G. Cheng, Y. Wu, Y. Nie, L. Yang, and J. Xia, "Journal of Intelligent Decision Making and Granular Computing Research on Mine Disaster Risk Monitoring Based

on Topic Model Retrieval Technology," *Journal of Intelligent Decision Making and Granular Computing*, vol. 2, no. 1, pp. 1–12, 2026, doi: 10.31181/jidmgc11202629.

[6] K. Zheng, W. Qin, and X. Du, "Global Land Surface Dry/Wet Conditions Mining Based on Spatial-Temporal Association Rules," *Earth and Space Science*, vol. 8, no. 9, Sep. 2021, doi: 10.1029/2020EA001501.

[7] W. S. Karurung, K. Lee, and W. Lee, "Assessment of forest fire vulnerability prediction in Indonesia: Seasonal variability analysis using machine learning techniques," *International Journal of Applied Earth Observation and Geoinformation*, vol. 138, Apr. 2025, doi: 10.1016/j.jag.2025.104435.

[8] R. Haldulakar and J. Agrawal, "Optimization of Association Rule Mining through Genetic Algorithm," *International Journal on Computer Science and Engineering*, vol. 3, no. 3, pp. 1252–1259, Mar. 2011.

[9] C. Pinheiro, S. Guerreiro, and H. S. Mamede, "A Survey on Association Rule Mining for Enterprise Architecture Model Discovery," *Business and Information Systems Engineering*, vol. 66, no. 6, pp. 777–798, Dec. 2024, doi: 10.1007/s12599-023-00844-5.

[10] A. Telikani, A. H. Gandomi, and A. Shahbahrami, "A survey of evolutionary computation for association rule mining," *Inf. Sci. (N. Y).*, vol. 524, pp. 318–352, Jul. 2020, doi: 10.1016/j.ins.2020.02.073.

[11] H. R. Qodmanan, M. Nasiri, and B. Minaei-Bidgoli, "Multi objective association rule mining with genetic algorithm without specifying minimum support and minimum confidence," *Expert Syst. Appl.*, vol. 38, no. 1, pp. 288–298, Jan. 2011, doi: 10.1016/j.eswa.2010.06.060.

[12] H. Hou, L. Shen, J. Jia, and Z. Xu, "An integrated framework for flood disaster information extraction and analysis leveraging social media data: A case study of the Shouguang flood in China," *Science of the Total Environment*, vol. 949, Nov. 2024, doi: 10.1016/j.scitotenv.2024.174948.

[13] J. Gao, H. Yu, and S. Zhang, "Joint event causality extraction using dual-channel enhanced neural network," *Knowl. Based. Syst.*, vol. 258, Dec. 2022, doi: 10.1016/j.knosys.2022.109935.

[14] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using Siamese BERT-networks," in *Proc. EMNLP-IJCNLP*, 2019, pp. 3982–3992, doi: 10.18653/v1/D19-1410.

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

[15] M. Di Napoli *et al.*, "On the estimation of landslide intensity, hazard and density via data-driven models," *Natural Hazards*, vol. 119, no. 3, pp. 1513–1530, Dec. 2023, doi: 10.1007/s11069-023-06153-0.

[16] Badan Pusat Statistik, "Penduduk, Laju Pertumbuhan Penduduk, Distribusi Persentase Penduduk, Kepadatan Penduduk, Rasio Jenis Kelamin Penduduk Menurut Provinsi, 2024.

[17] M. Choo and D. K. Yoon, "A meta-analysis of the relationship between disaster vulnerability and disaster damage," *International Journal of Disaster Risk Reduction*, vol. 102, Feb. 2024, doi: 10.1016/j.ijdrr.2024.104302.

[18] K. E. Lamb and S. R. White, "Categorisation of built environment characteristics: The trouble with tertiles," *International Journal of Behavioral Nutrition and Physical Activity*, vol. 12, no. 1, Feb. 2015, doi: 10.1186/s12966-015-0181-9.

[19] P. Zippenfenig, "Open-Meteo.com Weather API," *Zenodo*, 2024, doi: 10.5281/zenodo.14582479.

[20] Open-Meteo, "Historical Weather API," *Open-Meteo Documentation*. Accessed: May 25, 2026.

[21] X. Ding, H. Wan, G. Shi, C. Hong, and Z. Liu, "Predicting hazard degree levels of metro operation accidents based on ordered constraint Apriori-RF method," *International Journal of Transportation Science and Technology*, vol. 18, pp. 245–260, Jun. 2025, doi: 10.1016/j.ijtst.2024.06.008.

[22] G. B. Gebremeskel and T. W. Yilma, "Multilevel rules mining association for processing big data using genetic algorithm," *Computing and Artificial Intelligence*, p. 1819, Feb. 2025, doi: 10.59400/cai1819.

[23] D. Gangaramani and R. Londhe, "Survey on association rule analysis: Exploration using mining analysis," *Int. J. Hybrid Intell. Syst.*, vol. 21, no. 1, pp. 2–13, Feb. 2025, doi: 10.3233/his-240015.