IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 15 July 2025, accepted 19 August 2025, date of publication 29 August 2025, date of current version 16 September 2025.

Digital Object Identifier 10.1109/ACCESS.2025.3604068

RESEARCH ARTICLE

Transformer-Based Semantic Role Labeling for Crisis Events Using Semi-Supervised Learning on Low-Resource Language Twitter Texts

AMELIA DEVI PUTRI ARIYANTO[1,2], DIANA PURWITASARI[1], (Senior Member, IEEE), CHASTINE FATICHAH[1], (Member, IEEE), SRI DEVI RAVANA[3], ANDRIAN[1], AND ANAK AGUNG YATESTHA PARWATA[1]

[1]Department of Informatics, Faculty of Intelligent Electrical and Informatics Technology, Institut Teknologi Sepuluh Nopember, Surabaya 60111, Indonesia
[2]Information Technology and Systems Study Program, Widya Husada Semarang University, Semarang 50146, Indonesia
[3]Department of Information Systems, Faculty of Computer Science and Information Technology, Universiti Malaya, Kuala Lumpur 50603, Malaysia

Corresponding author: Diana Purwitasari (diana@if.its.ac.id)

This research is funded by the Directorate General of Higher Education, Ministry of Education and Culture of the Republic of Indonesia, for The Doctoral Dissertation Program Scheme, under Grant 038/E5/PG.02.00.PL/2024 and the Indonesian Endowment Fund for Education (LPDP) on behalf of the Indonesian Ministry of Higher Education, Science and Technology and managed under the EQUITY Program (Contract No. 4299/B3/DT.03.08/2025 and No. 3029/PKS/ITS/2025).

ABSTRACT Twitter texts related to crisis events often contain vital information about the impact of disasters. However, irregular language structures, slang, and character limitations pose challenges for automated information extraction related to disaster impact. The Semantic Role Labeling (SRL) task can assist in extracting this information by identifying semantic roles in a sentence, such as who was involved, what happened, when, and where. This task provides a more structured understanding of the information related to disaster management. However, traditional based SRL tends to use predicates as the main anchor, which is less flexible in capturing semantic role relationships in unstructured texts such as Twitter. In addition, SRL research in low-resource languages such as Indonesian is still limited due to the lack of labeled datasets. This study aims to develop a more flexible SRL without relying on predicates as the main anchor and applying a semi-supervised learning strategy with pseudo-labeling. This strategy lets the model gradually learn from unlabeled data, enhancing generalization without solely depending on labeled data. A filtering function is implemented to minimize noise from inaccurate pseudo-labels, eliminating low-confidence predictions and ensuring that only high-quality data is used for retraining. Transformer-based models were chosen because of their self-attention mechanisms that enable an adequate understanding of contextual semantic relationships. Experiments demonstrate that the transformer model with Bidirectional Encoder Representations from Transformers (BERT) based architecture adapted for a specific language achieves the highest performance, with an F1-score of 0.863 at a threshold of 0.9. This study contributes to the advancement of SRL for low-resource language (especially for Indonesian) and paves the way for using SRL results in disaster severity analysis to enhance social media-based emergency response.

INDEX TERMS Crisis event, low-resource language, semantic role labeling, semi-supervised learning, transformer model.

I. INTRODUCTION

A crisis event refers to a sudden or gradual incident posing significant risks to human safety, health, welfare, the environment, and assets, requiring immediate and effective responses to minimize damage [1]. Crisis events can occur due to human actions or from nature. For instance, a massive earthquake that arises from nature can result in building collapse and human fatalities. This situation critically endangers people's health and safety, necessitating immediate evacuation, medical treatment, and infrastructure repair. Remote sensing data and social media data have been employed in prior research

The associate editor coordinating the review of this manuscript and approving it for publication was Prakasam Periasamy.

to identify crisis events [1]. Remote sensing data can be used to monitor and analyze crisis events without the need for direct physical contact [2]. Extreme weather patterns, like heavy rainfall, are monitored using weather satellite data to provide timely flooding warnings [3]. Radar data from satellites can also be used to detect ground deformation due to earthquakes, thereby helping scientists understand the impact of earthquakes and prepare for further mitigation [4]. However, the use of remote sensing data, such as high-resolution satellite data, is expensive, and its accessibility is limited to certain organizations, inhibiting its widespread use in various institutions or organizations that require such data for rapid response [1].

Social media platforms, such as Twitter (now called X) and Facebook, have opened up huge opportunities for fast and efficient information gathering during crisis events [5]. Social media data can provide early warnings and damage reports that are critical to effective disaster response. Twitter has become increasingly popular because of its fast response time, which allows information to be disseminated in real time [6]. For example, when an earthquake occurs, affected residents can immediately post information, which is then responded to via retweets or comments. Authorities or emergency response organizations monitoring Twitter can obtain early information and begin collecting data to assess the situation. However, Twitter faces challenges in data processing due to its character limit of only 280 [7], which can lead to the loss of important context. In addition, the language used in social media texts is often informal and includes slang or abbreviations [8], unlike more structured news texts [9]. More advanced text processing methods are required to extract critical information related to crises from Twitter to facilitate enhanced emergency response.

Previous studies have used various approaches to extract important information related to crisis events, including topic modeling approaches to understand topics discussed during crisis events. For instance, topic modeling revealed that the main discussions during the South Carolina flood disaster centered around non-human damage, such as power outages and human casualties [10]. Another approach is clustering, which is used to group messages, news, or reports from sources such as social media (e.g., Twitter) that indicate a crisis event [11]. Clustering and topic modeling only group information based on common characteristics or broad topics without revealing specific details about actions, objects, or victims. In contrast, sentiment analysis only determines whether a message is positive, negative, or neutral [12] without providing rich details about specific text elements related to the crisis event.

Previous studies have also employed classification approaches to identify relevant messages from crisis events, addressing the challenge of noisy Twitter data [5]. For example, Blomeier's study [13] classified relevant tweets about the Ahr Valley flood in Germany to facilitate effective disaster response using various classifier algorithms, including random forest, naïve Bayes, and support vector machine. Developing a machine learning model to identify relevant messages (with binary labels such as relevant or irrelevant) is an important first step; however, it is not sufficient to fully and quickly understand a crisis situation [5]. There is a need to develop a more complete system to enhance situational awareness in crisis events, which is crucial for emergency response organizations' decision-making [14]. Situational awareness involves understanding a high-level picture of a disaster, including the overall scale of damage, casualties, injured individuals, and their urgent needs [5]. Unlike clustering, topic modeling, sentiment analysis, and classification—which focus on grouping, summarizing, or filtering relevant information—information extraction transforms unstructured text into structured data by identifying key details such as entities, semantic roles, and relationships. This structured representation facilitates a deeper understanding of crisis events, improving situational awareness [8].

In a crisis context, an SRL task for information extraction offers a promising solution for analyzing Twitter text data. By identifying semantic roles within unstructured text, SRL enables a more complete and structured representation of crisis events, enhancing situational awareness for effective emergency response [8]. Semantic roles are the roles played by words or phrases in a sentence that are usually associated with the main predicate of an event [15]. These semantic roles include information related to who performed the action, who received the action, where the action occurred, and when the action occurred [9]. For example, there is a tweet that shows a crisis event caused by human actions such as ''a train hits a house in northern Greece''. The semantic role of ''a train'' in the SRL approach is identified as Agent (the perpetrator of an action), which explains that the party that may be responsible for the collision is the railway company. The word ''a house'' is identified as having a semantic role as Patient (the recipient of the action taken by the perpetrator), so that information related to who and what was affected by the collision and damage evaluation can be carried out. The word ''northern Greece'' has a semantic role as the location where the incident occurred; thus, emergency response organizations can capture the specific location of a crisis event when planning rescue operations.

Recent studies on SRL have widely adopted Proposition Bank (PropBank) labeling rules, as shown in Table 1, where the labeling assigns specific semantic roles (ARG0–ARG5) to sentence elements depending on the verb and context [16]. The most commonly used labels are ARG0, which typically refers to the doer of an action (Agent), and ARG1, which represents the recipient of the action (Patient) [17]. The meanings of ARG2–ARG5 can vary significantly depending on the predicate, often causing ambiguity and confusion during label assignment [18]. Additionally, these labels lack sensitivity to domain-specific details, such as the number of victims or impacted objects, which are crucial

for understanding crisis events and supporting emergency response organizations. Due to its predicate-centric nature, traditional PropBank-based SRL assigns semantic roles to each predicate individually, which can result in irrelevant semantic role assignments in the context of crisis events. For instance, in complex sentences with multiple predicates, predicate-centric SRL can produce many semantic role label results based on all its predicates, making it difficult to interpret because it ignores the overall context of the sentence and focuses too much on each predicate [19].

In contrast to traditional predicate-centered SRL studies, we are motivated to develop SRL for information extraction on informal Twitter text data in the crisis event domain and do not make predicates as their main anchors (illustrated in Figure 1). Thus, the analysis of semantic role labels becomes more flexible and context-rich, as it provides a deeper understanding of how semantic roles are connected within events or actions described in the text as a whole rather than being constrained by direct associations with specific predicates. For example, in Figure 1, the phrase ''the truck'' plays the semantic role AFFECTEDOBJECTS-ARG because it is directly impacted in the crisis event, being involved in the accident and affecting other objects, such as motorbikes. Similarly, ''two motorbikes'' also take on the AFFECTEDOBJECTS-ARG semantic role, as they are directly affected by the accident caused by the truck, resulting in damage or further involvement in the incident. Assigning semantic role labels helps identify key aspects of a crisis event, enabling emergency response teams to prioritize rescue efforts effectively.

Several studies have applied SRL in specific domains, such as Access Control Policies, scientific literature on biology, or news articles (refer to Table 1). There are not many previous studies covering the crisis event domain. In addition, previous studies have successfully employed SRL methods in high-resource languages like English and Chinese due to the public datasets available for these languages. Data annotation is the first stage in building a model for SRL, particularly in texts containing low-resource languages such as Amharic [15], Sinhala [20], and Indonesian [9]. Manual data labeling requires human involvement, which is time-consuming and laborious. The process is very expensive and requires significant effort to produce large, high-quality datasets [21]. Thus, research related to the development of SRL in the crisis event domain using low-resource languages still have great potential. Semi-supervised learning offers an efficient solution for improving the performance of SRL models by using unlabeled data and can reduce the dependence on expensive and time-consuming manual labeling [22]. Therefore, due to the limited availability of labeled semantic role data, we plan to utilize semi-supervised learning for an SRL task related to crisis events in Indonesia. The semi-supervised learning strategy generates pseudo-labels from unlabeled data, which are then combined with the initial labeled dataset for retraining. However, semi-supervised learning can introduce noise into the dataset because pseudo-labels may contain errors. A filtering function is used to discard those below a certain threshold before retraining to maintain model performance.

Although semi-supervised learning strategies can improve the efficiency of SRL tasks, selecting the proper method for the SRL model is also an important factor. Based on the analysis of previous studies in Table 1, the commonly used methods in SRL tasks are still dominated by rule-based approaches. For example, research on narrative abstractive summarization in news documents utilizes rule-based SRL to enrich the summary timeline with semantic role information [23]. Similarly, research focusing on extracting events and relationships in genetic regulatory networks from scientific literature on molecular biology and biomedical sciences also uses rule-based methods [16]. However, rule-based approaches have significant limitations, especially in dealing with unstructured texts with high language variation, such as the text on Twitter used in our study. Rule-based models have difficulty handling this variation because they rely heavily on pre-defined rules, making them less flexible in dealing with new or unexpected contexts [9]. Our study utilizes a Transformer-based SRL method as an effective approach. This method leverages bidirectional processing and the self-attention mechanism, which assigns weights to each word based on its relevance to other words in the text, enabling a broader contextual understanding beyond sequential word order.

Some of our primary contributions are summarized as follows:

1) This study develops a more contextual SRL approach for the crisis event domain. Unlike previous studies that mostly use PropBank-based SRL annotations with ARG0-ARG5 labels that are less specific and irrelevant in capturing important nuances of crisis events, this study designs an SRL approach that does not make predicates the main anchor. This method facilitates a better comprehension of semantic roles and aligns with the context of crisis events.
2) This study contributes to filling the gap in developing SRL for low-resource languages. Most previous SRL studies have focused on high-resource languages such as English and Mandarin, while this study contributes to the development of SRL for Indonesian, which still has minimal public datasets. Thus, this study opens up wider opportunities for NLP in low-resource languages.
3) This study utilizes a semi-supervised learning approach to overcome the limitations of labeled data in the crisis event domain. Using unlabeled data enhances the SRL model, decreasing reliance on costly and time-consuming manual labeling.
4) This study evaluates and compares the performance of several transformer models for SRL tasks on Indonesian Twitter texts discussing crisis events. Twitter texts have unique characteristics, such as length limitations, use of slang, abbreviations, and language structures that are not always formal, which can affect the model's

effectiveness in understanding semantic roles. Through comparative analysis, this study identifies the most appropriate transformer-based SRL model for handling short and unstructured texts in the context of crisis events. The study's results aim to offer insights for developing more effective SRL methods and advancing NLP for the Indonesian language.

This study consists of five sections, including Section I, which introduces the background, motivation, and solution. Section II discusses related research on crisis events in social media and the development of SRL for information extraction. Section III describes the proposed method for developing SRL in the crisis event domain in low-resource languages. Section IV presents the experimental results and analysis. Section V presents the conclusions and future work directions.

II. RELATED WORKS

This section discusses previous studies related to crisis event detection using social media data and SRL information extraction methods.

A. CRISIS EVENT ON SOCIAL MEDIA DATA

Several approaches have been proposed for crisis event detection from social media data, such as topic modeling, clustering, sentiment analysis, classification, and information extraction. The topic modeling approach is used to understand the topics discussed in a crisis event, such as the case of a typhoon in Japan [24] and flooding in South Carolina [10], to add insights into how the public understands, feels, and responds to such crisis events through social media. Kitazawa and Hale [24] applied the Latent Dirichlet Allocation (LDA) algorithm in a topic modeling approach to crisis event detection during a typhoon in Japan. The reason for choosing LDA was that it is included in unsupervised learning, which does not require strong prior knowledge related to specific topics that appear in public conversations. After the topics were successfully identified by LDA, they were assigned to four predetermined categories: awareness, preparation, action, and impact. The awareness category relates to hazardous weather phenomena, such as hurricane intensity and weather forecast topics. The preparation category presents public disaster preparation activities, such as general and landslide warnings. The action category consists of evacuation actions. The impact category focuses on the impact of a disaster and does not refer to specific actions, such as train disruptions. Karami et al. [10] also used the LDA algorithm to detect crisis events during flooding in South Carolina. The topics that are often discussed in South Carolina's crisis-related research on flooding are non-human damage (such as loss of electricity) and human casualties, which are of primary public concern.

Another approach that can be used for crisis event detection is clustering, which uses unsupervised learning techniques to group data into several clusters based on the similarity or proximity of features without requiring labeled data [25]. Belcastro et al. [26] used a clustering approach to group social media (Twitter) user posts based on geographical similarity during the earthquake in Italy. The Density-Based Spatial Clustering of Applications with Noise (DBSCAN) algorithm

TABLE 1. Summary of differences in contributions between previous SRL studies and our research.

Ref.	Previous Studies Contributions	Text Data Source	Text Formality	Languages	Domain	Labeling Rules	Methods Used
[19]	This study extracted access control policies from natural language documents.	Access control policies documents	Formal	English	Access control policies	PropBank's annotation	SwiRL Tools
[16]	This study performs event and relation extraction for genetic regulatory networks using SRL.	Scientific literature in molecular biology and biomedical journals.	Formal	English	Biology	PropBank's annotation	Rule-based
[23]	This study develops a narrative abstractive summarization method that constructs chronological summaries from news documents, using SRL to enrich the timeline with semantic role information before abstraction and sentence selection.	Task 4 SemEval 2015 dataset, which comes from the Wikinews article	Formal	English	Several specific topics	PropBank's annotation	Rule-based
[27]	This study aims to improve SRL performance in Chinese, which has unique characteristics.	Public dataset: Chinese Proposition Bank (taken from news and magazine articles)	Formal	Chinese	Several specific topics	PropBank's annotation	LSTM model
[28]	This study aims to develop semantic role labeling rules for Bei sentences in Mandarin to improve accuracy.	News corpus from 2014 People's Daily	Formal	Chinese	Several specific topics	PropBank's annotation	Rule-based
Our study		**Twitter**	**Informal**	**Indonesian**	**Crisis event**	**Our proposed annotation**	**Transformer model**

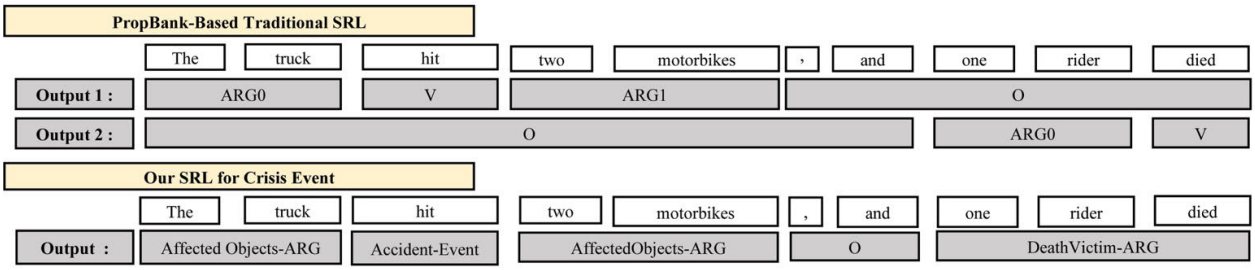


FIGURE 1. Differences between PropBank-based traditional SRL annotation and our SRL annotation for crisis event.

was used for crisis event detection in the case of an earthquake in Italy due to its reliability in detecting clusters of various sizes and shapes and its tolerance to noise. DBSCAN can help filter noise from highly variable social media data because it only focuses on spatially significant data clusters. After the data is grouped into clusters based on geographical similarity, the next step is to understand the contents of each cluster by extracting keywords from each cluster. Ruz et al. [29] also used a clustering approach to detect crisis events in the case of Hurricane Florence. The spatiotemporal DBSAN algorithm is used to group post messages based on time and location. Then, topic modeling is used to identify topics discussed in a single spatiotemporal cluster.

The sentiment analysis approach is also included in the crisis event detection approach, which focuses on identifying whether the sentiment expressed in a tweet is positive, negative, or neutral so that it can reflect the feelings or reactions of social media users toward a crisis event [29]. Neppalli et al. [30] used sentiment from text data on Twitter based on the geographical location of its users and measured how sentiment could change based on the relative distance from the epicenter of Hurricane Sandy in the Atlantic Ocean. They assumed that tweets from users closer to the center of the storm showed a more negative sentiment than tweets from users farther from the epicenter. The algorithm used by Neppalli et al. to detect tweet sentiment is a machine learning algorithm, such as Support Vector Machine (SVM) and Naïve Bayes, with input feature representation a combination of bag-of-words and polarity scores. Wu and Cui [31] also conducted a sentiment analysis of the crisis event due to Hurricane Sandy in the Atlantic Ocean using the SentiWordNet 3.0 algorithm, which is a lexical resource for calculating sentiment scores. Each word in the tweet was matched with its corresponding entry in SentiWordNet. If a word was found in SentiWordNet, then positive and negative scores were taken. The scores for each word were then used to calculate the overall sentiment score of the tweet.

Classification approaches in the crisis event domain are widely used to classify tweets as relevant or irrelevant or to classify requests for help, which require labeled data in the classification process. Zhou et al. [32] used the Bidirectional Encoder Representations from Transformers (BERT) algorithm to classify tweet text messages in the case of Hurricane Harvey. One tweet text can be classified into several categories: requests for help, tweets that provide complete addresses, and tweets that mention demographic information from victims. Bhoi et al. [33] proposed a hybrid Long Short-Term Memory (LSTM) algorithm with Convolutional Neural Network (CNN) to classify tweets as relevant or irrelevant related to crisis events about the earthquake in Nepal. Upadhyay et al. [34] adopted a classification approach to identify whether or not there was a crisis event from a tweet text. The dataset used by Upadhyay et al. includes several public datasets related to crisis events that occur in English-speaking countries (Canada, Australia and the United States), and propose a hybrid algorithm (BERT, CNN, Bidirectional LSTM, and Self-Attention) to capture semantic relationships from a crisis text while preserving important information.

Previous studies have used topic modeling, clustering, sentiment analysis, and classification approaches. However, few information extraction approaches have been developed for crisis events. The clustering approach only groups texts based on feature similarities without capturing the deeper meaning of the relationships between key components in the text. As a result, it fails to identify crucial details such as who was involved, what happened, where, when, and why the crisis occurred. The topic modeling approach is less effective for short texts posted on social media, where topics cannot be clearly identified due to limited content [35]. The sentiment analysis approach is limited to positive, negative, or neutral emotional content of texts posted on social media without providing in-depth information about the crisis event. Existing classification approaches that only classify tweets as relevant, irrelevant, or related to specific help requests do not adequately capture the complete context of a crisis event.

The information extraction approach can extract specific information from the text that can be used for further action, such as identifying the location of the incident or determining the number of victims. We found one study that used an information-extraction approach, namely, the study conducted by Huang et al. [36], who used an information extraction approach for crisis event detection in the case of an earthquake in China. However, the extracted information only included location information, which was obtained using the FoolNLTK tool, and time information, which was extracted using regular expressions. In contrast to previous studies,

we use an information extraction approach (through SRL task) for crisis event detection that can identify semantic roles between words in sentences so that the context of text on social media can be done comprehensively (such as who was involved, what happened, how, when, and where) from a crisis event.

B. SEMANTIC ROLE LABELING FOR INFORMATION EXTRACTION

SRL is one of the main techniques in Natural Language Processing (NLP) designed to identify the semantic role of each element in a sentence. Through SRL, each component of a sentence is given a specific role to provide a complete picture of who did what, to whom, when, and where [37]. SRL is effective for information extraction, especially from unstructured texts such as news, reports, and social media posts. For example, in the case of a crisis event such as a flood from Twitter text, SRL can help identify who was affected, where the flood occurred, and what caused the flood. Several methods are used in SRL for information extraction such as traditional methods, machine learning, and deep learning methods [8]. The use of tools was also discussed in previous SRL research.

Tools are practical applications of various NLP methods that combine rule-based methods, machine learning, and/or deep learning to facilitate application on an industrial or research scale [38]. Narouei et al. [19] used SRL tools to identify Access Control Policies (ACP) from English natural language documents. The tools used in this study were SwiRL tools, which provide ready-to-use SRLs that allow users to run a few lines of code without developing a model from scratch. The labeling performed by Narouei et al. refers to the PropBank annotation, where the information extracted in the ACP domain includes the verb, ARG0 (actor who gives the task), ARG1 (thing assigned by the actor), ARG2 (object that is given the task), and ARG-MOD (modal). Alzahrani and Aljuaid [39] used SRL tools developed by the University of Illinois to extract verbs and their associated semantic roles as features for measuring cross-language text similarity between Arabic and English. There are weaknesses in using tools when performing SRL for information extraction, such as the lack of flexibility for certain domains. Flexibility is reduced because the available tools tend to be trained on very general datasets, such as news or non-specific texts; thus, if applied to more specific domains, the performance of the tools will decrease [38].

The rule-based method (without tools) was carried out by Wang [28] to improve the accuracy of automatic SRL on passive sentences in Chinese, especially sentences using the passive marker ''Bei'' because of its complicated structure. This study extracted information pertaining to the semantic role in passive sentences, including determining who is the agent (doer of the action), who is the patient (recipient of the action), and the semantic relationship between nouns and verbs in sentences using the passive marker ''Bei''. Another study by Hou and Ceesay [16] also used the rule-based method to extract events and relations in genetic regulatory networks from English-language scientific literature on molecular biology and biomedicine by utilizing SRL. The genetic regulatory network's semantic roles are identified through information extraction, including core labels (ARG0-ARG5) and modifier tags, such as ARG-LOC (indicating location) and ARG-TMP (indicating time). The advantage of the rule-based method for SRL is that it does not require a large training dataset. However, rule-based methods are based on linguistic rules that are specific to a particular domain or language styles. When these rules are applied to different domains, there is a concern that performance will decrease because the rules may not be sufficiently general enough to handle a wider range of language variations [40].

Unlike the rule-based method, which tends to be rigid and only works well for texts that conform to previously defined rules and are difficult to generalize to different texts, the SRL machine learning method can create a model to learn directly from training data [8]. The more data provided, the better the model was at recognizing patterns in the text. With machine learning methods, generalization capabilities can be increased to make better predictions of texts that have not previously been found before [9]. Lazemi et al. [41] developed a machine learning-based SRL model that can be applied to Persian. The MaxEnt classifier algorithm considers several features, such as word position features with verbs (before or after the verb), voice features of the verb, and distance features to the verb. This study extracted information related to semantic roles in the Persian language domain, including semantic roles such as the doer of the action (ARG0), the recipient of the action (ARG1), the location where an action occurs (ARG-LOC), and the purpose or direction of the action (ARG-DIR). Careful feature selection and design for manual text data extraction is a time-consuming and error-prone process in machine learning methods.

In recent years, deep learning methods have become popular as information extraction approaches in NLP because it can automatically extract complex patterns or features from text data without the need for explicit human definitions [40]. One of the widely used deep learning variants is Long Short-Term Memory (LSTM), which is a type of Recurrent Neural Network (RNN) designed to handle long-term dependencies in sequential data [8]. Xia et al. [27] introduced a deep bidirectional highway LSTM and combined it with Conditional Random Fields (CRFs) and POS tagging to improve SRL performance in complex Chinese. CRFs are used in the upper layer to improve predictions in sequence-to-sequence tasks, especially in handling how one semantic role label relates to the following label. POS tagging provides information about the syntactic class of a word in a sentence, thus helping the model to determine the semantic role of each word in a sentence, such as whether a word functions as an agent or patient in an action (verb). POS tagging is also used for embedding. In many modern deep learning-based models, POS tags can be transformed into vector representations (embeddings) and

combined with word embeddings to provide additional context about the syntactic structure of a sentence.

In addition to LSTM, transformer-based deep learning methods are increasingly used in information extraction approach [8]. One well-known transformer models is BERT (Bidirectional Encoder Representations from Transformers) which has been widely used, including in conversational SRL tasks for Chinese, to extract semantic role information from conversations by considering the sequential context of the entire dialog [17]. Other variants of BERT, such as RoBERTa, have also been used specifically in the Robotic-surgical domain to extract surgical procedural actions from the English literature [42]. Transformer-based deep learning methods have the advantage of handling the global context in text because they are able to process all tokens simultaneously so that the context between words throughout the sentence can be better integrated [8].

We are motivated to explore information extraction using the SRL task in crisis events for low-resource languages like Indonesian, particularly in Twitter texts, as it offers a new direction for advancing NLP in this domain. To achieve this, we use a Transformer-based SRL method, which leverages bidirectional context understanding and a self-attention mechanism that assigns weights to words based on relevance. Additionally, Transformers can adapt to language variations without relying on linguistic rules or manual feature extraction, making them well-suited for low-resource languages like Indonesian.

III. METHODOLOGY

Figure 2 illustrates the overall flowchart of our proposed work for performing SRL on Indonesian Twitter text data related to crisis events. Several main phases are carried out: dataset development, self-training with filtering function, and SRL model evaluation. Detailed explanations of each phase are described in the following subsections.

A. PHASE I: DATASET DEVELOPMENT

1) DATA PREPARATION

The first stage in the dataset development phase is data preparation by crawling text data (tweets) using Twitter access tokens, specifying a one-month time span, and implementing a time lag to comply with Twitter API limitations during the tweet collection. The search keywords included ''*kebakaran* (fire)'', ''*terbakar* (burning)'', ''*kecelakaan* (accident)'', ''*banjir* (flood)'', and ''*gempa bumi* (earthquake)'', with a search period from 2018 to 2023. Several keywords were chosen because they represent four disaster events that often occur in Indonesia and have quite a significant impact, such as floods, earthquakes, fires, and accidents. Floods and earthquakes are natural disasters that often occur in Indonesia due to its geographical conditions in the Pacific Ring of Fire, causing high seismic activity (earthquakes) [43] and high rainfall that often triggers flooding. Fires and accidents are closely related to non-natural factors, where forest fires and residential fires often occur in Indonesia, especially during the dry season or due to uncontrolled human activities [44]. Meanwhile, traffic and transportation accidents are also one of the leading causes of death in Indonesia due to poor driving behavior [45]. Thus, these four events can represent natural disasters (earthquakes, floods) and non-natural disasters (fires, accidents), providing broad coverage in disaster analysis in Indonesia using text data from social media.

Tweet crawling was conducted using several Python modules such as calendar (to calculate monthly date ranges), datetime (for time adjustment), and time (to delay execution and handle API rate limits). When access limitations occurred, the program paused for two minutes before retrying, while successful retrievals were followed by a 20-second delay before moving to the next date range. In total, 269,652 raw tweets were collected, consisting of 39,030 related to accidents, 43,028 to floods, 149,308 to fires, and 38,286 to earthquakes.

The raw tweets were preprocessed through case folding (converting all text to lowercase), removing URLs, mentions, and hashtags, followed by text normalization to replace slang with formal words. Lemmatization (changing words into basic forms [46]) and stopword removal were intentionally excluded to preserve sentence context, particularly for location names. For instance, the word ''*berlari* (running)'' would typically be reduced to its root ''*lari* (run)'', but in cases such as ''*Jl Seturan* (Seturan Street)'', lemmatization could incorrectly transform it into a non-place form, thus distorting the original meaning.

Duplicate tweets were removed during preprocessing by calculating cosine similarity between tweet pairs, with a threshold of 0.8 used to identify and delete duplicates. In addition, tweets referring to disaster events outside Indonesia were excluded. For instance, a tweet such as ''*Menelan korban 179 jiwa, banjir di daerah Sikkim India* (Took 179 lives, floods in the Sikkim area of India)'' was considered irrelevant to the Indonesian context and removed. This filtering was performed using string matching with lists of Indonesian place names and foreign country names. After preprocessing and filtering, 56,135 tweets remained: 8,155 accidents, 7,321 floods, 33,251 fires, and 7,408 earthquakes.

2) SPLITTING DATA

A total of 10,000 tweets were randomly selected from 56,135 clean tweets to build an SRL model with a semi-supervised learning approach. The proportion of training and testing data used is 90%-10% because 90% of training data allows the model to learn more examples, thus increasing its ability to capture patterns and relationships in the data [47]. Meanwhile, 10% of the data (1,000 tweets) is used as testing data to evaluate model performance, ensuring that the model can generalize well to new data that has never been seen before.

Furthermore, a further division is carried out from 90% of the training data into 35% labeled data (3,150 tweets) and 65% unlabeled data (5,850 tweets). This division is carried

out because the SRL process is done with a semi-supervised learning approach, where only a tiny portion of the data has manual annotations. In contrast, most data remains unlabeled and will be used in semi-supervised learning. The proportion of 35% labeled data is chosen so that the model has enough annotated examples to build an initial representation, while 65% unlabeled data is used to improve the model's generalization through semi-supervised learning techniques.

The labeled training data by event type includes 1,000 tweets for floods, 656 for earthquakes, 878 for fires, and 616 for accidents. This distribution reflects the natural pattern of events on Twitter, where some categories (such as floods or fires) have more reports than others. In the real world, the number of reported events in Indonesia is not always evenly distributed—for example, floods tend to be reported more often than earthquakes. By maintaining this distribution in the training data, the model is better prepared to deal with variations in the frequency of actual events.

To ensure that the labeled training data used in the semi-supervised learning process covers sufficient linguistic diversity, an analysis of the linguistic characteristics of the manually annotated tweets was conducted (Table 2). This analysis covers four major crisis event categories: floods, earthquakes, fires, and accidents. The fire category has the highest average tweet length of around 26.83 tokens, followed by earthquakes with 25.52 tokens, floods with 23.69 tokens, and accidents with the shortest length of around 19.99 tokens per tweet. The significant differences in minimum and maximum tweet lengths across categories (e.g., fires range from 5 to 126 tokens) indicate that the data contains very short to very long sentences, reflecting the diversity of syntactic structures in Twitter texts.

In addition, vocabulary diversity was also analyzed using the Type-Token Ratio (TTR) metric, which is the ratio between the number of unique words (types) and the total number of words (tokens) in the text. The TTR value is calculated by dividing the number of unique words by the number of words that appear. TTR values range from 0 to 1, where values closer to 1 indicate that almost all words used are unique (very high vocabulary diversity), while values closer to 0 indicate that most words are repeated, reflecting low vocabulary diversity) [48]. In this dataset, TTR values range from 0.215 to 0.266, indicating a good level of vocabulary variation within each category. For example, the accident

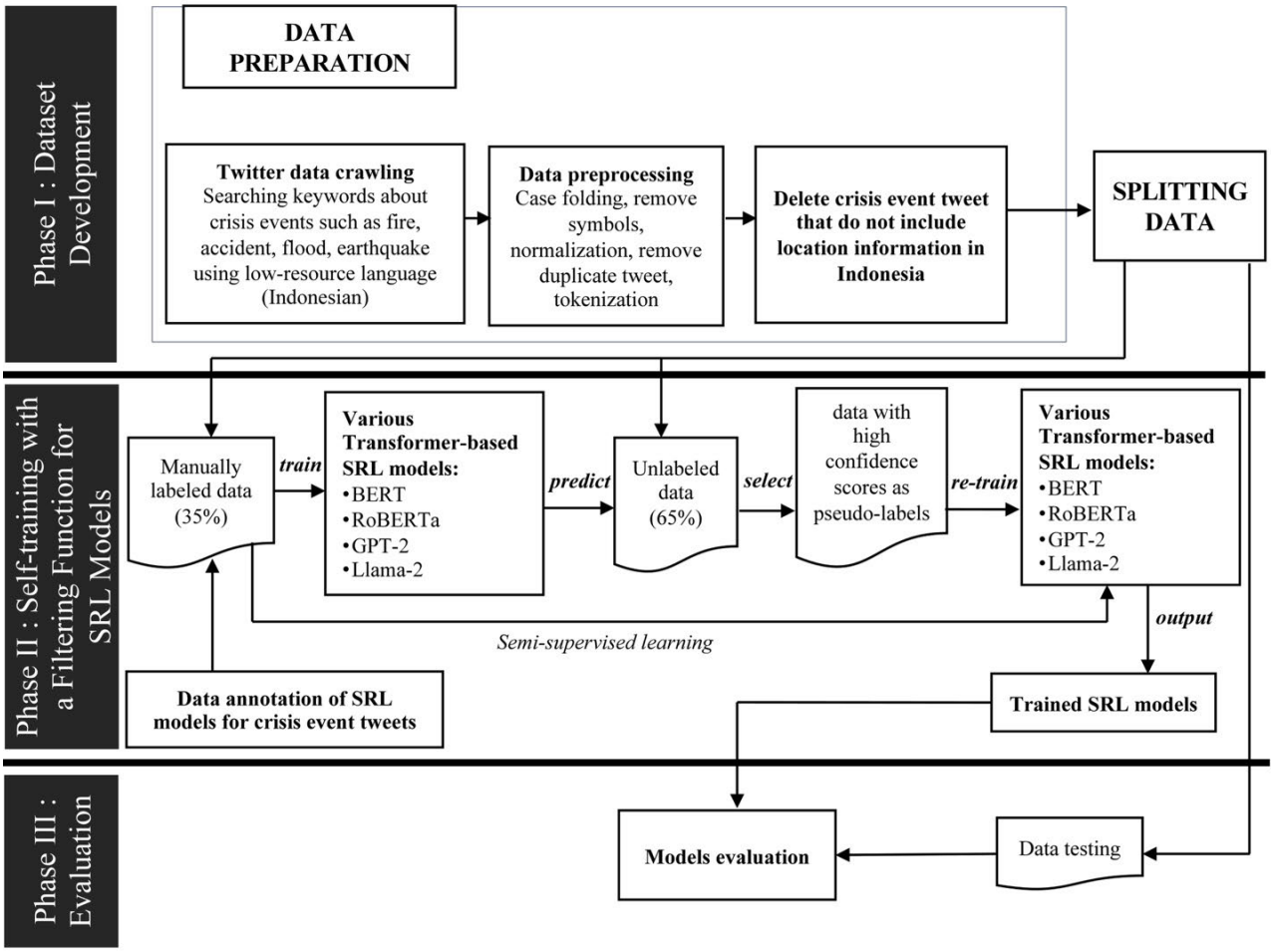


FIGURE 2. Overall flowchart of proposed work.

category has the highest TTR (0.266), indicating high word variation despite its shorter average tweets. In contrast, the fire category has the lowest TTR (0.215), but still shows adequate lexical richness. These values indicate that the annotated dataset covers substantial linguistic diversity in terms of sentence structure and vocabulary variation.

Thus, the labeled training data used in this study have been carefully selected and analyzed to realistically represent the distribution of crisis events and capture essential linguistic variation. This ensures that even though only about 3,150 tweets were manually annotated, this subset is still representative and sufficient to train the model to capture complex semantic patterns in Twitter text related to crisis events. Meanwhile, the testing data is randomly selected with a balanced distribution of 250 sentences for each event type (floods, earthquakes, fires, and accidents). This approach ensures that the model evaluation is carried out fairly across all types of events so that the model's performance not only depends on the more dominant categories in training but also remains accurate for less frequent events.

B. PHASE II: SELF-TRAINING WITH A FILTERING FUNCTION FOR SRL MODELS

1) DATA ANNOTATION

The SRL approach utilizing semi-supervised learning necessitates labeled training data. According to the data division from the previous stage, around 3,150 tweets in the training dataset require annotation, and also 1,000 tweets in the testing dataset need to be labeled as ground truth. Because SRL works at the token level, some tweets are tokenized before being given semantic role labels. Overall, the annotation process produces 76,298 labeled tokens in the training data and 22,908 tokens in the testing data.

To systematically label these tokens, this study defines fifteen semantic role labels designed to detect crisis events in Indonesian texts, which are classified as low-resource languages [49]. Table 4 defines various semantic role labels, including main event labels such as FALSE-EVENT, FLOOD-EVENT, FIRE-EVENT, EARTHQUAKE-EVENT, and ACCIDENT-EVENT. FALSE-EVENT is used to identify texts that convey information that does not describe an actual event. For example, in the text "*fenomena gerhana bulan total memang indah tapi warga harus waspada banjir rob punya tips atau berita warga lainnya?* (the phenomenon of total lunar eclipse is indeed beautiful, but residents must be aware of rob floods, have tips or other news from residents?)," the phrase "*banjir rob* (rob floods)" is labeled FALSE-EVENT because, in the context of the sentence, this phrase does not describe an event that is happening, but only a warning or potential event in the future. Meanwhile, FLOOD-EVENT, FIRE-EVENT, EARTHQUAKE-EVENT, and ACCIDENT-EVENT highlight the main events in the text according to the event type. For instance, FLOOD-EVENT is given to words that indicate the actual occurrence of a flood, such as in the sentence "*hujan deras sejak sabtu akibatkan banjir besar di jember 1 orang meninggal ratusan rumah terendam*(heavy rain since Saturday caused a major flood in Jember, one person died, hundreds of houses were submerged)," where there is factual information such as time, place, and impact of the flood. Besides the main event label, additional semantic roles offer more information regarding an event. PLACE-ARG is used to mark the event's location, TIME-ARG indicates when the event occurred, and the AFFECTEDOBJECTS-ARG label is used to identify the objects affected.

We conducted a comprehensive validation process to ensure that these designed semantic role labels are consistently and correctly applied. First, the overall consistency between annotators was assessed using the Inter-Annotator Agreement (IAA) measured by Cohen's Kappa, providing a global indication of agreement across all labeled tokens to ensure the reliability of the argument labels assigned by the two annotators. The annotation team consisted of two fluent Indonesian speakers: an expert in disaster management and a PhD student under the guidance of two computer science experts. The IAA results show a value of 0.924, close to 1, indicating that the semantic role labeling results are objective and demonstrate a very high level of agreement [50].

In addition, to gain more granular insights into the reliability of individual labels, we calculated per-label IAA using classification metrics, including precision, recall, and F1-score (see Table 3). This approach is crucial because Cohen's Kappa provides only an overall aggregate value, failing to pinpoint specific labels that exhibit consistency and those that tend to have discrepancies. Consequently, this classification metric allows a more focused analysis of low-frequency labels. Technically, all token labels from both annotators are arranged in two parallel vectors. The labels are mapped to a consistent numeric format and then calculated using the classification_report function from the sklearn library. This function builds a one-vs-rest evaluation scheme for each label and produces precision, recall, and F1-score metrics for each label. This process is symmetric; if the position of the annotator is reversed, the F1-score value will remain the same. For example, although WOUNDVICTIM-ARG is a low-frequency label, the evaluation results show that the annotators still have a relatively high level of agreement with an F1-score of 0.84, a precision of 0.88, and a recall of 0.81. This score indicates that both annotators are relatively consistent in recognizing and assigning the label, even though the number of occurrences is limited. Similarly, the DEATHVICTIM-ARG label, which is also classified as a low-frequency class, achieved a high F1-score of 0.91, reflecting the clarity of the label definition and strong understanding among annotators of the concept. These reliability assessments provide the foundation for understanding how the defined labels are applied in practice, which can be further illustrated with concrete annotation examples.

Figure 3 provides an example of token annotation in a tweet, illustrating how these labels are applied in text. During the annotation process, several terms are utilized, including

TABLE 2. Linguistic variation summary per category (from labeled training data).

Category	Number of Sentences	Average Tokens	Min	Max	Range (Max-Min)	TTR
Flood	1000	23.69	4	56	52	0.248
Earthquake	656	25.52	4	51	47	0.222
Fire	878	26.83	5	126	121	0.215
Accident	616	19.99	4	54	50	0.266

TABLE 3. IAA results per label.

Semantic Role Label	Precision	Recall	F1-score
O	0.97	0.99	0.98
FALSE-EVENT	0.88	0.85	0.87
PLACE-ARG	0.98	0.96	0.97
TIME-ARG	0.90	0.94	0.92
OFFICER-ARG	0.97	0.85	0.91
FLOOD-EVENT	0.63	0.93	0.75
STREET-ARG	0.91	0.92	0.91
INFORMATION-ARG	0.95	0.75	0.84
AFFECTEDOBJECTS-ARG	0.96	0.95	0.96
REASON-ARG	0.93	0.94	0.94
WOUNDVICTIM-ARG	0.88	0.81	0.84
DEATHVICTIM-ARG	0.87	0.94	0.91
EARTHQUAKE-EVENT	0.96	0.62	0.76
FIRE-EVENT	0.91	0.91	0.91
ACCIDENT-EVENT	0.94	0.96	0.95

text_id, which serves as a unique identifier for each tweet entry in the dataset. The text_id format consists of an event code followed by a text sequence number in the dataset, for example, BAN-01183. The event code in text_id is derived from Indonesian words that indicate the type of event: BAN for flood *(BANJIR),*GEM for earthquake *(GEMPA),* KEB for fire *(KEBAKARAN),* and KEC for accident *(KECELAKAAN).* Thus, the text_id BAN-01183 indicates that the tweet is on row 01183 for the flood event.

Token_id represents the identity of each token in a tweet, with a format that follows the sequence of tokens in the text (for example, BAN-01183.006 for the sixth token, and BAN- 01183.007 for the seventh token). Tokens are word units to be annotated, while labels indicate the semantic role assigned to each token based on its context in the text. For example, in the sentence "*banjir besar di jember 1 orang meninggal ratusan rumah terendam* (big flood in jember one person died hundreds of houses submerged)," the word "*banjir besar* (big flood)" is marked as FLOOD-EVENT, "*jember* (one of the city names in the East Java province in Indonesia is Jember City)" is labeled PLACE-ARG, while "*1 orang meninggal* (one person died)" is labeled DEATHVICTIM- ARG because it indicates the number of fatalities. Meanwhile, "*ratusan rumah terendam* (hundreds of houses submerged)" is labeled AFFECTEDOBJECTS-ARG, indicating the event's impact on the property. Thus, the SRL model built in this study can identify texts related to crisis events more contextually and informatively by leveraging the semantic roles in Indonesian text, compared to traditional SRL, which only uses the ARG0-ARG5 scheme. Overall, the combination of quantitative agreement measurements and practical annotation examples validates that the designed semantic role labels are both conceptually sound and reliably applicable in this dataset.

The frequency distribution of semantic role labels based on event types in labeled training and testing data is presented in Table 5. The table shows that the O label has the most significant proportion in both training (69.74%) and testing (67.85%) data, indicating that most of the tokens in the text do not have semantic roles relevant to crisis events. Several semantic role labels are found in all event types, including INFORMATION-ARG, STREET-ARG, AFFECTEDOBJECTS-ARG, TIME-ARG, PLACE-ARG, O, REASON-ARG, DEATHVICTIM-ARG, WOUNDVICTIM-ARG, and OFFICER-ARG. These labels are widespread because their roles are general and not limited to a particular event type. In addition, FALSE- EVENT has a relatively large number and appears in various types of events due to the presence of text that does not report an event that is or has occurred but rather a warning, speculation, or unverified information. Meanwhile, labels that indicate the main event type, such as FLOOD-EVENT, FIRE-EVENT, ACCIDENT-EVENT, and EARTHQUAKE-EVENT, tend to only appear in the appropriate event category. However, there are cases where a single tweet contains more than one event type, such as in the text "*tol jombang mojokerto - untuk update kecelakaan yang menyebabkan kendaraan terbakar* (jombang mojokerto toll - for updates on accidents causing vehicles to burn)." This case shows that vehicle accidents can lead to fires in some situations, especially if a fuel leak or a hard impact triggers a fire. Therefore, texts reporting accidents can also contain information about fires. This label distribution reflects the complexity of processing crisis event texts, where a single report can include more than one event type and the presence of speculative or unconfirmed information.

Figure 4 displays a word cloud illustrating the most frequently used words (in Indonesian) associated with the REASON-ARG semantic role labels based on different event types: (a) flood, (b) accident, (c) fire, and (d) earthquake. Meanwhile, Figure 5 presents a word cloud that translates these words into English. From the visualization, it can be seen that for floods, words such as "*hujan deras* (rain torrential)," "*luapan sungai* (river overflowing)," and "*drainase rusak*(broken drainage)" often appear as the leading causes. Words that are often associated with the cause of accidents are "*menabrak pembatas* (hit barrier)" and "*pengemudi mengantuk* (driver sleepy)." Terms like "*korsleting listrik* (short circuit)," "*ledakan gas* (gas explosion)," and "*puntung rokok* (cigarette butts)" are prevalent, indicating common fire causes. Meanwhile, words such as "*pergerakan lempeng* (plate movement)," "*zona subduksi* (subduction zone)," and "*aktivitas sesar* (cesarean activity)" are the causes of earthquakes. This visualization provides insight into how the causes of various crisis events are represented in tweets.

2) SEMI-SUPERVISED LEARNING

The training process of this SRL model begins by loading the labeled training data stored in CSV format. The data is then processed through the df_to_dataset function to run a series of preprocessing stages, such as padding, which transforms the raw data into a format suitable for model training by ensuring that all token sequences have uniform lengths. The labeled training data entered into the df_to_dataset function is named train_val and then separated into two parts, D_train and D_val, according to Step 2 of Algorithm 1. This division aims to monitor model performance during training and prevent overfitting. The transformer-based model from Hugging Face is trained using the train_val data. After the initial model is trained, a self-training technique is applied by utilizing unlabeled data, and the filter_threshold function plays a crucial role, as seen in the pseudocode in Algorithm 1. The function loads the trained model and uses it to predict token labels on the unlabeled dataset. The model generates a token label prediction and an associated confidence score for each text in the unlabeled dataset.

One key aspect of Algorithm 1 is using a threshold to filter out predictions with high confidence. This study explores three threshold values: 0.7, 0.8, and 0.9, chosen to represent common selection strategies used in self-training. A threshold of 0.7 allows the model to include more pseudo-labeled data, even with lower confidence, which may be helpful in early training rounds to expand data coverage. A threshold of 0.8 reflects a more balanced selection, filtering out low-confidence predictions while retaining sufficient instances. The 0.9 threshold enforces stricter filtering by accepting only highly confident predictions, which can help reduce label noise, especially in imbalanced datasets. This selection allows the observation of performance trends under recall-oriented, balanced, and precision-oriented conditions. Algorithm 1 provides a general illustration with a threshold of 0.9. The experiment repeated the process for thresholds of 0.7, 0.8, and 0.9 to compare model performance in recall-oriented, balanced, and precision-oriented scenarios.

Then, the confidence scores of each token in the text are averaged to determine the overall confidence of a prediction. This averaging process is done because the model is trained at the sentence level, while the SRL task operates at the word level. Therefore, averaging the confidence scores of each token is necessary to align with how the model processes and understands text. In addition, in a text, some tokens have very high or very low confidence scores compared to others. By taking the average, the influence of extreme tokens can be minimized, making data selection more stable. This averaging process also plays an important role in data selection for retraining because only texts with high overall confidence levels are selected. Thus, the data used in retraining is of better quality, which can ultimately improve the model's performance in subsequent iterations.

Then, suppose the average confidence score of a text exceeds a predetermined threshold (for example, in Algorithm 1, threshold 0.9). In that case, the predicted token labels of the text are considered valid enough and are added to the training dataset as pseudo-labeled data. The pseudo-labeled data that meets this threshold is then combined with the initial training dataset, and the model is retrained using the updated dataset. This process is repeated iteratively, where each iteration adds new pseudo-labeled data to the training dataset. This iterative loop stops when no more pseudo-labeled data meets the threshold criteria. Once the self-training process is complete, the final model is evaluated using the testing dataset. This evaluation involves calculating performance metrics such as accuracy, precision, recall, and F1-score. In the context of Algorithm 1, x_i refers to the text in the unlabeled dataset, y_i is the predicted token label, and $conf_i$ is a list of confidence scores for each token.

3) SRL MODELING

This study explored several transformer-based models for SRL tasks related to crisis events in low-resource languages, specifically Indonesian. Details of the transformer models are presented below:

a: BIDIRECTIONAL ENCODER REPRESENTATIONS FROM TRANSFORMERS (BERT)

BERT is a bidirectional Transformer-based NLP model that understands words by considering context from both directions (left and right) simultaneously. The BERT architectural structure is formed through stacking encoders designed to produce contextual representations of text. Each encoder layer in BERT consists of a self-attention mechanism and a feed-forward layer. The self-attention mechanism allows the

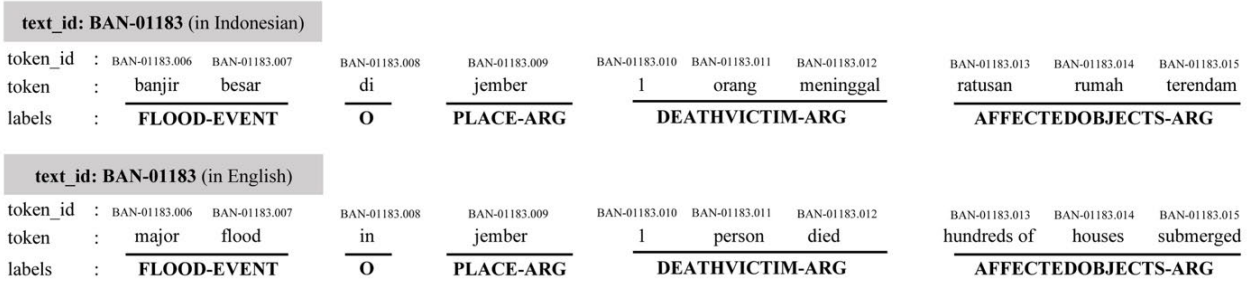


FIGURE 3. Examples of sentences labeled with semantic role for SRL tasks related to crisis events on low-resource language text data (Indonesian).

TABLE 4. Definition of semantic role labels related to crisis events in low-resource languages (Indonesian).

No.	Semantic Role Label	Definition	Example of Words/Phrases	Example of Complete Sentences in Indonesian	Example of Complete Sentences in English
1	EARTHQUAKE-EVENT	A semantic role indicates that the event discussed in the text is related to an earthquake.	*"gempa* (earthquake)*"*	*gempa 4,9 sr guncang mamuju tengah*	4.9 magnitude earthquake shakes central Mamuju
2	INFORMATION-ARG	A semantic role that indicates the details of a report explaining an event.	*"4,9 sr* (4.9 magnitude)*"*		
3	FALSE-EVENT	A semantic role indicates that a statement in the text does not describe an actual event.	*"banjir rob* (rob floods)*"*	*fenomena gerhana bulan total memang indah tapi warga harus waspada banjir rob punya tips atau berita warga lainnya? yuk posting di atmago!*	The phenomenon of total lunar eclipse is indeed beautiful, but residents must be aware of rob floods, have tips or other news from residents?
4	STREET-ARG	A semantic role that refers to the name of a street in an area where the event took place.	*"jl. h. rusini 2* (h. rusini 2 street)*"*	*update: kebakaran rumah di jl. h. rusini 2, kel. meruya selatan, kec. kembangan, jakarta barat. api sudah padam pukul 2.53 wib. tidak ada korban.*	Update: House fire on h. rusini 2 street, Meruya Selatan Village, Kembangan District, West Jakarta. The fire was extinguished at 2:53 WIB. There were no victims.
5	FIRE-EVENT	A semantic role indicates that the event discussed in the text is related to a fire.	*"kebakaran* (fire)*"*		
6	AFFECTEDOBJECTS-ARG	A semantic role that refers to objects or assets affected by the event.	*"ratusan rumah terendam* (hundreds of houses submerged)*"*	*hujan deras sejak sabtu akibatkan banjir besar di jember 1 orang meninggal ratusan rumah terendam*	Heavy rain since saturday caused major flood in Jember, 1 person died, hundreds of houses submerged
7	TIME-ARG	A semantic role that indicates the time of the event.	*"sabtu* (saturday)*"*		
8	O	A word or phrase that does not have a semantic role in the sentence structure related to any event.	*"sejak* (since)*", "akibatkan* (caused)*", "di* (in)*"*		
9	PLACE-ARG	A semantic role that indicates the location or place where an event occurs in the text, such as a city, village, or specific area mentioned in a tweet.	*"jember* (one of the city names in Indonesia, Jember City)*"*		
10	REASON-ARG	A semantic role that indicates the cause or reason for an event.	*"hujan deras* (heavy rain)*"*		
11	DEATHVICTIM-ARG	A semantic role that indicates the number or identity of victims who died in an event, including individuals or groups who were fatally affected.	*"1 orang meninggal* (1 person died)*"*		
12	FLOOD-EVENT	Semantic roles indicate that the events in the text are related to flooding.	*"banjir besar* (major flood)*"*		
13	OFFICER-ARG	Semantic roles that indicate official officers involved in an event, such as firefighters, police, or medical personnel.	*"kakorlantas polri irjen* (Head of Traffic Police of the National Police, Inspector General)*"*	*kakorlantas polri irjen lumowa menyebut ada beberapa korban luka akibat kecelakaan di tol cikampek*	Head of Traffic Police of the National Police, Inspector General Lumowa, said several victims were injured due to an accident on the Cikampek toll road.
14	ACCIDENT-EVENT	Semantic roles indicate that the events in the text are related to accidents, such as traffic accidents.	*"kecelakaan* (accident)*"*		
15	WOUNDVICTIM-ARG	Semantic roles that indicate the number or identity of victims injured in an event.	*"ada beberapa korban luka* (several victims were injured)*"*		

FIGURE 4. The most frequent words (in Indonesian) for the semantic role label "REASON-ARG" in the training data by event type: (a) flood, (b) accident, (c) fire, (d) earthquake.

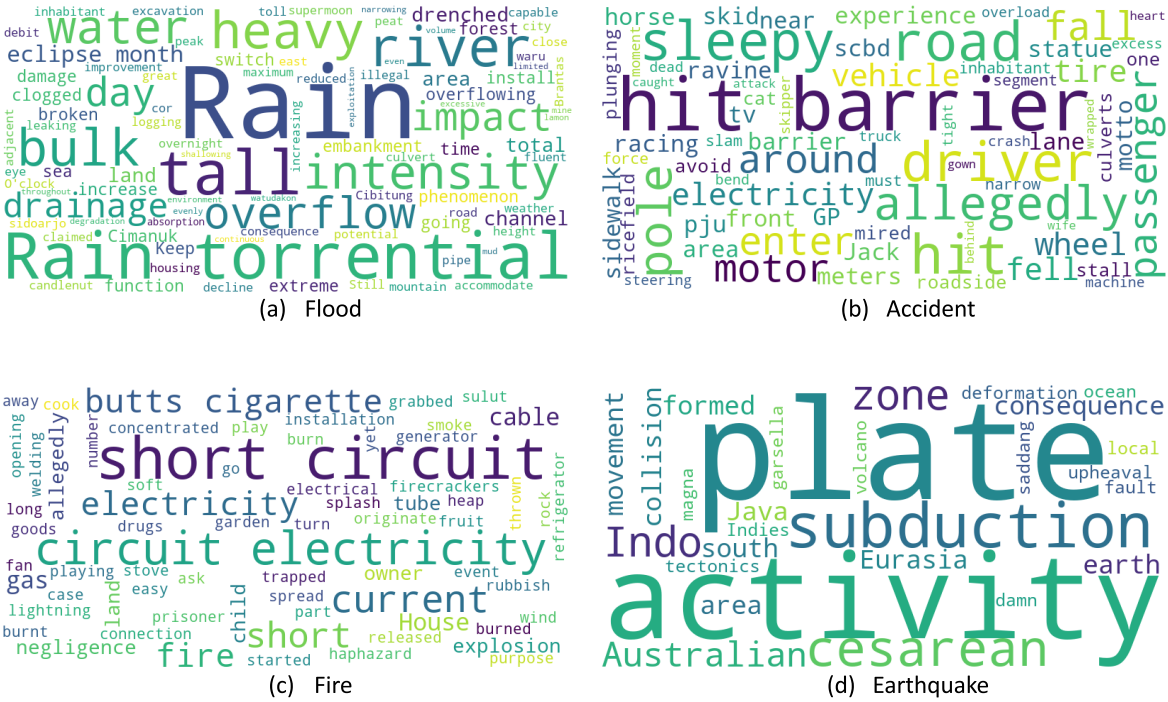


FIGURE 5. The most frequent words (in Indonesian translated into English) for the semantic role label "REASON-ARG" in the training data by event type: (a) flood, (b) accident, (c) fire, (d) earthquake.

model to calculate the relevance weight between a token and another token in a sentence so that the model can understand the relationship between words more contextually. After the self-attention process, a feed-forward layer applies non-linear transformations to each token's representation, refining its features and improving the model's ability to capture complex semantic relationships in the text [51].

Our research will explore IndoBERT [52] for SRL tasks on crisis events. IndoBERT is a transformer-based model specially trained in Indonesia that adopts the BERT-Base

TABLE 5. Overall frequency distribution of semantic roles labels on labeled training data and testing data.

Semantic Roles Labels	Amount of Data	Event				Total Frequency	Total %
		Flood	Accident	Fire	Earthquake		
O	#training	18302	8030	14952	11925	53209	69.74
	#testing	3743	3060	4335	4406	15544	67.85
PLACE-ARG	#training	2340	894	2934	1697	7865	10.31
	#testing	554	342	914	554	2364	10.32
FALSE-EVENT	#training	982	221	450	1017	2670	3.50
	#testing	143	100	101	434	949	4.14
STREET-ARG	#training	145	991	1062	7	2205	2.89
	#testing	60	546	334	9	778	3.40
AFFECTEDOBJECTS-ARG	#training	190	592	1279	60	2121	2.78
	#testing	70	207	360	31	668	2.92
TIME-ARG	#training	260	230	943	392	1825	2.39
	#testing	94	101	275	129	599	2.61
OFFICER-ARG	#training	415	104	705	444	1668	2.19
	#testing	112	34	192	107	445	1.94
INFORMATION-ARG	#training	335	390	153	683	1561	2.05
	#testing	50	124	30	212	416	1.82
FIRE-EVENT	#training	0	1	644	0	645	0.85
	#testing	0	0	189	0	189	0.83
REASON-ARG	#training	253	128	170	46	597	0.78
	#testing	70	39	37	16	162	0.71
ACCIDENT-EVENT	#training	0	472	0	0	472	0.62
	#testing	0	211	0	0	211	0.92
DEATHVICTIM-ARG	#training	76	145	192	27	440	0.58
	#testing	17	109	43	23	192	0.84
EARTHQUAKE-EVENT	#training	2	0	0	438	440	0.58
	#testing	1	0	0	138	139	0.61
FLOOD-EVENT	#training	300	0	3	0	303	0.40
	#testing	167	0	0	0	167	0.73
WOUNDVICTIM-ARG	#training	86	119	68	4	277	0.36
	#testing	13	40	23	9	85	0.37

architecture. This model consists of 12 hidden layers with a dimension of 768, 12 attention heads, and a feed-forward layer with a hidden size 3,072. To ensure better contextual understanding in Indonesian, IndoBERT is trained using 220 million words collected from various sources, including news articles (55 million words), the Indonesian Web Corpus (90 million words), and the Indonesian Wikipedia (74 million words). During the training process, this model uses Indonesian WordPiece vocabulary tokenization with a size of 31,923 words and processes data in document blocks containing 512 tokens per batch, allowing for more efficient and contextual text modeling.

Figure 6 illustrates the IndoBERT architecture used in our research to perform SRL on crisis events. The bottom layer represents the raw input sentences that the IndoBERT model will process. The sentence is first split into subtokens using WordPiece tokenization, with special tokens such as [CLS] to mark the beginning of the sequence and [SEP] as a separator between sentences in the input. Adding these tokens aims to help the model distinguish between different input parts. These subtokens are then converted into numeric vector representations in the embedding layer. Next, the BERT encoder layer processes the input embedding using self-attention mechanisms and feed-forward networks to capture the context and meaning of the text. Above the encoder layer, a dense layer is added to generate a probability distribution over the semantic role labels for each token. This process finally produces semantic role labels that match the semantic structure of the input sentence.

b: ROBUSTLY OPTIMIZED BERT PRE-TRAINING APPROACH (ROBERTA)

RoBERTa is a Transformer-based model developed as a variation of BERT with the main difference in the pretraining strategy [53]. This model retains the basic architecture of BERT but changes some aspects of its training, such as eliminating the Next Sentence Prediction (NSP) task, using a larger dataset and batch, and implementing dynamic masking. In the dynamic masking scheme, the tokens are masked during a pretraining change in each epoch, unlike the static approach used in BERT. This difference affects how the model learns text representations, especially in capturing relationships between words in different contexts.

RoBERTa can be adapted for various languages, including the IndoRoBERTa model, which is specifically designed to understand Indonesian text. This study uses IndoRoBERTa for the SRL task on crisis events. IndoRoBERTa is a masked language model based on the RoBERTa architecture with around 84 million parameters. It is trained using 3.1 GB of text from Indonesian Wikipedia.

Algorithm 1 Self-Training with Filtering Function for SRL Models

Input:
L ← Labeled dataset (CSV)
U ← Unlabeled dataset
T ← Confidence threshold (e.g., 0.9)
M ← Transformer-based SRL model
Output:
M_final ← Trained model after self-training
Procedure:
D_L ← Preprocess(L) // padding, tokenization, formatting
Split D_L into D_train and D_val (e.g., 80/20)
M ← Train(D_train, D_val) // Train initial model on labeled data
repeat
  P ← Predict(M, U) // Predict token labels and confidence scores
  S ← ∅ // High-confidence pseudo-labeled samples
  for each (x_i, y_i, conf_i) in P do
    avg_conf ← Average(conf_i) // Average token-level confidence
    if avg_conf ≥ T then
      S ← S ∪ {(x_i, y_i)} // Add to pseudo-labeled set
    end if
  end for
  if S ≠ ∅ then
    L ← L ∪ S // Add pseudo-labeled data to training set
    D_L ← Preprocess(L)
    Split D_L into D_train and D_val
    M ← Retrain(M, D_train, D_val) // Retrain with expanded data
  end if
until S = ∅ // Stop if no new pseudo-labeled data
Evaluate(M, TestSet) // Final evaluation on held-out test set
Return:
M_final ← M

c: GENERATIVE PRE-TRAINED TRANSFORMER (GPT)

GPT is a language model developed based on the Transformer architecture with a decoder-only approach. Unlike traditional encoder-decoder models, GPT consists of several Transformer-decoder layers that work autoregressively, where the prediction of the next word in a sentence only depends on the previous words [54].

As a more sophisticated version, GPT-2 presents several improvements compared to its predecessor (GPT). This model has 48 Transformer layers with a hidden dimension of 1600, and the number of parameters reaches 1.5 billion. In addition, the vocabulary size is expanded to 50,257 tokens, with the maximum context window increasing from 512 to 1024 tokens. These improvements make GPT-2 capable of handling longer and more complex texts and performing NLP tasks, such as text classification and generation [55].

In this study, we use a GPT-2 model adapted for Indonesian, namely GPT2-Indonesian (cahya/gpt2-small-indonesian-522M), to perform SRL tasks on crisis events using text data from Twitter. The GPT2-Indonesian small model is a variant of GPT-2 that has been trained using 522MB of Indonesian Wikipedia text, which has been processed using Byte Pair Encoding (BPE)-based tokenization with a vocabulary size of 52,000 tokens. This model can process input sequences of 128 consecutive tokens and has been optimized for various NLP tasks, such as classification and text generation. This model is uncased, so it does not distinguish between uppercase and lowercase letters.

d: LARGE LANGUAGE MODEL (LLM)

LLM is a transformer-based artificial intelligence model trained on a tremendous amount of text data to understand and generate language contextually [56]. This model has revolutionized various NLP tasks, such as translation, text summarization, classification, and interactive dialogue. Several popular LLMs, such as GPT-3.5, Llama-2, and Mixtral, have performed highly in various language-based tasks, especially in English [57]. There are developments in LLM models for low-resource languages, such as the Komodo language model [58], which was developed as a more specific solution for Indonesian. Komodo-7B is a large language model with 7 billion parameters, built on the Llama-2 architecture. The model is trained on various datasets, including

textbooks, everyday language usage, multilingual sources, and cross-language datasets.

Our research will explore the Komodo language model for SRL tasks in crisis events. However, the large size of the Komodo language model can present challenges, especially regarding memory usage and computational efficiency. We utilize quantization techniques to address the size issue as a solution [59]. Quantization refers to reducing the precision of numerical representations in a model. In the case of Komodo, quantization is performed using BitsAndBytesConfig to convert the model into a 4-bit format. This process occurs when the model is loaded into memory, allowing all model layers to be represented in a more compact format. Quantization does not change the basic architecture of the model. However, it significantly reduces the model size and speeds up computation, allowing Komodo to run on devices with limited resources without drastically sacrificing performance.

The basic architecture of the Komodo language model used in this study starts by taking a raw input sentence and breaking it down into meaningful tokens using a tokenizer from "Yellow-AI-NLP/komodo-7b-base." These tokens are then transformed into numerical representations through an embedding layer, which is the foundation for the model's understanding. Next, the encoder layer utilizes a self-attention mechanism to understand the context of each token. Low-rank adaptation is applied directly in the self-attention layers, allowing for more efficient fine-tuning without changing all model parameters. After going through the encoder layer, the results are forwarded to the dense layer, which classifies tokens. Finally, the model generates semantic role labels for each token in the input.

C. PHASE III: SRL MODELS EVALUATION

After the SRL model is trained, the next crucial step is to evaluate its performance. This evaluation is important to understand how well the model can identify semantic role labels from words in a sentence. In this study, we employ precision, recall, F1-score, and accuracy, which are standard metrics in sequence labeling tasks and provide a precise measure of the model's effectiveness. Since the proposed SRL framework emphasizes direct role extraction in crisis-related texts rather than predicate-centered analysis, metrics such as predicate accuracy or argument identification/classification are not utilized, as they are less relevant to our task formulation.

Precision measures the proportion of correct semantic role label predictions compared to all labels predicted by the model. Equation 1 shows the precision formula, where TP (True Positive) is the number of tokens correctly labeled as semantic role, while FP (False Positive) is the number of tokens incorrectly labeled as semantic role by the model.

$$Precision = \frac{TP}{TP + FP} \tag{1}$$

Meanwhile, recall measures how well the model recognizes all correct semantic role labels in the data, reflecting its ability to capture relevant labels. Equation 2 shows the recall formula, where FN (False Negative) is the number of tokens that should have been labeled as semantic roles but were not recognized by the model.

$$Recall = \frac{TP}{TP + FN} \tag{2}$$

The F1-score metric (in Equation 3) is calculated as the harmonic mean between precision and recall, which provides a balance between the two to avoid the dominance of one metric. Ultimately, accuracy represents the percentage of correctly predicted tokens in the dataset, providing a general measure of the model's performance. The accuracy formula can be seen in Equation 4, with TN (True Negative) meaning the number of correct tokens not labeled with semantic role.

$$F1 - score = 2 \times \frac{PrecisionRecall}{Precision + Recall} \tag{3}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{4}$$

IV. EXPERIMENTS AND ANALYSIS

This section assesses the effectiveness of Transformer-based SRL methods related to crisis events in Indonesian Twitter texts. We test various Transformer models and analyze

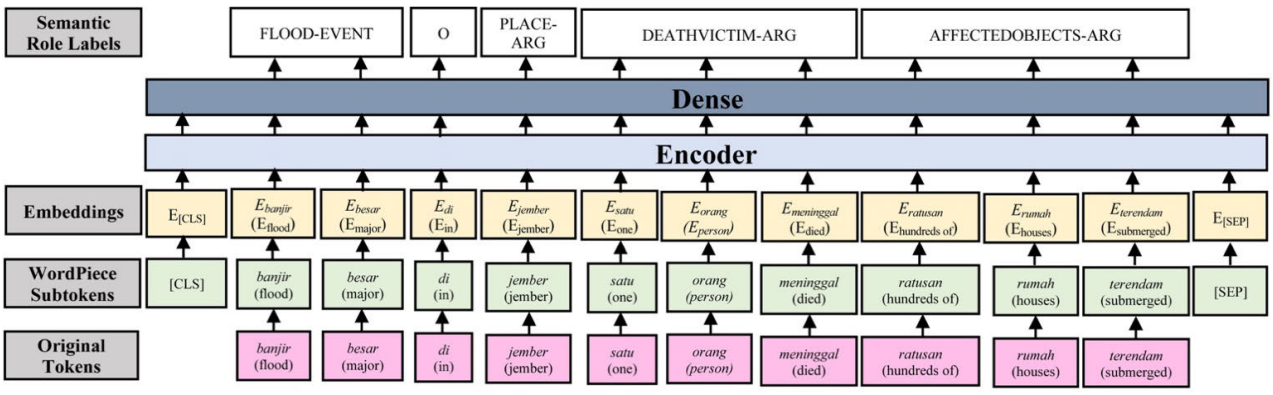


FIGURE 6. Architecture for SRL on crisis events using IndoBERT.

their performance based on relevant evaluation metrics. The experimental setup subsection covers dataset configurations, models, and training parameters for fair comparison. The results analysis compares model performance based on threshold values and identifies the most optimal model. In addition, we analyze prediction error patterns to understand the challenges in applying SRL to Twitter texts and its potential improvements in the future.

A. EXPERIMENTAL SETUP

We evaluate transformer-based models, with BERT, RoBERTa, GPT-2 and Llama-2 based architectures, for the SRL task on crisis events by utilizing the same training and testing data, following the split proportions outlined in Section III. In addition, we also apply the same hyperparameters (Table 6) to ensure a fair comparison between models. The learning rate is set to 2e-5, a commonly effective value for transformer-based models [60]. The batch size 16 is chosen to optimize the use of GPU memory without sacrificing computational efficiency. Training is carried out for 10 epochs with an early stopping technique to prevent overfitting, while a weight decay of 0.01 is applied to reduce model complexity and improve generalization. We also used a seed of 42 to ensure the experiment's reproducibility so that the results obtained could be relied upon and compared consistently. With this configuration, we ensure that each model is evaluated under optimal and fair conditions so that the experimental results can provide good insight into the effectiveness of each architecture in the SRL task on crisis events.

TABLE 6. Hyperparameters of models used to perform SRL on crisis events.

Hyperparameter	Definition	Value
Learning rate	A value determines how much the model weights are adjusted during each training step.	2e-5
Batch size	The number of data samples processed simultaneously in a single training iteration.	16
Epoch	A complete training cycle where the entire training dataset is passed through once.	10
Weight decay	A regularization technique that reduces the model weights to prevent overfitting.	0.01
Seed	The initial value is used to generate random numbers in the algorithm.	42

B. RESULTS ANALYSIS

In this study, we evaluate and compare the performance of several Transformer-based models in the SRL task on Twitter texts discussing crisis events. The models tested in this study consist of several Transformer architectures, namely: IndoBERT (indobert-base-uncased) based on BERT, IndoRoBERTa (indo-roberta-small) based on RoBERTa, GPT2-Indonesian (gpt2-small-indonesian) based on GPT-2, and Komodo (komodo-7b-base) based on LLaMA-2. The terms ''IndoBERT,'' ''IndoRoBERTa,'' ''GPT2-Indonesian'', and ''Komodo'' are used in this study to refer to each of these models.

We apply a semi-supervised learning approach, where the model is trained using a combination of labeled and unlabeled data iteratively. This process aims to improve the model's generalization by utilizing unlabeled data that is gradually labeled based on the confidence score obtained from the previous iteration. For instance, the token ''*Jalan Soekarno - Hatta* (Soekarno - Hatta Street)'' initially had a confidence score of 0.687. However, after several iterations, this score increased to 0.868. This improvement indicates that the model is becoming more confident in its predictions as it undergoes re-training. To assess the performance of each model, we used three threshold values (0.7, 0.8, and 0.9) as the minimum limit of the model's confidence level in determining the label. The evaluation was carried out by measuring accuracy, precision, recall, and F1-score to gain a deeper understanding of the effectiveness of each model in the SRL task on Indonesian texts.

Table 7 shows that the IndoBERT model performs best among all tested models, reaching an F1-score of 0.863 at a threshold of 0.9. IndoBERT performs exceptionally well because it is specifically optimized for the Indonesian language, enabling it better to capture sentence structure and context in Indonesian texts. Furthermore, BERT's bidirectional architecture enhances its ability to model semantic relationships across tokens, making it particularly effective for SRL tasks. A sensitivity analysis across thresholds of 0.7, 0.8, and 0.9 illustrates this trend. At 0.7, the F1-score is 0.721, reflecting noisy pseudo-labels; at 0.8, the score improves to 0.818, striking a balance between coverage and reliability; and at 0.9, the score peaks at 0.863, confirming that stricter confidence filtering strengthens SRL accuracy in noisy, low-resource Twitter texts. Overall, this sensitivity analysis demonstrates that threshold selection directly shapes IndoBERT's labeling quality: higher thresholds enhance precision by reducing noise, while lower thresholds expand recall at the cost of reliability.

On the other hand, the Komodo model shows competitive performance in SRL tasks, although it is still below IndoBERT. One factor influencing this difference is the quantization process during the implementation of Komodo, which aims to reduce model size and increase computational efficiency. However, the effect of this quantization does not significantly reduce model performance. At a threshold of 0.9, Komodo records an F1-score of 0.857, slightly different from IndoBERT. Komodo's main advantage lies in its significant architecture, with 7 billion parameters, allowing the model to capture more complex semantic patterns, including word-to-word relationships in SRL. In addition, the large model capacity and its adaptation to Indonesian provide good generalization capabilities, especially in dealing with unstructured texts such as tweets on Twitter, which often contain informal language, abbreviations, and non-standard sentence structures in Indonesian.

Meanwhile, IndoRoBERTa shows a slight advantage over IndoBERT at a low threshold (0.7), with an F1-score of 0.755 compared to 0.721 for IndoBERT. However, when the threshold increases to 0.9, IndoBERT experiences a more stable and significant increase, with the F1-score increasing to 0.863, while IndoRoBERTa only increases to 0.844. This trend shows that at a low threshold, IndoRoBERTa makes more frequent predictions than IndoBERT, so the F1-score appears higher. However, as the threshold increases, the model becomes more selective, only retaining highly confident predictions. As a result, IndoRoBERTa's recall decreases more sharply than IndoBERT, indicating that many instances previously classified with low confidence no longer meet the threshold and are eventually ignored.

The GPT2-Indonesian model shows a similar trend to IndoRoBERTa, with lower performance than IndoBERT and Komodo, especially at a high threshold. At a threshold of 0.9, the F1-score of GPT2-Indonesian only reaches 0.824, lower than IndoBERT (0.863) and Komodo (0.857). The low value of GPT2-Indonesian is due to the autoregressive architecture that only reads text from left to right to predict the next word in a sequence. In contrast, IndoBERT uses a bidirectional architecture, which allows the model to understand the context from both directions in a sentence. A thorough understanding of the relationships between words is essential in SRL tasks. Therefore, bidirectional models like IndoBERT are superior in capturing semantic patterns compared to autoregressive models like GPT2-Indonesian, which have limitations in seeing the overall context.

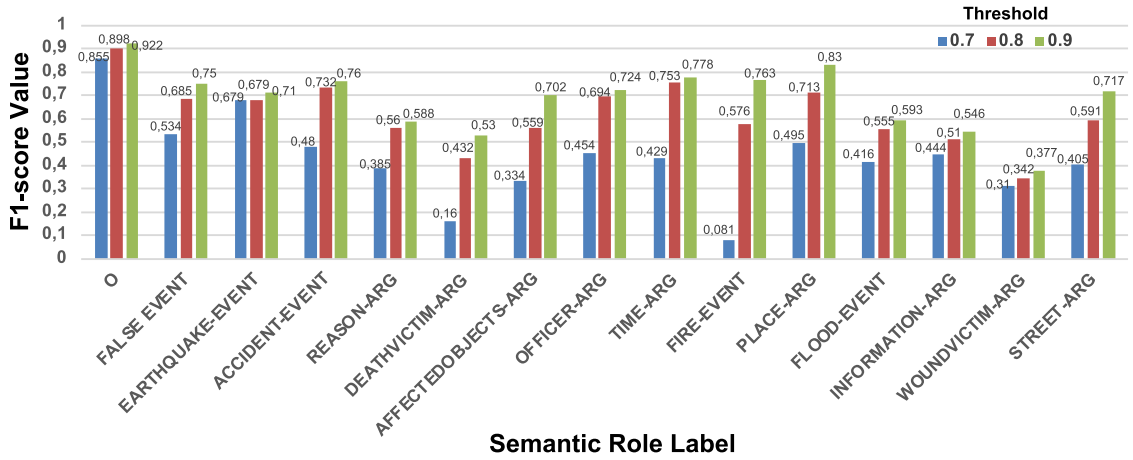
In addition to using Transformer-based models, we also implement a non-Transformer model in the form of BiLSTM. The architecture used consists of an Embedding layer (128-dimensional) that maps each token to a vector representation, followed by a 64-unit Bidirectional LSTM layer to capture context from the left and right directions simultaneously, and then ends with a TimeDistributed-Dense soft-max of size n_tags to predict the label of each token. BiLSTM was chosen because it is relatively lightweight (parameters <5M), easily reproducible, and with a bidirectional architecture, it can still utilize sequential context, even without a self-attention mechanism.

In semi-supervised self-training experiments, BiLSTM produces very high accuracy but much lower precision, recall, and F1-score. This phenomenon arises due to class imbalance: more than 90% of tokens in the corpus are labeled "O" (outside the semantic role). A model that guesses "O" for almost all tokens will achieve accuracy close to the majority label proportion, while failing to recognize important minority labels (e.g., ACCIDENT-EVENT, PLACE-ARG). When the threshold is increased from 0.7 to 0.9, the precision value does fluctuate slightly. However, the F1-score is almost stagnant because the decrease in recall offsets the increase in precision, indicating that BiLSTM's confidence in non-"O" labels is relatively low, so only a few additional tokens pass the pseudo-label selection in each iteration.

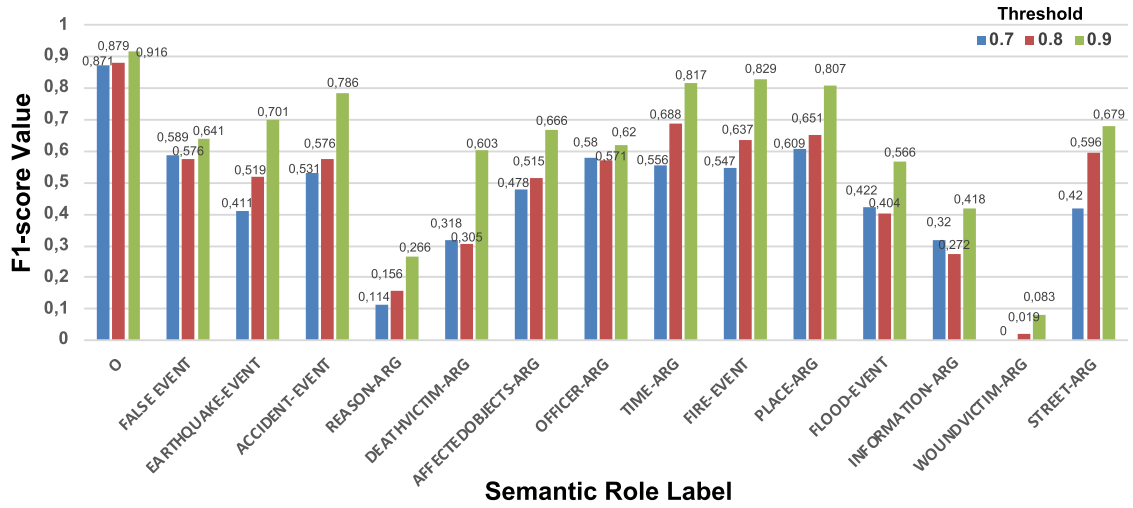
Compared with BiLSTM models, transformer-based models such as IndoBERT show consistent superiority in SRL tasks, especially in imbalanced crisis data that uses informal language. Transformer uses a bidirectional self-attention mechanism that can understand the relationship between tokens, both close and distant, to form a more accurate context representation sensitive to variations in meaning. In contrast,

TABLE 7. Performance of several transformer and non-transformer-based models for SRL tasks on crisis events.

Model Type	Model Name	Threshold	Accuracy	Precision	Recall	F1-Score
Transformer	BERT (indobert-base-uncased)	0.7	0.764	0.779	0.764	0.721
		0.8	0.832	0.831	0.832	0.818
		0.9	**0.868**	**0.865**	**0.868**	**0.863**
	RoBERTa (indo-roberta-small)	0.7	0.789	0.794	0.789	0.755
		0.8	0.802	0.800	0.802	0.777
		0.9	0.855	0.847	0.855	0.844
	GPT-2 (gpt2-small-indonesian)	0.7	0.788	0.788	0.788	0.755
		0.8	0.823	0.817	0.823	0.805
		0.9	0.839	0.832	0.839	0.824
	Llama-2 (komodo-7b-base)	0.7	0.778	0.780	0.778	0.738
		0.8	0.863	0.856	0.863	0.855
		0.9	0.865	0.859	0.865	0.857
Non-Transformer	BiLSTM	0.7	0.950	0.760	0.500	0.560
		0.8	0.950	0.730	0.550	0.590
		0.9	0.950	0.750	0.530	0.590



(a) BERT model (indobert-base-uncased)



(b) RoBERTa model (indo-roberta-small)

FIGURE 7. Comparison of F1-scores from various Transformer-based models for each semantic role label on Twitter text data. Subfigures (a) to (b) represent the performance of BERT, and RoBERTa models under confidence thresholds of 0.7, 0.8, and 0.9.

the BiLSTM model produces predictions with a more even level of confidence between labels, making it difficult to distinguish which tokens are truly confident and which are not. As a result, when the threshold is increased, the number of tokens that pass the pseudo-label selection does not increase much, and the quality of the added labels does not increase significantly. This causes the BiLSTM F1-score value only to increase slightly or even tend to stagnate in the range of 0.56–0.59, even though the threshold is increased from 0.7 to 0.9. This finding confirms that high performance in SRL is not only determined by the semi-supervised learning strategy but is highly dependent on the representational capacity of the model architecture used, where Transformer excels in capturing complex language structures and contexts.

In general, the experimental results in Table 7 show that each model experiences an increase in performance as the threshold increases. However, when analyzed in more depth based on the semantic role label, as presented in Figure 7 and Figure 8, it is found that not all labels experience a consistent increase in performance. Increasing the threshold can cause a decrease in the F1-score on some semantic role labels, indicating that the model becomes more selective in determining predictions so that some instances previously classified as positive no longer meet the confidence threshold.

One prominent example is the FLOOD-EVENT label on GPT2-Indonesian, which shows a fluctuating trend. At a threshold of 0.7, the F1-score reaches 0.253, then increases significantly at a threshold of 0.8 to 0.528 but decreases again to 0.236 at a threshold of 0.9. A fluctuating pattern is also seen in the DEATHVICTIM-ARG label for the IndoRoBERTa model, where the F1-score at a threshold of 0.7 is 0.318 but decreases to 0.305 at a threshold of 0.8 before finally increasing significantly to 0.603 at a threshold of 0.9. This fluctuation shows that increasing the threshold does not always guarantee a consistent increase in performance in a particular model.

In contrast, the IndoBERT model shows a stable increasing trend across all labels without experiencing performance fluctuations like other models. This increasing trend shows that IndoBERT is better able to maintain consistent performance

at each threshold, making it the most reliable model for the SRL task on Twitter texts related to crisis events. Thus, although increasing the threshold contributes to increasing model performance overall, the analysis per label semantic role results show that the effect of increasing the threshold can vary depending on the characteristics of the label and the model used. Models with bidirectional architecture, such as IndoBERT, are superior in maintaining

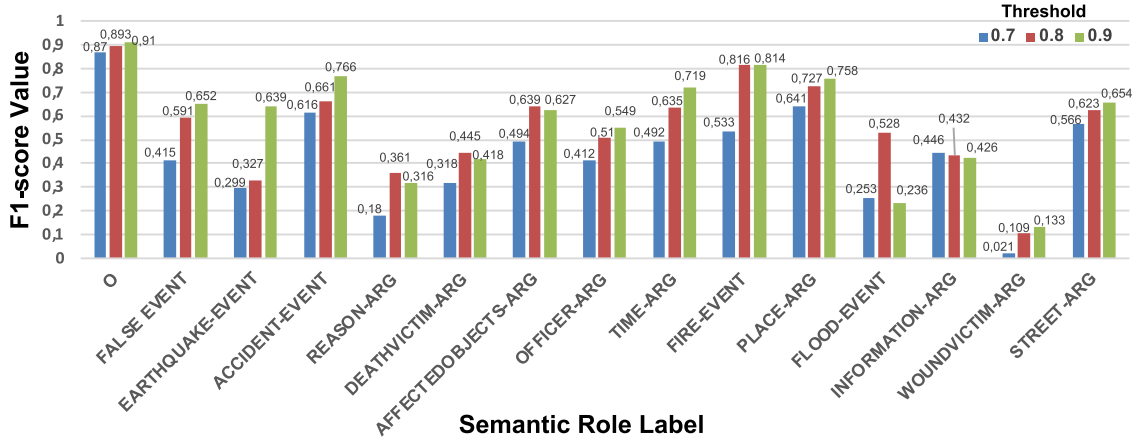
F1-score stability at high thresholds. Other models, like IndoRoBERTa, Komodo, and especially autoregressive models such as GPT2-Indonesian, exhibit more significant performance fluctuations between thresholds, resulting in inconsistent prediction outcomes across various semantic role labels.

C. ERROR ANALYSIS

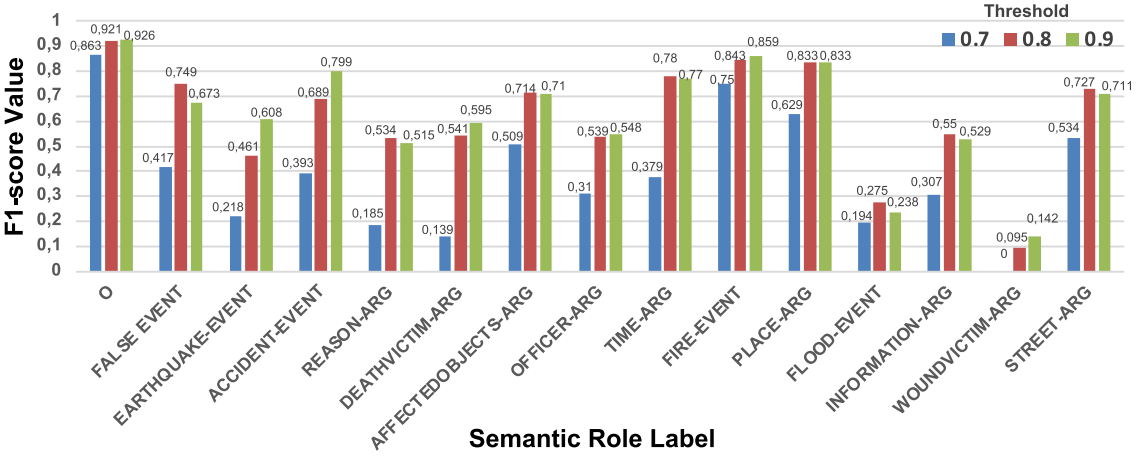
This error analysis subsection aims to identify patterns of mistakes that occur during the evaluation process. It seeks to understand the factors leading to incorrect predictions and to provide insights into the challenges of conducting SRL on crisis event texts from Indonesian Twitter. This error analysis will be accomplished by examining the confusion matrix and providing examples of misclassifications from the best-performing model, IndoBERT, using a threshold of 0.9.

The confusion matrix value in Figure 9 shows that there is a tendency for the model to predict flood events (FLOOD-EVENT) as FALSE-EVENT incorrectly. In our analysis of the semantic role label prediction results in Table 7, we found that negative contexts contributed to specific errors. For instance, the phrase "*mengantisipasi banjir* (anticipating floods)" caused the model to interpret the event as non-existent, resulting in its classification as FALSE-EVENT. Thus, long and informal sentence structures and descriptive phrases like "*mengantisipasi* (anticipating)" in Twitter texts negatively affect the model's understanding.

In addition, based on the F1-score in Figure 7, the semantic role label WOUNDVICTIM-ARG has the lowest value. The confusion matrix shows that WOUNDVICTIM-ARG is often misclassified as DEATHVICTIM-ARG. This misclassification (as seen in Table 8) is likely influenced by Twitter's character limitations, which often lead to concise and less detailed disaster-related information. Due to these constraints, tweets may not indicate whether the victims



(c) GPT-2 model (gpt2-small-indonesian)



(d) Llama-2 model (Komodo-7b-base)

FIGURE 8. Comparison of F1-scores from various Transformer-based models for each semantic role label on Twitter text data. Subfigures (c) to (d) represent the performance of GPT-2, and Llama-2 models under confidence thresholds of 0.7, 0.8, and 0.9.

TABLE 8. Example of prediction results of semantic role labels by BERT model with threshold 0.9 for SRL task on crisis event.

No	Text_id	Token in Indonesia	Token in English	Actual Label	Predicted Label	Full Sentence in Indonesian	Full Sentence in English
1	BAN-00172	*banjir*	floods	FLOOD-EVENT	FALSE-EVENT	*setelah air surut sahabat banser dengan sigap mmbuat penanganan untuk meng-antisipasi banjir susulan di keclosari brebes*	After the water receded, Banser friends swiftly made handling to anticipate floods aftershock in Keclosari Brebes
2		*susulan*	aftershock	FLOOD-EVENT	FALSE-EVENT		
3	KEC-01123	*kuningan*	kuningan	PLACE-ARG	STREET-ARG	*430 : rt : kecelakaan mobil pick up mitsubishi l 300 b 9124 uai di ruas tol dalam kota (kuningan) arah cawang sudah dalam penanganan petugas*	430: RT: Mitsubishi L 300 B 9124 UAI pick-up car accident in toll road section in the city (Kuningan) towards Cawang is already being handled by officers
4	KEB-00490	*pasar*	market	AFFECTEDOBJECTS-ARG	PLACE-ARG	*polsek medan baru salurkan bantuan sembako kepada korban kebakaran di pasar meranti*	Medan Baru Police distribute essential food assistance to fire victims at Meranti Market
5		*meranti*	meranti	AFFECTEDOBJECTS-ARG	PLACE_ARG		
6	KEB-00634	*7*	seven	WOUNDVICTIM-ARG	DEATHVICTIM-ARG	*tadi pagi ada kecelakaan mobil yang ngebut, lepas kendali, dan menabrak 7 orang,*	This morning, there was a car accident that was speeding, out of control, and hit seven people,
7		*orang*	people	WOUNDVICTIM-ARG	DEATHVICTIM-ARG		
8	BAN-01104	*gempa*	earthquake	EARTHQUAKE-EVENT	EARTHQUAKE-EVENT	*habis diguncang gempa 5,3 sr pada 6/12/2018 pagi sorenya diterjang banjir di kecamatan kempo kabupaten dompu ntb 9 rumah rusak dan 120 rumah terendam banjir hutan makin rusak saat banjir sungai berisi kayu & sampah salah satu penyebab banjir di ntb adalah maraknya penebangan hutan*	After being shaken by a 5.3 SR earthquake on 6/12/2018 in the morning and evening, floods hit it in Kempo District, Dompu Regency, NTB. Nine houses were damaged, and 120 houses were submerged in floods. The forest was further damaged during the flood. The river was filled with wood and garbage. One of the causes of flooding in NTB is the rampant logging
9		*5,3*	5.3	INFORMATION-ARG	INFORMATION-ARG		
10		*sr*	sr	INFORMATION-ARG	INFORMATION-ARG		
11		*6/12/2018*	6/12/2018	TIME-ARG	TIME-ARG		
12		*banjir*	flood	FLOOD-EVENT	FLOOD-EVENT		
13		*di*	in	O	O		
14		*kecamatan*	district	PLACE-ARG	PLACE-ARG		
15		*kempo*	kempo	PLACE-ARG	PLACE-ARG		
16		*kabupaten*	regency	PLACE-ARG	PLACE-ARG		
17		*dompu*	dompu	PLACE-ARG	PLACE-ARG		
18		*ntb*	ntb (west nusa tenggara province)	PLACE-ARG	PLACE-ARG		
19		*9*	nine	AFFECTEDOBJECTS-ARG	AFFECTEDOBJECTS-ARG		
20		*rumah*	houses	AFFECTEDOBJECTS-ARG	AFFECTEDOBJECTS-ARG		
21		*rusak*	damaged	AFFECTEDOBJECTS-ARG	AFFECTEDOBJECTS-ARG		

mentioned are injured or deceased. As a result, the model assigns the DEATHVICTIM-ARG label more frequently, reflecting the ambiguity present in the textual data.

There is also an error in labeling AFFECTEDOBJECTS-ARG to PLACE-ARG, which can occur due to word ambiguity in Twitter text. The words ''*pasar* (market)'' and ''*meranti* (the name of the market is meranti)'' can be understood as locations by the model, while the actual context is the affected object. In addition, training data often shows a pattern where words after prepositions such as ''*di* (in)'' have a semantic role as a location, so the model prioritizes PLACE-ARG.

The misclassification of PLACE-ARG label as STREET-ARG is likely due to the ambiguity of location names in Twitter text. One example is the word ''Kuningan,'' which the model predicts as STREET-ARG, even though, in context, ''Kuningan'' refers to an administrative area (PLACE-ARG). This error arises because ''Kuningan'' has two possible meanings: it can refer to a district in West Java (Kuningan District, Kuningan Regency) or a street in Semarang City, Central Java (Kuningan Street). In the analyzed text, ''Kuningan'' appears in the phrase ''*ruas tol dalam kota (Kuningan) arah Cawang* (toll road section in the city (Kuningan) towards

Cawang).'' The surrounding context likely leads the model to associate ''Kuningan'' with the highway system, classifying it as STREET-ARG. The model heavily relies on nearby words like ''*tol* (toll),'' ''*ruas* (section),'' and ''*arah Cawang* (towards Cawang),'' which are more commonly linked to roads rather than administrative areas.

Despite some errors in label prediction, transformer-based models with BERT architecture in a threshold of 0.9 still perform very well in the SRL task on Twitter texts related to crisis events. This model can identify multiple events within a single sentence (in Table 8 with example sentences from text_id BAN-01104), such as EARTHQUAKE-EVENT and FLOOD-EVENT, while consistently assigning relevant semantic roles like TIME-ARG, INFORMATION-ARG, PLACE-ARG, and AFFECTED OBJECTS-ARG. This capability demonstrates the model's strong contextual understanding of complex disaster information. The self-attention mechanism further enhances its effectiveness. Thus, the model can distinguish two events in one sentence based on keywords such as ''*diguncang gempa*(rocked by an earthquake)'' and ''*diterjang banjir* (hit by a flood),'' each of which refers to a separate event. Moreover, a higher threshold reduces prediction errors by ensuring only high-confidence classifications, improving accuracy.

D. CROSS-DOMAIN EVALUATION ON NEWS ARTICLES

To evaluate the extent to which the SRL framework developed in this study can be generalized outside its training domain, cross-domain testing was conducted using the best-performing model, IndoBERT, with a confidence threshold of 0.9. Although this model was trained entirely using Indonesian Twitter data in the context of crisis events, we tested it on Indonesian news texts to assess the model's robustness to more formal text domain variations.

The test data consisted of five news documents obtained from IEEE Dataport [61], covering various types of crisis events: fire (1 document), accident (1 document), flood (2 documents), and earthquake (1 document). The total tokens in the test data reached 1,055, with an average of 12.6 sentences per document, 16.75 tokens per sentence, and 211 tokens per document. The model was used directly without retraining or additional domain adjustments.

The evaluation showed that the model improved performance on news texts compared to Twitter texts. This improvement is indicated by the increase in all primary evaluation metrics, including accuracy (from 0.868 to 0.896), precision (from 0.865 to 0.909), recall (from 0.868 to 0.896), and F1-score (from 0.863 to 0.899) as shown in Table 9. This increase is due to clearer sentence structures, consistent use of grammar, and more explicit expressions of events in the news text, which make it easier for the model to identify semantic roles.

However, not all labels showed improved performance. One example is the WOUNDVICTIM-ARG label, which saw its F1-score drop from 0.377 on the Twitter data to just 0.200 on the news data (Figure 10). This drop is likely due to variations in how injured victims are conveyed in a journalistic style that differs from Twitter, such as using medical terms or metaphors. The AFFECTEDOBJECTS-ARG label also saw its performance drop from 0.702 to 0.519. This drop could be attributed to more complex descriptive expressions in news stories, such as using long phrases or metaphors to describe the affected object, which do not always match the patterns the model learned from the Twitter data.

In contrast, the DEATHVICTIM-ARG label saw its F1-score increase from 0.530 to 0.625. This increase is likely because news stories tend to be explicit and consistent in terms of mentions of fatalities, such as the phrases ''*dua orang tewas* (two people died)'' or ''*korban meninggal dunia* (victims died),'' which provide linguistic signals that the model more easily recognizes. Overall, this cross-domain experiment shows that SRL models trained on social media data can maintain — and even improve — their performance when applied to news texts, without retraining. This finding is an early indication that the proposed approach has the potential to be generalized to formal domains with different language structures.

TABLE 9. Comparison of IndoBERT model performance with a threshold of 0.9 between Twitter data (in-domain) and news data (cross-domain).

Metrics	Twitter (in-domain)	News (cross-domain)
Precision	0.865	0.909
Recall	0.868	0.896
F1-score	0.863	0.899
Accuracy	0.868	0.896

E. QUALITATIVE COMPARISON WITH PROPBANK-BASED TRADITIONAL SRL

SRL performance evaluation is usually done quantitatively by comparing metrics such as F1-score against datasets with a uniform annotation scheme. However, in the context of this study, a direct numerical comparison between our proposed SRL system and the traditional PropBank-based approach would not provide meaningful insight. This problem is caused by the fundamental differences in the annotation schemes: PropBank uses the generic ARG0–ARG5 framework and relies heavily on the presence of explicit verbal predicates, while our system is designed for the crisis domain with more specific semantic labels such as FLOOD-EVENT, FALSE-EVENT, AFFECTEDOBJECTS-ARG, and PLACE-ARG. These schemes are intended to capture contextual and immediately useful information in emergency response, especially from informal texts in Indonesian. Attempting to align the two schemes would instead forcefully simplify our domain-specific labels into a generic framework, which could obscure the value of the contribution of our system. Therefore, qualitative analysis was chosen as the most appropriate approach to demonstrate the superiority of our system over PropBank-based traditional SRL, especially in understanding the semantic structure of crisis event texts. This analysis allows us to highlight the extent to which each system can or fails to capture the whole semantic meaning, including in

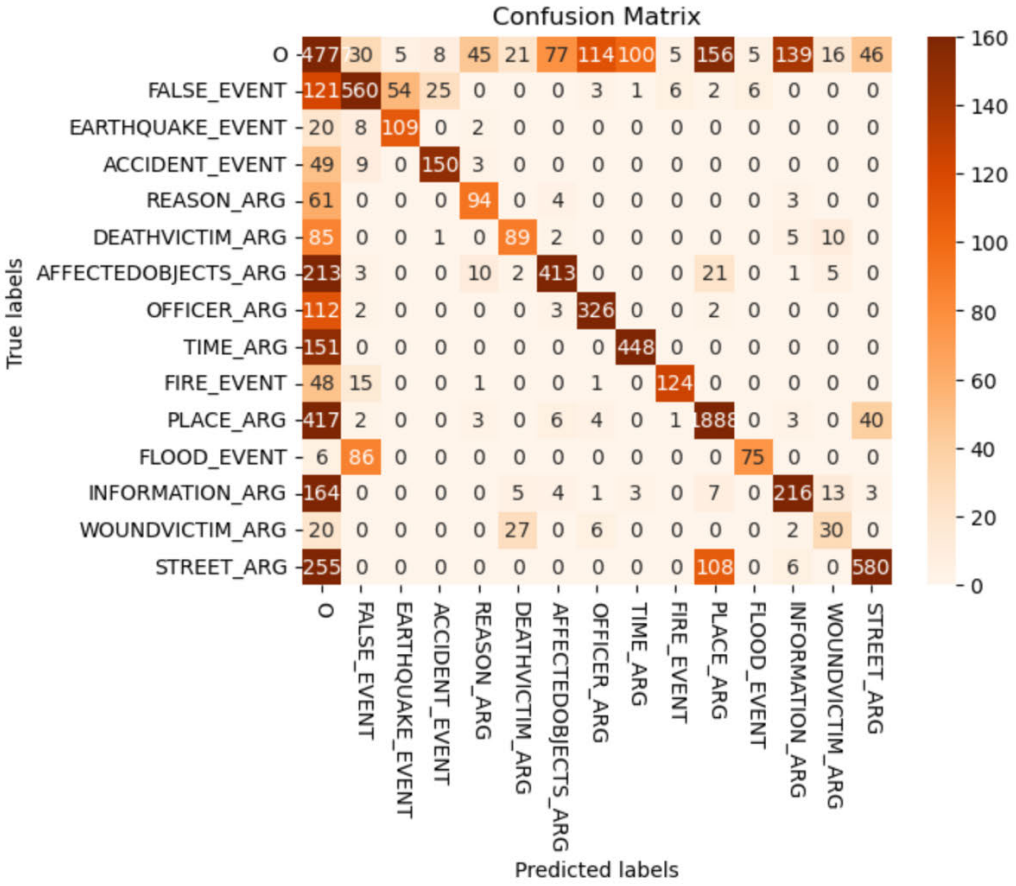


FIGURE 9. Confusion matrix of BERT model with threshold 0.9 for SRL task on crisis event.

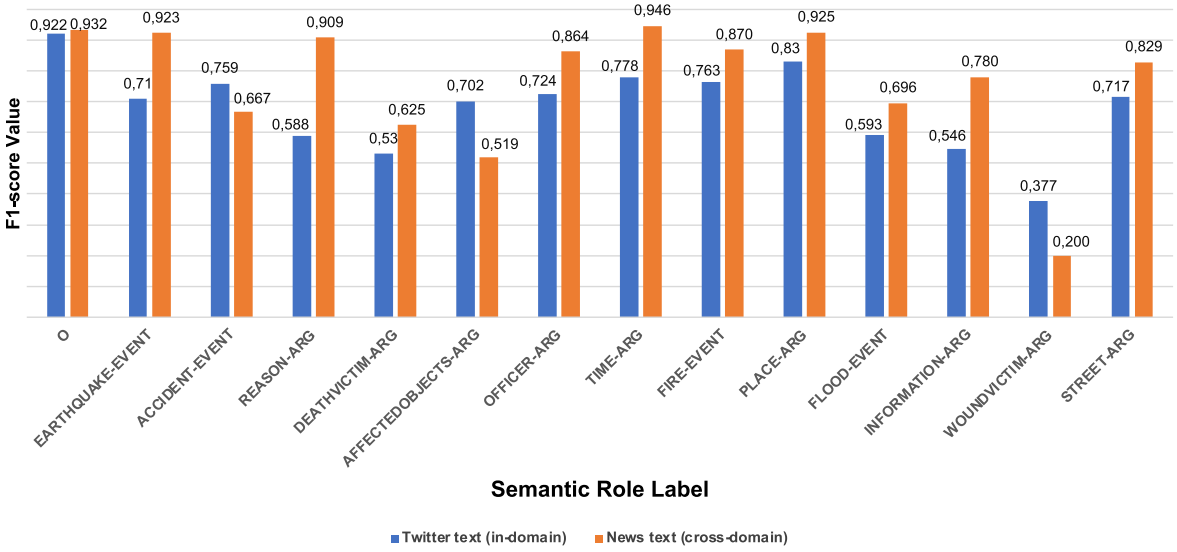


FIGURE 10. Distribution of F1-Score per Label in In-Domain and cross-domain testing.

sentences that do not have explicit predicates, contain factual ambiguity, or have complex structures commonly found in social media texts during the crisis.

To support this qualitative analysis, we compare the semantic annotation results of two different SRL systems: our proposed SRL system and the PropBank-based SRL system

(Table 10). For the analysis of our system, we first took 50 random tweet samples from the output of the best model, namely the IndoBERT model trained using a semi-supervised learning approach with a confidence threshold of 0.9. These samples were selected based on the predictions with the highest confidence level, which is assumed to represent the system's performance under optimal conditions. Furthermore, for the PropBank-based traditional SRL system analysis, we used the same tweet samples so that the comparison could be carried out equally. Each tweet was then translated into English using automatic machine translation, given that the PropBank-based SRL system used—HanLP v2.1.0 [62] —shows more stable and accurate performance on English input. These translated tweets are then processed using HanLP's SRL module to generate semantic annotations based on the PropBank scheme.

From 50 output pairs between HanLP and our system, four representative sentence examples are selected based on different linguistic categories: (1) sentences with nominalized events and no verbal predicates, (2) sentences with potentially non-factual information (FALSE-EVENT), (3) simple sentences with one explicit predicate, and (4) complex sentences with multiple predicates. The first category refers to sentences where the event is conveyed not through a verb (such as ''flooded'' or ''occurred''), but through a noun form of the event itself (e.g., ''flood''), without any accompanying verb that would typically serve as the predicate. These constructions pose particular challenges for predicate-centered SRL frameworks like PropBank, which require explicit verbal anchors to assign semantic roles. The selection of these four cases is intended to represent the diverse semantic challenges commonly found in crisis texts on social media and evaluate the system's ability to capture event information completely and meaningfully. Through direct comparison between the systems, we can demonstrate the strengths and limitations of each approach in detail and measurably.

The comparison results between PropBank-based SRL (using HanLP with English translation) and our proposed SRL show significant differences in semantic output, especially on Indonesian texts in the crisis event domain. In the first case, there is no explicit predicate in the sentence, so the PropBank-based system cannot produce any SRL output. The absence of a verb causes the system to fail to identify any semantic role. In contrast, our proposed SRL system is still able to identify ''*banjir* (flood)'' as FLOOD-EVENT and ''*kalibuntu village, losari sub-district, brebes* (kalibuntu village, losari sub-district, brebes)'' as PLACE-ARG, demonstrating the ability to understand crisis events expressed nominally or implicitly.

In the second case, the sentence ''*ini upaya pemkab sidoarjo atasi banjir di tropodo (odp-pr)* (this is the sidoarjo district government's effort to overcome flooding in tropodo (odp-pr))'' shows semantic ambiguity. Although PropBank can extract some semantic roles based on predicates such as ''is'', ''effort'', and ''overcome'', the output generated is fragmented and does not directly capture that the context is a potential event or unverified information. On the other hand, our system not only successfully recognizes ''*pemkab* (district government's)'' as OFFICER-ARG and ''*sidoarjo* (Sidoarjo is one of the cities in Indonesia)'' as PLACE-ARG, but is also able to label ''*banjir* (flooding)'' as FALSE-EVENT, indicating that our system can understand the context of information that is a hoax or not happening—an important feature in handling crisis information sourced from social media.

In the third case, a regular sentence with one main predicate, the PropBank system can generate annotations for the predicate was and embed the entire event as ARG1. However, the semantic labels generated are general and do not highlight important information structures in the crisis context, such as the affected objects or the location of the incident. In contrast, our system assigns the label FIRE-EVENT to ''*kebakaran* (fire)'', AFFECTEDOBJECTS-ARG to ''*sebuah gudang distributor furniture* (a furniture distributor warehouse)'', and PLACE-ARG to ''kawasan akong kec. sepatan kabupaten tangerang (akong area, sepatan district, tangerang regency)'', providing a more explicit and informative situational framework.

In the fourth case, which is a complex sentence with several predicates and two main clauses, ''*sudah hampir seminggu…warga pun terpaksa..*(it's been almost a week…residents have been forced …)'', PropBank-based traditional SRL produces five different predicate structures (for ''submerged'', ''been'', ''forced'', ''leave'', and ''take'') with semantic role fragmentation. This fragmentation makes it challenging to interpret the event as a whole. In contrast, our SRL system can produce a single semantic structure that describes the event as a whole: ''*banjir* (flood)'' as FLOOD-EVENT, ''*tiga dukuh* (three hamlets)'' as AFFECTEDOBJECTS-ARG, and ''*desa jati wetan kabupaten kudus* (Jati Wetan village, Kudus district)'' as PLACE-ARG, thus offering a more informative and relevant representation for crisis management.

The analysis of the four cases above strengthens the argument that the PropBank-based SRL approach, which relies heavily on explicit predicates, has significant limitations when applied to informal texts in the crisis event domain, especially in Indonesian on social media. The absence of predicates in nominal sentences, complex sentence structures with many clauses, and factual ambiguity in crisis information cause the PropBank system to be unable to produce meaningful annotations or even fail to produce results. The reliance on explicit syntactic structures and the generic ARG0–ARG5 annotation scheme makes this approach less sensitive to the context and needs of the crisis domain.

In contrast, our proposed SRL approach shows flexibility in handling variations in sentence structure and embedding domain-specific semantic roles such as FLOOD-EVENT, AFFECTEDOBJECTS-ARG, FALSE-EVENT, and PLACE-ARG. This system not only overcomes the limitations in detecting implicit events and nominalizations, but is also more relevant in identifying important information

TABLE 10. Comparison of PropBank-based and proposed SRL annotations across different case types of crisis-related tweets.

Case Type	Example Tweet	PropBank Annotation	Proposed Annotation
Nominalized sentence	*banjir di desa kalibuntu kecamatan losari brebes* (flood in kalibuntu village, losari sub-district, brebes)	-	**Output 1:** *Banjir* (FLOOD-EVENT) *di* (O) *desa kalibuntu* (PLACE-ARG) *kecamatan losari brebes* (PLACE-ARG)
Non-factual information sentence	*ini upaya pemkab sidoarjo atasi banjir di tropodo (odp-pr)* (This is the Sidoarjo district government's effort to overcome flooding in Tropodo (odp-pr))	• **Output 1 for "is" predicate:** This (ARG1) is (PRED) the Sidoarjo district (ARG2) government's effort to overcome flooding (ARG2) in Tropodo (odp-pr) (ARG2) • **Output 2 for "overcome" predicate:** This is the Sidoarjo district government's effort to overcome (PRED) flooding in Tropodo (odp-pr) (ARG1) • **Output 3 for "effort" predicate:** This is the Sidoarjo district government's (ARG0) effort (PRED) to overcome flooding in Tropodo (odp-pr)	**Output 1:** *ini upaya* (O) *pemkab* (OFFICER_ARG) *sidoarjo* (PLACE-ARG) *atasi* (O) *banjir* (FALSE_EVENT) *di* (O) *tropodo* (PLACE-ARG) *(odp-pr)* (O)
Simple sentence (with one predicate)	*telah terjadi kebakaran sebuah gudang distributor furniture di kawasan akong kec. sepatan kabupaten tangerang* (There was a fire at a furniture distributor warehouse in the Akong area, Sepatan district, Tangerang regency)	**Output 1 for "was" predicate:** There was (PRED) a fire at a furniture distributor warehouse in the Akong area, Sepatan district, (ARG1) Tangerang regency.) (ARG1)	**Output 1:** *telah terjadi* (O) *kebakaran* (FIRE_EVENT) *sebuah gudang distributor furniture* (AFFECTEDOBJECTS_ARG) *di* (O) *kawasan akong kec . sepatan kabupaten* (PLACE_ARG) *tangerang* (PLACE_ARG)
Complex sentence	*sudah hampir seminggu banjir masih merendam tiga dukuh di desa jati wetan kabupaten kudus. warga pun terpaksa meninggalkan rumahnya dan mengungsi ke posko banjir* (It's been almost a week since the floods have submerged three hamlets in Jati Wetan village, Kudus district. Residents have been forced to leave their homes and take refuge at the flood post)	• **Output 1 for "submerged" predicate-Phrase 1:** It's been almost a week since the floods (ARG0) have submerged (PRED) three hamlets in Jati Wetan village, Kudus district. • **Output 2 for "been" predicate-Phrase 1:** it's (ARG1) been (PRED) almost a week since the floods (ARGM.TMP) have submerged three hamlets in Jati Wetan (ARGM.TMP) village, Kudus district. (ARGM.TMP) • **Output 3 for "leave" predicate-Phrase 2:** Residents (ARG0) have been forced to leave (PRED) their homes and take refuge at the flood post. (ARG1) • **Output 4 for "forced" predicate-Phrase 2:** Residents (ARG1) have been forced (PRED) to leave (ARG2) their homes and take refuge at the flood post. (ARG2) • **Output 5 for "take" predicate-Phrase 2:** Residents have been forced to leave their homes and take (PRED) refuge (ARGM-PRR) at the flood post. (ARGM-LOC)	**Output 1:** *sudah hampir seminggu* (O) *banjir* (FLOOD_EVENT) *masih merendam* (O) *tiga dukuh* (AFFECTEDOBJECTS_ARG) *di* (O) *desa jati wetan kabupaten kudus.* (PLACE_ARG) *warga pun terpaksa meninggalkan* (O) *rumahnya* (PLACE_ARG) *dan mengungsi ke posko* (O) *banjir* (FLOOD_EVENT)

for emergency response purposes. The ability to capture the complete event structure from informal text is a significant added value in low-resource languages such as Indonesian, especially on platforms such as Twitter, which

are often used as data sources in crises. This flexibility is achieved by leveraging contextualized embeddings produced by a transformer-based model (IndoBERT), which captures the meaning of words depending on their context. This enables the SRL system to infer semantic roles from surrounding words even when predicates are absent or implicit. Furthermore, unlike PropBank-based SRL, which relies on the presence of explicit agents (predicate/ARG0), our model does not treat missing agents as a failure case. This is because the use of domain-specific semantic roles (e.g., DEATHVICTIM-ARG, WOUNDVICTIM-ARG, AFFECTEDOBJECTS-ARG, and PLACE-ARG) allows the system to directly extract crisis-relevant information, such as the number of deceased victims (DEATHVICTIM-ARG), the number of injured victims (WOUNDVICTIM-ARG), damaged infrastructures (AFFECTEDOBJECTS-ARG), and affected locations (PLACE-ARG), without relying on the presence of an explicit predicate. In this way, the resulting event structures remain meaningful even when explicit agents are absent in the sentence, which is crucial in crisis-related tweets where agents are often omitted.

V. CONCLUSION AND FUTURE WORKS

In this study, we develop and evaluate an SRL model focused on crisis events in Indonesian Twitter texts. Unlike the traditional PropBank-based approach, which classifies semantic roles based on predicates within sentences using labels ARG0-ARG5, our model does not rely on predicates as the main anchor. This allows greater flexibility in recognizing and capturing semantic role relationships across sentence structures.

To evaluate the effectiveness of this approach, we compare several Transformer-based SRL models: IndoBERT based on the BERT architecture, IndoRoBERTa based on the RoBERTa architecture, GPT2-Indonesian based on the GPT-2 architecture, and Komodo based on the Llama-2 architecture. Due to the limited labeled data in this domain, we apply a semi-supervised learning strategy through iterative pseudo-labeling. The experimental results show that a transformer model with a BERT-based architecture, adapted for a specific low-resource language, achieves the best performance, with the highest F1-score of 0.863 at a threshold of 0.9. Its superiority is mainly due to its bidirectional architecture and language-specific adaptation, which enhance its ability to capture semantic relationships and contextual nuances more effectively than other models.

Further analysis related to the selection of thresholds shows that increasing the threshold value generally improves model performance, although the effect varies depending on the model and labels tested. At a threshold of 0.7, the model becomes less strict in assigning labels, resulting in more instances classified as positive, even if some are misclassified. At a threshold of 0.8, the balance between recall and precision becomes more stable, but some labels still fluctuate. A threshold of 0.9 allows the model to maintain predictions with high confidence. Semi-supervised learning can improve model performance, especially when labeled data is scarce. The model can improve its prediction confidence by gradually incorporating unlabeled data. To ensure that the generated pseudo-labels are more accurate, our approach has been enhanced with a filtering function to filter out noise in the data so that only highly confident predictions are used in further training.

Despite the promising results, our study has three main limitations. First, its generalizability to other low-resource languages beyond Indonesian remains untested, raising questions about how well the proposed SRL framework adapts to different linguistic structures. Second, the dataset is constrained by Twitter API restrictions, which may limit scalability and representativeness. When applying the model to broader or more diverse disaster communication scenarios, these limitations should be considered. Third, the error analysis results revealed several common error patterns. One of the main factors is ambiguity in Twitter text, especially in place names with multiple meanings. For example, the misprediction of PLACE-ARG as STREET-ARG often occurs because of location names such as ''Kuningan,'' which could refer to a district in West Java or a street name in Semarang City, Central Java. In addition, the lack of explicit information in some texts also causes misclassification, such as the misprediction of DEATHVICTIM-ARG as WOUNDVICTIM-ARG, which occurs due to the context of the number of victims not being conveyed in the text.

As a direction for further research, several aspects can be improved. First, cross-lingual transfer is an important direction. Although the SRL framework proposed in this study was designed for Indonesian, it is technically transferable to other languages, either typologically similar ones such as Malay, or more widely used English. However, this adaptation process is insufficient, requiring simply translating the annotation guide. Several important steps are required, including: (i) compiling a new corpus relevant to the crisis event domain in the target language, taking into account content filtering and geolocation constraints; (ii) re-annotating a strategically selected subset of the data using the same SRL labeling scheme, to establish ground truth in the target language; and (iii) fine-tuning or further training an appropriate pre-trained language model (e.g. MalayBERT or RoBERTa for English) to maintain semantic role classification accuracy. Further research should systematically evaluate across languages—starting from structurally similar languages such as Malay, to languages with different characteristics such as English—to measure performance changes, identify specific linguistic challenges (such as differences in morphology or use of idiomatic expressions), and assess the extent to which the SRL framework can be generalized across languages.

The second line of future work concerns data scalability. Since the dataset in this study was constrained by Twitter API restrictions, future research could investigate alternative or complementary data sources, such as news articles, governmental disaster reports, or other social media platforms (e.g., Facebook or Instagram). Combining multi-platform

data streams alleviates dependency on a single API and enriches the dataset's coverage, enabling more representative and scalable disaster-related corpora. Beyond textual sources, multimodal integration with images or videos shared during disaster events could also be explored. Such integration has the potential to provide complementary contextual cues (e.g., visual evidence of damage or affected areas), thereby enhancing both the richness and reliability of semantic role extraction for disaster analysis.

Third, one potential direction for future work is to enhance the model's ability to resolve lexical ambiguity, particularly for labels such as PLACE-ARG and AFFECTEDOBJECTS-ARG, which often overlap semantically in informal crisis-related texts. To address this, external linguistic and geographical resources could be integrated during pre-processing or model fine-tuning. For example, incorporating gazetteers (lists of known place names) can help disambiguate location references. At the same time, part-of-speech (POS) tags may aid in differentiating between objects and locations based on syntactic roles. Such lexical enrichment is expected to improve the model's ability to distinguish between semantically similar roles, especially in low-resource or noisy settings.

Beyond enhancing model performance, further research can explore how information extracted through SRL can be utilized to assess disaster severity. Our model successfully extracts key semantic roles, such as DEATHVICTIM-ARG, WOUNDVICTIM-ARG, and AFFECTEDOBJECTS-ARG, which provide critical details about fatalities, injuries, and affected assets in a crisis event. Disaster severity can be evaluated using fuzzy logic, where the number of deaths (DEATHVICTIM-ARG), injuries (WOUNDVICTIM-ARG), and estimated material losses (AFFECTEDOBJECTS-ARG) serve as the main variables in the analysis. The estimated material losses can be calculated from SRL-based information extraction by identifying objects labeled with the semantic role AFFECTEDOBJECTS-ARG and multiplying them by the estimated damage cost associated with each affected object. By implementing this system, SRL extraction improves semantic role understanding in text and enables automated disaster severity assessment. This system can assist stakeholders in making faster and more informed emergency response decisions, leading to more effective disaster management.

REFERENCES

[1] H. Senaratne, M. Mühlbauer, S. Götzer, T. Riedlinger, and H. Taubenböck, "Detecting crisis events from unstructured text data using signal words as crisis determinants," *Int. J. Digit. Earth*, vol. 16, no. 2, pp. 4601–4620, Dec. 2023, doi: 10.1080/17538947.2023.2278714.

[2] Y. Ma, F. Chen, J. Liu, Y. He, J. Duan, and X. Li, "An automatic procedure for early disaster change mapping based on optical remote sensing," *Remote Sens.*, vol. 8, no. 4, p. 272, Mar. 2016, doi: 10.3390/rs8040272.

[3] S. V. Razavi-Termeh, M. Seo, A. Sadeghi-Niaraki, and S.-M. Choi, "Flash flood detection and susceptibility mapping in the monsoon period by integration of optical and radar satellite imagery using an improvement of a sequential ensemble algorithm," *Weather Climate Extremes*, vol. 41, Sep. 2023, Art. no. 100595, doi: 10.1016/j.wace.2023.100595.

[4] M. Akhoondzadeh and D. Marchetti, "Developing a fuzzy inference system based on multi-sensor data to predict powerful earthquake parameters," *Remote Sens.*, vol. 14, no. 13, p. 3203, Jul. 2022, doi: 10.3390/rs14133203.

[5] M. Imran, F. Ofli, D. Caragea, and A. Torralba, "Using AI and social media multimodal content for disaster response and management: Opportunities, challenges, and future directions," *Inf. Process. Manage.*, vol. 57, no. 5, Sep. 2020, Art. no. 102261, doi: 10.1016/j.ipm.2020.102261.

[6] D. Istanbulluoglu, "Complaint handling on social media: The impact of multiple response times on consumer satisfaction," *Comput. Hum. Behav.*, vol. 74, pp. 72–82, Sep. 2017, doi: 10.1016/j.chb.2017.04.016.

[7] J. A. Wahid, L. Shi, Y. Gao, B. Yang, L. Wei, Y. Tao, S. Hussain, M. Ayoub, and I. Yagoub, "Topic2Labels: A framework to annotate and classify the social media data through LDA topics and deep learning models for crisis response," *Expert Syst. Appl.*, vol. 195, Jun. 2022, Art. no. 116562, doi: 10.1016/j.eswa.2022.116562.

[8] A. D. P. Ariyanto, D. Purwitasari, and C. Fatichah, "A systematic review on semantic role labeling for information extraction in low-resource data," *IEEE Access*, vol. 12, pp. 57917–57946, 2024, doi: 10.1109/ACCESS.2024.3392370.

[9] A. D. P. Ariyanto, C. Fatichah, and D. Purwitasari, "Semantic role labeling for information extraction on Indonesian texts: A literature review," in *Proc. Int. Seminar Intell. Technol. Its Appl. (ISITIA)*, Jul. 2023, pp. 119–124, doi: 10.1109/isitia59021.2023.10221008.

[10] A. Karami, V. Shah, R. Vaezi, and A. Bansal, "Twitter speaks: A case of national disaster situational awareness," *J. Inf. Sci.*, vol. 46, no. 3, pp. 313–324, Jun. 2020, doi: 10.1177/0165551519828620.

[11] L. Huang, G. Liu, T. Chen, H. Yuan, P. Shi, and Y. Miao, "Similarity-based emergency event detection in social media," *J. Saf. Sci. Resilience*, vol. 2, no. 1, pp. 11–19, Mar. 2021, doi: 10.1016/j.jnlssr.2020.11.003.

[12] P. Y. Win Myint, S. L. Lo, and Y. Zhang, "Unveiling the dynamics of crisis events: Sentiment and emotion analysis via multi-task learning with attention mechanism and subject-based intent prediction," *Inf. Process. Manage.*, vol. 61, no. 4, Jul. 2024, Art. no. 103695, doi: 10.1016/j.ipm.2024.103695.

[13] E. Blomeier, S. Schmidt, and B. Resch, "Drowning in the information flood: Machine-learning-based relevance classification of flood-related tweets for disaster management," *Information*, vol. 15, no. 3, p. 149, Mar. 2024, doi: 10.3390/info15030149.

[14] A. Munir, A. Aved, and E. Blasch, "Situational awareness: Techniques, challenges, and prospects," *AI*, vol. 3, no. 1, pp. 55–77, Jan. 2022, doi: 10.3390/ai3010005.

[15] B. M. Hailu, Y. Assabie, and Y. B. Sinshaw, "Semantic role labeling for amharic text using multiple embeddings and deep neural network," *IEEE Access*, vol. 11, pp. 33274–33295, 2023, doi: 10.1109/ACCESS.2023.3263147.

[16] W. J. Hou and B. Ceesay, "Domain transformation on biological event extraction by learning methods," *J. Biomed. Informat.*, vol. 95, Jul. 2019, Art. no. 103236, doi: 10.1016/j.jbi.2019.103236.

[17] K. Xu, H. Wu, L. Song, H. Zhang, L. Song, and D. Yu, "Conversational semantic role labeling," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 29, pp. 2465–2475, 2021, doi: 10.1109/TASLP.2021.3074014.

[18] Q. Zhou, Z. Jiang, and F. Yang, "Sentences similarity based on deep structured semantic model and semantic role labeling," in *Proc. Int. Conf. Asian Lang. Process. (IALP)*, Dec. 2020, pp. 40–44, doi: 10.1109/IALP51396.2020.9310496.

[19] M. Narouei, H. Takabi, and R. Nielsen, "Automatic extraction of access control policies from natural language documents," *IEEE Trans. Dependable Secure Comput.*, vol. 17, no. 3, pp. 506–517, May 2020, doi: 10.1109/TDSC.2018.2818708.

[20] S. Gunasekara, D. Chathura, C. Jeewantha, and G. Dias, "Using annotation projection for semantic role labeling of low-resourced language: Sinhala," in *Proc. Int. Conf. Asian Lang. Process. (IALP)*, Dec. 2020, pp. 98–103, doi: 10.1109/IALP51396.2020.9310468.

[21] D. Purwitasari, A. F. Abdillah, S. Juanita, I. K. E. Purnama, and M. H. Purnomo, "A comparison of transformer and BiLSTM based BioNER model with self-training on low-resource language texts of online health consultatio," *Int. J. Intell. Eng. Syst.*, vol. 16, no. 6, pp. 213–224, 2023, doi: 10.22266/ijies2023.1231.18.

[22] C. Wen, T. Chen, X. Jia, and J. Zhu, "Medical named entity recognition from un-labelled medical records based on pre-trained language models and domain dictionary," *Data Intell.*, vol. 3, no. 3, pp. 402–417, Sep. 2021, doi: 10.1162/dint_a_00105.

[23] C. Barros, E. Lloret, E. Saquete, and B. Navarro-Colorado, "NATSUM: Narrative abstractive summarization through cross-document timeline generation," *Inf. Process. Manage.*, vol. 56, no. 5, pp. 1775–1793, Sep. 2019, doi: 10.1016/j.ipm.2019.02.010.

[24] K. Kitazawa and S. A. Hale, "Social media and early warning systems for natural disasters: A case study of Typhoon Etau in Japan," *Int. J. Disaster Risk Reduction*, vol. 52, Jan. 2021, Art. no. 101926, doi: 10.1016/j.ijdrr.2020.101926.

[25] S. R. Chowdhury, S. Basu, and U. Maulik, "A survey on event and subevent detection from microblog data towards crisis management," *Int. J. Data Sci. Anal.*, vol. 14, no. 4, pp. 319–349, Oct. 2022, doi: 10.1007/s41060-022-00335-y.

[26] L. Belcastro, F. Marozzo, D. Talia, P. Trunfio, F. Branda, T. Palpanas, and M. Imran, "Using social media for sub-event detection during disasters," *J. Big Data*, vol. 8, no. 1, pp. 1–14, Dec. 2021, doi: 10.1186/s40537-021-00467-1.

[27] Q. Xia, C.-H. Yeh, and X.-Y. Chen, "A deep bidirectional highway long short-term memory network approach to Chinese semantic role labeling," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2019, pp. 1–6. [Online]. Available: http://www.ieee.org/publications

[28] M. Wang, "Rule-based semantic role labeling of bei-sentences," in *Proc. Int. Conf. Asian Lang. Process. (IALP)*, Dec. 2021, pp. 178–182, doi: 10.1109/IALP54817.2021.9675157.

[29] G. A. Ruz, P. A. Henríquez, and A. Mascareño, "Sentiment analysis of Twitter data during critical events through Bayesian networks classifiers," *Future Gener. Comput. Syst.*, vol. 106, pp. 92–104, May 2020, doi: 10.1016/j.future.2020.01.005.

[30] V. K. Neppalli, C. Caragea, A. Squicciarini, A. Tapia, and S. Stehle, "Sentiment analysis during hurricane sandy in emergency response," *Int. J. Disaster Risk Reduction*, vol. 21, pp. 213–222, Mar. 2017, doi: 10.1016/j.ijdrr.2016.12.011.

[31] D. Wu and Y. Cui, "Disaster early warning and damage assessment analysis using social media data and geo-location information," *Decis. Support Syst.*, vol. 111, pp. 48–59, Jul. 2018, doi: 10.1016/j.dss.2018.04.005.

[32] B. Zhou, L. Zou, A. Mostafavi, B. Lin, M. Yang, N. Gharaibeh, H. Cai, J. Abedin, and D. Mandal, "VictimFinder: Harvesting rescue requests in disaster response from social media with BERT," *Comput., Environ. Urban Syst.*, vol. 95, Jul. 2022, Art. no. 101824, doi: 10.1016/j.compenvurbsys.2022.101824.

[33] A. Bhoi, S. P. Pujari, and R. C. Balabantaray, "A deep learning-based social media text analysis framework for disaster resource management," *Social Netw. Anal. Mining*, vol. 10, no. 1, pp. 1–14, Dec. 2020, doi: 10.1007/s13278-020-00692-1.

[34] A. Upadhyay, Y. K. Meena, and G. S. Chauhan, "SatCoBiLSTM: Self-attention based hybrid deep learning framework for crisis event detection in social media," *Expert Syst. Appl.*, vol. 249, Sep. 2024, Art. no. 123604, doi: 10.1016/j.eswa.2024.123604.

[35] J. Rashid, S. M. A. Shah, and A. Irtaza, "Fuzzy topic modeling approach for text mining over short text," *Inf. Process. Manage.*, vol. 56, no. 6, Nov. 2019, Art. no. 102060, doi: 10.1016/j.ipm.2019.102060.

[36] L. Huang, P. Shi, H. Zhu, and T. Chen, "Early detection of emergency events from social media: A new text clustering approach," *Natural Hazards*, vol. 111, no. 1, pp. 851–875, Mar. 2022, doi: 10.1007/s11069-021-05081-1.

[37] D. Fernández-González, "Transition-based semantic role labeling with pointer networks," *Knowl.-Based Syst.*, vol. 260, Jan. 2023, Art. no. 110127, doi: 10.1016/j.knosys.2022.110127.

[38] I. Lauriola, A. Lavelli, and F. Aiolli, "An introduction to deep learning in natural language processing: Models, techniques, and tools," *Neurocomputing*, vol. 470, pp. 443–456, Jan. 2022, doi: 10.1016/j.neucom.2021.05.103.

[39] S. Alzahrani and H. Aljuaid, "Identifying cross-lingual plagiarism using rich semantic features and deep neural networks: A study on Arabic-English plagiarism cases," *J. King Saud Univ.-Comput. Inf. Sci.*, vol. 34, no. 4, pp. 1110–1123, Apr. 2022, doi: 10.1016/j.jksuci.2020.04.009.

[40] Y. Yang, Z. Wu, Y. Yang, S. Lian, F. Guo, and Z. Wang, "A survey of information extraction based on deep learning," *Appl. Sci.*, vol. 12, no. 19, p. 9691, Sep. 2022, doi: 10.3390/app12199691.

[41] S. Lazemi, H. Ebrahimpour-Komleh, and N. Noroozi, "Improving Persian dependency-based semantic role labeling using semantic and structural relations," in *Proc. 4th Int. Conf. Pattern Recognit. Image Anal. (IPRIA)*, Mar. 2019, pp. 163–167.

[42] M. Bombieri, M. Rospocher, S. P. Ponzetto, and P. Fiorini, "Machine understanding surgical actions from intervention procedure textbooks," *Comput. Biol. Med.*, vol. 152, Jan. 2023, Art. no. 106415, doi: 10.1016/j.compbiomed.2022.106415.

[43] N. Sabah and A. Sil, "A comprehensive report on the 28th September 2018 Indonesian Tsunami along with its causes," *Natural Hazards Res.*, vol. 3, no. 3, pp. 474–486, Sep. 2023, doi: 10.1016/j.nhres.2023.06.003.

[44] S. Heo, S. Park, and D. K. Lee, "Multi-hazard exposure mapping under climate crisis using random forest algorithm for the Kalimantan Islands, Indonesia," *Sci. Rep.*, vol. 13, no. 1, pp. 1–13, Aug. 2023, doi: 10.1038/s41598-023-40106-8.

[45] Y. Umniyatun, M. I. Nurmansyah, Y. Farradika, T. B. Purnama, and D. N. Hidayat, "Motorcycle risky behaviours and road accidents among adolescents in Jakarta metropolitan area, Indonesia," *Int. J. Injury Control Saf. Promotion*, vol. 28, no. 3, pp. 339–346, Jul. 2021, doi: 10.1080/17457300.2021.1928229.

[46] O. Toporkov and R. Agerri, "On the role of morphological information for contextual lemmatization," *Comput. Linguistics*, vol. 50, no. 1, pp. 157–191, Mar. 2024, doi: 10.1162/coli_a_00497.

[47] S. Verma, S. Vieweg, W. Corvey, L. Palen, J. Martin, M. Palmer, A. Schram, and K. Anderson, "Natural language processing to the rescue? extracting 'situational awareness' tweets during mass emergency," in *Proc. Int. AAAI Conf. Web Social Media*, 2021, vol. 5, no. 1, pp. 385–392, doi: 10.1609/icwsm.v5i1.14119.

[48] F. deBoer, "Evaluating the comparability of two measures of lexical diversity," *System*, vol. 47, pp. 139–145, Dec. 2014, doi: 10.1016/j.system.2014.10.008.

[49] A. D. P. Ariyanto, D. Purwitasari, B. Amaliah, C. Fatichah, M. G. Taqiuddin, and Haikal, "Annotated data for semantic role labeling of crisis events in Indonesian tweets," *Data Brief*, vol. 61, Aug. 2025, Art. no. 111688, doi: 10.1016/j.dib.2025.111688.

[50] M. T. Uliniansyah, I. Budi, E. Nurfadhilah, D. I. N. Afra, A. Santosa, A. D. Latief, A. Jarin, Gunarso, M. A. Jiwanggi, N. N. Hidayati, R. Fajri, R. R. Suryono, S. Pebiana, S. Shaleha, T. W. Ramdhani, and T. Sampurno, "Twitter dataset on public sentiments towards biodiversity policy in Indonesia," *Data Brief*, vol. 52, Feb. 2024, Art. no. 109890, doi: 10.1016/j.dib.2023.109890.

[51] S. Thara and P. Poornachandran, "Transformer based language identification for malayalam-english code-mixed text," *IEEE Access*, vol. 9, pp. 118837–118850, 2021, doi: 10.1109/ACCESS.2021.3104106.

[52] F. Koto, A. Rahimi, J. Lau, and T. Baldwin, "IndoLEM and IndoBERT: A benchmark dataset and pre-trained language model for Indonesian NLP," in *Proc. 28th Int. Conf. Comput. Linguistics*, 2020, pp. 757–770, doi: 10.18653/v1/2020.coling-main.66.

[53] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A robustly optimized BERT pretraining approach," 2019, *arXiv:1907.11692*.

[54] A. Gasparetto, M. Marcuzzo, A. Zangari, and A. Albarelli, "A survey on text classification algorithms: From text to predictions," *Information*, vol. 13, no. 2, p. 83, Feb. 2022, doi: 10.3390/info13020083.

[55] I. Botunac, M. B. Bakarić, and M. Matetić, "Comparing fine-tuning and prompt engineering for multi-class classification in hospitality review analysis," *Appl. Sci.*, vol. 14, no. 14, p. 6254, Jul. 2024, doi: 10.3390/app14146254.

[56] Y. Chang, X. Wang, J. Wang, Y. Wu, L. Yang, K. Zhu, H. Chen, X. Yi, C. Wang, Y. Wang, W. Ye, Y. Zhang, Y. Chang, P. S. Yu, Q. Yang, and X. Xie, "A survey on evaluation of large language models," *ACM Trans. Intell. Syst. Technol.*, vol. 15, no. 3, pp. 1–45, Jun. 2024, doi: 10.1145/3641289.

[57] H. Zhao, H. Chen, F. Yang, N. Liu, H. Deng, H. Cai, S. Wang, D. Yin, and M. Du, "Explainability for large language models: A survey," *ACM Trans. Intell. Syst. Technol.*, vol. 15, no. 2, pp. 1–38, Apr. 2024, doi: 10.1145/3639372.

[58] L. Owen, V. Tripathi, A. Kumar, and B. Ahmed, "Komodo: A linguistic expedition into Indonesia's regional languages," 2024, *arXiv:2403.09362*.

[59] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer, "QLORA: Efficient finetuning of quantized LLMs," in *Proc. 37th Int. Conf. Neural Inf. Process. Syst.*, Red Hook, NY, USA, 2023, pp. 1–28.

[60] F. M. Plaza-del-Arco, M. D. Molina-González, L. A. Ureña-López, and M. T. Martín-Valdivia, "Comparing pre-trained language models for Spanish hate speech detection," *Expert Syst. Appl.*, vol. 166, Mar. 2021, Art. no. 114120, doi: 10.1016/j.eswa.2020.114120.

[61] A. Dewandaru, "Event geoparsing Indonesian news dataset," *IEEE Dataport*, May 2020, doi: 10.21227/s5rh-rn19.
[62] H. He. *HanLP: Han Language Processing*. Accessed: Jul. 6, 2025. [Online]. Available: https://hanlp.hankcs.com/en/demos/srl.html

AMELIA DEVI PUTRI ARIYANTO received the master's degree in informatics from the Institut Teknologi Sepuluh Nopember (ITS), Surabaya, Indonesia, in 2022, where she is currently pursuing the Ph.D. degree. Additionally, she is a Lecturer with Widya Husada Semarang University, Semarang, Indonesia. She has published several papers in computer science in various journals and conferences. Her research interests include artificial intelligence, natural language processing, and data mining.

DIANA PURWITASARI (Senior Member, IEEE) received the Ph.D. degree in computer science and engineering from the Multimedia Computing Laboratory, Department of Electrical Engineering, Faculty of Intelligent Electrical and Informatics Technology (ELECTICS), Institut Teknologi Sepuluh Nopember (ITS), Surabaya, Indonesia, in 2020.

She is currently a Professor with the Department of Informatics, ELECTICS, ITS. Her research interests include information retrieval, social network analysis, computational intelligence, and web mining. Since March 2020, she has been actively participating in several professional societies, including the design task force for AI Toward Indonesia Vision 2045, the University Center of Excellence on Artificial Intelligence for Healthcare and Society (UCE AIHeS) in Indonesia, the Industrial Electronics Society (IES) Chapter of the Indonesia Section, and Indonesia AI Research Consortium under the Ministry of Research and Technology.

CHASTINE FATICHAH (Member, IEEE) received the Ph.D. degree from Tokyo Institute of Technology, Japan, in 2012. She is currently a Professor and the Director of the Undergraduate and Postgraduate Education at the Sepuluh Nopember Institute of Technology, Surabaya, Indonesia. Her research interests include artificial intelligence, image processing, and data mining.

SRI DEVI RAVANA is an Associate Professor at Universiti Malaya and Deputy Head of the Malaya Artificial Intelligence Centre of Research at Universiti Malaya. She has published widely, led funded projects, and actively fosters innovation through academic, industry, and community collaborations. Her research focuses on information retrieval, knowledge management, and artificial intelligence.

ANDRIAN is currently pursuing the bachelor's degree with the Department of Informatics Engineering, Faculty of Intelligent Electrical and Informatics Technology (ELECTICS), Institut Teknologi Sepuluh Nopember (ITS), Surabaya, Indonesia. He actively contributes as a Teaching Assistant with the Intelligent Computing and Vision Laboratory, assisting in academic and research activities. His research interests include artificial intelligence, text mining, and automation processes.

ANAK AGUNG YATESTHA PARWATA received the bachelor's degree in computer science from the Department of Informatics Engineering, Institut Teknologi Sepuluh Nopember (ITS), Surabaya, Indonesia, in 2024. His research interests include social network analysis and natural language processing.