VOL. 10. NO. 4 MAY 2025
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480 /jitk.v10i4.6194

JITK (JURNAL ILMU PENGETAHUAN DAN TEKNOLOGI KOMPUTER)

FOREST FIRE LOCATION AND TIME RECOGNITION IN SOCIAL MEDIA TEXT USING XLM-ROBERTA

Hafidz Sanjaya[1]; Kusrini Kusrini[2*]; Kumara Ari Yuana[3]; Arief Setyanto[4]; I Made Artha Agastya[5]; Simone Martin Marotta[6]; José Ramón Martínez Salio[7]

Magister of Informatics[1,2,3,4,5]
Universitas AMIKOM Yogyakarta, Indonesia[1,2,3,4,5]
https://amikom.ac.id[1,2,3,4,5]
hafidzsanjaya@students.amikom.ac.id[1], kusrini@amikom.ac.id[2*], kumara.a@amikom.ac.id[3], arief_s@amikom.ac.id[4], artha.agastya@amikom.ac.id[5]

Expert AI, Napoly, Italy[6]
https://www.expert.ai[6]
smarotta@expert.ai[6]

Eviden, Madrid, Spain[7]
https://eviden.com[7]
jose.martinezs@eviden.com[7]

(*) Corresponding Author
(Responsible for the Quality of Paper Content)

The creation is distributed under the Creative Commons Attribution-NonCommercial 4.0 International License.

Abstract—*Forest fires have become a serious global threat, significantly impacting ecosystems, communities, and economies. Although remote sensing technology shows potential, limitations such as time delays, limited sensor coverage, and low resolution reduce its effectiveness for real-time forest fire detection. Additionally, social media can serve as a multimodal sensor, presenting multilingual text data with rapid and global coverage. However, it may encounter challenges in obtaining location and time information on forest fires due to limitations in datasets and model generalization. This study aims to develop a multilingual named entity recognition (NER) model to identify location and time entities of forest fires in social media texts such as tweets. Utilizing a transfer learning approach with the XLM-RoBERTa architecture, fine-tuning was performed using the general-purpose Nergrit corpus dataset containing 19 entities, which were relabeled into 3 main entities to detect location, date, and time entities from tweets. This approach significantly improves the model's ability to generalize to disaster domains across multiple languages and noisy social media texts. With a fine-tuning accuracy of 98.58% and a maximum validation accuracy of 96.50%, the model offers a novel capability for disaster management agencies to detect forest fires in a scalable, globally inclusive manner, enhancing disaster response and mitigation efforts.*

Keywords: *entity recognition, forest fire, social media text, xlm-roberta.*

Intisari—*Kebakaran hutan telah menjadi ancaman global yang serius, memberikan dampak signifikan pada ekosistem, masyarakat, dan ekonomi. Meskipun teknologi penginderaan jauh memiliki potensi, keterbatasan seperti keterlambatan waktu, cakupan sensor yang terbatas, dan resolusi rendah mengurangi efektivitasnya untuk deteksi kebakaran hutan secara real-time. Media sosial sebagai sensor dapat menjadi solusi multimodal yang menyajikan data teks multibahasa yang memiliki cakupan global dan cepat tetapi dapat menghadapi masalah dalam mendapatkan informasi lokasi dan waktu kebakaran hutan karena keterbatasan dataset dan generalisasi model. Penelitian ini bertujuan untuk mengembangkan suatu model named entity recognition (NER) multibahasa untuk mengenali informasi lokasi dan waktu kebakaran hutan pada teks media sosial seperti tweets. Dengan memanfaatkan pendekatan transfer learning pada arsitektur XLM-RoBERTa, fine-tuning dilakukan menggunakan suatu dataset serbaguna Nergrit corpus yang memiliki 19 entitas, kemudian*

dilabeli ulang menjadi 3 entitas utama untuk mendeteksi entitas lokasi, tanggal dan waktu dari tweets. Pendekatan ini secara signifikan meningkatkan kemampuan model untuk melakukan generalisasi ke domain bencana pada berbagai bahasa dan teks media sosial yang bising. Dengan akurasi fine-tuning sebesar 98,58% dan akurasi validasi tertinggi sebesar 96,50%, model ini memberikan kemampuan baru bagi badan pengelola bencana untuk deteksi kebakaran hutan yang skalabel dan mencakup global untuk meningkatkan respons dan mitigasi bencana.

***Kata Kunci**: pengenalan entitas, kebakaran hutan, teks media sosial, xlm-roberta.*

INTRODUCTION

Recent evidence reveals that forest fires have become a global problem that has worsened in recent years, significantly impacted forest ecosystems, and contributed to climate change [1]. Efficient forest management is critical because forest fires pose a significant threat to all creatures in the region [2]. The most recent data indicates that fires caused a total forest loss of 0.69 million hectares in 2023 across several countries including Indonesia, Spain, Italy, Greece, Slovak, and others [3].

Disaster management efforts for forest fires are highly urgent, given the increasing prevalence of this issue and its negative impacts, which can occur anywhere in the world. Disaster management's preparedness, detection, emergency response, and mitigation stages widely utilize remote sensing technology [4,5,6]. The efficacy of remote sensing data is significantly hampered by the inherent limitations in temporal, geographic, and spectral resolution of earth observation sensors. This shortcoming becomes especially critical in applications that require swift action, as timely and precise information is essential for effective early detection. Addressing these constraints is vital to enhancing responsiveness and decision-making in critical situations [7].

To overcome these constraints, recent research has enhanced remote sensing data by integrating social media data, leveraging big data analytics, and applying it across multiple stages of disaster management, including preparedness, detection, emergency response, mitigation, and recovery [8]. However, the implementation of these approaches is often limited to specific domains, and their multilingual generalization remains a significant challenge.In the big data era, a new paradigm has emerged where humans act as adaptive, cost-effective sensors, sharing their insights about the natural world through social media, a practice known as social media sensing [9].

Several studies have explored the application of social media sensing in managing forest fire disasters, particularly for preparedness [10], and early detection [11]. Twitter (X) is a social media platform for delivering textual warnings and information pinpointing the precise location and time of the incident by leveraging big data [12]. This method is effective, but obtaining location and time can be difficult when the information is only in text-based data. Named entity recognition is a possible task to extract information from text-based data [13]. However, named entity recognition tasks using natural language processing approaches are sometimes limited in their ability to be applied only to certain language cases [14]. Generally, this task is successful for languages with extensive and high-quality datasets. Creating high-quality datasets also requires a lot of resources and effort. Meanwhile, using the same model in different languages can save time and resources [15].

Several studies have applied named entity recognition (NER) models for social media sensing. Koshy et al. [16] proposed RoBERTa-BiLSTM-CRF-attention model, demonstrating robust performance in identifying locations in Twitter datasets across various disaster types (flood, hurricane, earthquake, explosion) and regions with an average F1-score of 0.935. Eligüzel et al. [17] implemented Recurrent Neural Network (RNN) models, including Long Short-Term Memory (LSTM), Gated Recurrent Unit (GRU), and Bidirectional LSTM (BiLSTM), for entity recognition of locations, persons, and organizations in earthquake-related tweets, achieving an accuracy of 92%.

Experiments with activation functions such as Sigmoid, SoftPlus, and SoftMax demonstrated improved performance. Suwaileh et al. [18] implemented transfer learning for Location Mention Recognition (LMR) on Twitter using a multilingual BERT-based model, achieving an F1-score above 0.7 even in zero-shot scenarios. In another study, Berragan et al. [19] conducted a comparative study evaluating BERT [20], DistilBERT [21], RoBERTa [22], BiLSTM-CRF models [23], Stanza and SpaCy for place name extraction in UK-based Wikipedia data. The F1-score obtained by BERT was 0.939, DistilBERT was 0.924, RoBERTa was 0.923, CRF-BiLSTM was 0.859, Stanza was 0.730 and SpaCy was 0.554. The comparative results show that BERT outperforms

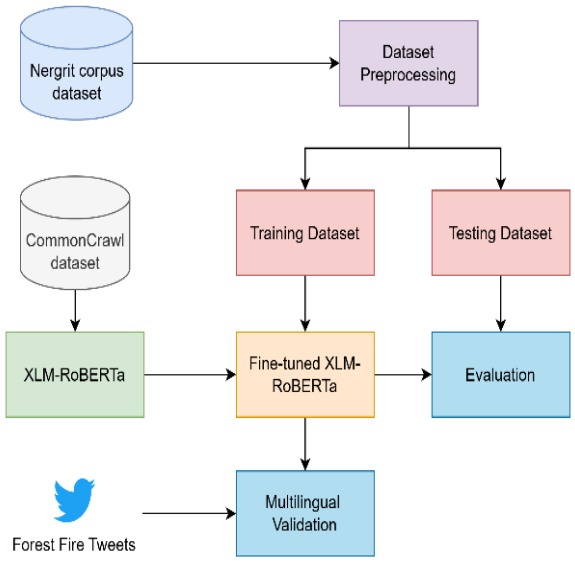
Accredited Rank 2 (Sinta 2) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti No.225/E/KPT/2022, December 07, 2022. Published by LPPM Universitas Nusa Mandiri

the other models. Girsang et al. [24] implemented the StanfordNER prebuilt model for the extraction of six location classes in Indonesia such as province, district/city, sub-district, village, road, and place names for natural disaster mapping from tweet data. The experimental results showed that StanfordNER achieved an F1-score of 0.8565.

Although previous research has shown the potential of machine learning in implementing named entity recognition, there are still limitations in cross-language generalization that can be improved. To address these limitations, this study seeks to answer how can XLM-RoBERTa with transfer learning efficiently improve the accuracy of location and time (including date) entity recognition in multilingual social media texts, what challenges in existing NER models can be addressed through the integration of pre-trained multilingual datasets, how well does the proposed model generalize across languages, particularly for low-resource languages, in location and time entity recognition scenarios and how can improvements in multilingual NER capabilities optimize social media sensing for forest fire disaster management. To answer these questions, the study develops a location and time entity recognition model using XLM-RoBERTa with transfer learning, leveraging a diverse pre-trained dataset of over 100 languages. Multilingual validation is conducted to evaluate the model's cross-language generalization capabilities using social media texts such as tweets.

MATERIALS AND METHODS

The research method in this study can be seen in Figure 1.



Source: (Research Results, 2025)

Figure 1. Research Flowchart

The entire workflow is illustrated in Figure 1, based on figure 1, the research method in this study consists of acquiring the general-purpose dataset Nergrit corpus from the hugging face website, where common tasks such as tokenization, stopword removal, and stemming are no longer performed on the dataset. Furthermore, dataset preprocessing is performed to improve the focus and relevance of the dataset for location and time recognition scenarios. The next step includes the development and training of the research model built on the XLM-RoBERTa architecture. Finally, concluding the model performance results including training evaluation and multilingual validation.

A. Dataset Acquisition

The dataset used in this study is the general-purpose dataset Nergrit corpus, obtained from the Hugging face website. This dataset is primarily collected from various general news websites, ensuring that it reflects the real-world news context [25]. By leveraging a well-documented dataset sourced from several news sites, this dataset provides a high-quality and realistic data source that reflects what is typically used in real-world events, which is critical in social media sensing that leverages text-based data. In addition, it can streamline subsequent processes efficiently.

This dataset uses the B-I-O (Beginning-Inside-Outside) annotation scheme, a widely used method in entity recognition tasks, which classifies parts of a sentence into a particular entity or non-entity. In this scheme, "B-" marks the beginning of an entity, "I-" indicates continuation within an entity, and "O" indicates words that do not fall under a particular entity label. Examples of data in the Nergrit corpus dataset are presented in Table 1.

Table 1. Sample data in the dataset

No	Tokens	Tags
1	['Kebakaran', 'hutan', 'di', 'Desa', 'Karangpatihan', ',', 'Kecamatan', 'Balong']	[O, O, O, B-GPE, I-GPE, O, B-GPE, I-GPE]
2	['Rabu' '16' 'Mei' '2018' ',' '09' ':' '36' 'WIB']	[B-DAT, I-DAT, I-DAT, I-DAT, O, B-TIM, I-TIM, I-TIM, I-TIM]
3	['Kebakaran', 'Rumah', 'di', 'Senen', 'Jakpus', ',', '15', 'Unit', 'Damkar', 'Dikerahkan']	[O, O, O, B-LOC, B-GPE, O, B-CRD, O, O, O]

Source : (Nergrit Corpus [25], 2021)

This dataset is divided into distinct train, validation, and test sets to streamline subsequent processing steps. It contains 12,532 training data, 2,399 validation data, and 2,521 testing data. Furthermore, the dataset includes 19 entities, as

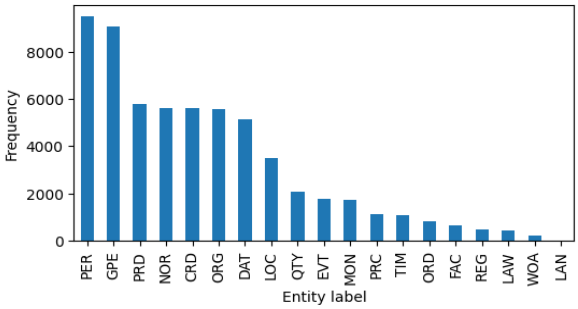
shown in Table 2. Each entity represents a unique category of information commonly found in text, enabling the system to accurately recognize and classify diverse data types. With these 19 distinct entities, the dataset offers both flexibility and precision, making it versatile and highly effective for various NLP applications. Thus, entities like location (non-administrative place names), geopolitical entity (administrative area), date, and time play a significant role in this study in detecting and tracking forest fires.

Table 2. List of entities in the dataset

No	Entity	Label	Description
1	Cardinal	CRD	Represents a number
2	Date	DAT	Date format (day, month, year)
3	Event	EVT	Name of an event
4	Facility	FAC	Name of facility
5	Geopolitical Entity	GPE	Administrative area
6	Law Entity	LAW	Legal document
7	Location	LOC	Non-administrative location (non-GPE), mountain, forest, park, etc.
8	Money	MON	Amount of money with currency
9	Political Organization	POL	Name of a political organization
10	Ordinal	ORD	Indicates position in an order
11	Organization	ORG	Company, agency, etc.
12	Person	PER	Person's name
13	Percent	PRC	Percentage (including %)
14	Product	PRD	Vehicle, weapon, food, etc.
15	Quantity	QTY	A number with a unit
16	Religion	REG	Religion name
17	Time	TIM	Hour, minute, second
18	Work of Art	WOA	An artistic work
19	Language	LAN	Language

Source : (Nergrit Corpus [25], 2021)

The frequency of each label is shown in the Figure 2. The most frequent label is “Person” followed by “Geopolitical Entity”.



Source : (Research Results, 2024)

Figure 2. Entity distribution in the dataset

B. Dataset Preprocessing

In preparing the dataset for our specific task, a targeted preprocessing step was undertaken to refine and adapt the data effectively. A key task is data relabeling to enhance the dataset's relevance and focus. To achieve this, we are relabeling the original 19 entities in the dataset into three main entities: location, date, and time.

To create the location label, geopolitical entities (GPE) and location entities (LOC) were merged into a unified category. This decision was driven by the observation that both GPE and LOC represent critical location-based information but are often indistinguishable in social media text when tracking forest fire events. For instance, a place like "*Kota Palangkaraya*" might be labeled as either GPE or LOC in different contexts, but both convey essential spatial information for the task. By unifying these entities into the single location (LOC) label, we simplified the label set, and made it easier for the model to recognize location mentions consistently. In addition to location labels, date (DAT) and time (TIM) labels were preserved without modification, as they are vital for constructing a temporal understanding of forest fire events. These labels can serve as an alternative way by allowing models to detect when an incident occurred, a critical component of situational awareness. An illustration of dataset preprocessing on sample data is shown in Table 3, where the original entity label in the dataset (red) is changed to the new entity label (blue).

Table 3. Illustration of dataset preprocessing

No	Tokens	Tags
1	['Kebakaran', 'hutan', 'di', 'Desa', 'Karangpatihan', ',', 'Kecamatan', 'Balong']	[O, O, O, B-GPE→B-LOC, I-GPE→I-LOC, O, B-GPE→B-LOC, I-GPE→I-LOC]
2	['Rabu' '16' 'Mei' '2018' ',' '09' ':' '36' 'WIB']	[B-DAT, I-DAT, I-DAT, I-DAT, O, B-TIM, I-TIM, I-TIM, I-TIM]
3	['Kebakaran', 'Rumah', 'di', 'Senen', 'Jakpus', ',', '15', 'Unit', 'Damkar', 'Dikerahkan']	[O, O, O, B-LOC, B-GPE→B-LOC, O, B-CRD→O, O, O, O]

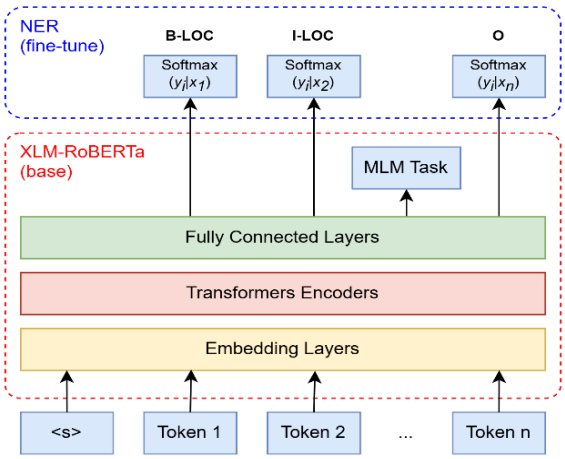
Source : (Research Results, 2024)

C. XLM-RoBERTa and Transfer Learning Approach

XLM-RoBERTa (XLM-R) is a multilingual language model proposed by Facebook [26]. This model is built using the RoBERTa architecture and trained using Masked Language Modeling (MLM) on a massive multilingual dataset covering over 100 languages. This robust training makes XLM-R particularly effective in cross-language generalization and zero-shot classification tasks. XLM-R architecture consists of embedding layers to represent input tokens, transformer encoders to understand the context between tokens using the self-attention mechanism, and fully connected layers to produce the final representation [22].

Accredited Rank 2 (Sinta 2) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti No.225/E/KPT/2022, December 07, 2022. Published by LPPM Universitas Nusa Mandiri

VOL. 10. NO. 4 MAY 2025
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480 /jitk.v10i4.6194
JITK (JURNAL ILMU PENGETAHUAN DAN TEKNOLOGI KOMPUTER)



Source : (Research Results, 2024)
Figure 3. Fine-tuning of XLM-R model architecture

In this study, transfer learning is used for fine-tuning XLM-R. Fine-tuning is the process of adapting a pre-trained model for a specific task [27]. The fine-tuned architecture is shown in the Figure 3. In this process, the model is retrained on labeled data relevant to the task, adjusting weights to recognize task-specific patterns and contexts. Rather than freezing the pre-trained layers, model parameters are fine-tuned alongside the new task-specific layers. This end-to-end fine-tuning allows model to adjust all of its weights for the specific task, leveraging the knowledge from pre-training to improve performance. The model's final layer is often modified to suit the desired output (shown in the blue dashed rectangle for the fine-tuned layer and the red dashed rectangle for the pre-trained model). In NER, the model output is configured as a sequence of labels $y = \{y_1, y_2, \dots, y_n\}$ for each token, with a softmax function applied to generate the probability of each label for each token in a sentence. The softmax function is shown in Equation (3).

$$P(y_i|x_i) = \frac{\exp(z_i)}{\sum_j \exp(z_j)} \quad (3)$$

where $P(y_i|x_i)$ is the probability of assigning label y_i to token x_i, z_i represents the raw score (logit) for token x_i for class y_i, and $\sum_j \exp(z_j)$ serves as the normalization factor.

D. Multilingual Validation

Multilingual validation is conducted to evaluate the cross-language generalization capabilities of the fine-tuned model on social media texts, such as tweets. Specifically, the model's ability to predict location, date and time labels is tested on 2,326 forest fire-related tweets in five languages: Indonesian, English, Spanish, Italian, and Slovak. These tweets were collected using crawling techniques over a period from August 2019 to November 2023. This validation process involves determining whether the predicted labels for each token are correct, providing insight into the robustness and effectiveness of the model's generalization across diverse linguistic and cultural contexts. Examples of tweets in each language for multilingual validation are presented in Table 4.

Table 4. Multilingual tweet samples for validation

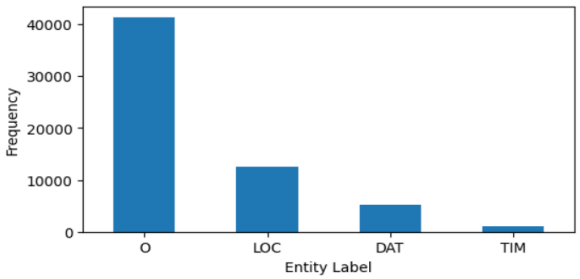
No	Tweets	Language
1	*Kebakaran hutan dan lahan terjadi di Taman Nasional Sebangau, Palangka Raya, Kalimantan Tengah, Rabu, 01 November 2023 Pukul 13.00 WIB*	*Indonesian*
2	*Wildfire rages for a second day in Evia destroying a Natura 2000 protected pine forest. – 5:51 PM Aug 14, 2019*	*English*
3	*3 nov 2023 21:57 - Incendio forestal obliga a la evacuación de hasta 850 personas cerca del pueblo de Montichelvo en Valencia.*	*Spanish*
4	*Allerta incendi in Sicilia, Schifani: "Richiamati in servizio 1.600 operai forestali e dichiarato lo stato di crisi"*	*Italian*
5	*Lesné požiare na Sicílii si vyžiadali dva ľudské životy an evakuáciu hotela – 23. septembra 2023 20:57*	*Slovak*

Source : (Research Results, 2024)

RESULTS AND DISCUSSION

This study uses the following main experimental setup: The transfer learning framework for fine-tuning the model is Transformers version 4.35.0, PyTorch version 2.1.0, optimizing with CUDA version 118. The software environment runs on Python 3.10, operating on Ubuntu 22.04 with an NVIDIA Titan V GPU (12 GB of memory) and a 9th Gen Intel Processor 9900K.

In this study, the crucial step in developing a multilingual model for recognizing location, date and time entities from social media text is dataset preprocessing to improve the focus and relevance of the task. The results of dataset preprocessing show that the distribution of entities in the dataset has changed. Figure 4 shows the distribution of entities after dataset preprocessing.



Source : (Research Results, 2024)
Figure 4. Entity distribution after preprocessing

To provide a clearer illustration of the preprocessing results, Table 5 presents sample data consisting of tokens and their corresponding entity labels that have undergone relabeling.

Table 5. Sample data from preprocessing results

No	Tokens	Tags
1	['Kebakaran', 'hutan', 'di', 'Desa', 'Karangpatihan', ',', 'Kecamatan', 'Balong']	[O, O, O, B-LOC, I-LOC, O, B-LOC, I-LOC]
2	['Rabu' '16' 'Mei' '2018' ',' '09' ':' '36' 'WIB']	[B-DAT, I-DAT, I-DAT, I-DAT, O, B-TIM, I-TIM, I-TIM, I-TIM]
3	['Kebakaran', 'Rumah', 'di', 'Senen', 'Jakpus', ',', '15', 'Unit', 'Damkar', 'Dikerahkan']	[O, O, O, B-LOC, B-LOC, O, O, O, O, O]

Source : (Research Results, 2024)

The next crucial stage is modeling. In this study, experiments were conducted using transfer learning techniques on the XLM-R and multilingual BERT (mBERT) models used in [18] to comprehensively evaluate the model comparison on cross-language generalization. However, during fine-tuning, the mBERT model used the same parameters as XLM-R. The chosen fine-tuning parameters were selected to strike a balance between computational efficiency and model performance. A batch size of 8 was chosen to address GPU memory limitations while maintaining the model's generalization ability. A learning rate of 0.00005 was used, as it is a commonly recommended value for fine-tuning large models [20], allowing for careful weight updates without disrupting the pre-trained representations. The Adam optimizer was selected for its effectiveness in handling noisy gradients and ensuring efficient convergence. With 3 epochs, this number is sufficient to achieve convergence without the risk of overfitting, allowing the model to learn effectively from the new data while retaining strong generalization capabilities. The performance comparison of the fine-tuned models can be seen in Table 6.

Table 6. Performance comparison of fine-tuned models in percent (%)

Model	Acc.	Precision	Recall	F1-score
XLM-R	**98.58**	**91.89**	**92.72**	**92.30**
mBERT-cased	98.51	90.68	92.01	91.34
mBERT-uncased	98.38	90.22	91.89	91.05

Source : (Research Results, 2024)

Table 6 highlights the performance of XLM-R, mBERT-cased, and mBERT-uncased with transfer learning technique. All three models exhibit high accuracy, precision, recall, and F1-score, indicating that this configuration of epochs and batch size is effective for achieving strong model performance. XLM-R demonstrates the highest performance, with an accuracy of 98.58%, precision of 91.89%, recall of 92.72%, and F1-score of 92.30%. The final stage of this research involves multilingual validation on social media texts such as tweets. The experiment was conducted to evaluate the comparison of model's cross-language generalization results in real-world scenarios. The results of location and time entity recognition in forest fire tweets are shown in Table 7.

Table 7. Tweet extraction results using XLM-R

No	Tweets	Results
1	*Kebakaran hutan dan lahan terjadi di Taman Nasional Sebangau, Palangka Raya, Kalimantan Tengah, Rabu, 01 November 2023 Pukul 13.00 WIB*	*[LOC Taman Nasional Sebangau, Palangka Raya, Kalimantan Tengah], [DAT Rabu, 01 November 2023], [TIM 13.00 WIB]*
2	*Wildfire rages for a second day in Evia destroying a Natura 2000 protected pine forest. – 5:51 PM Aug 14, 2019*	*[LOC Evia], [DAT Aug 14, 2019], [TIM 5:51 PM]*
3	*3 nov 2023 21:57 - Incendio forestal obliga a la evacuación de hasta 850 personas cerca del pueblo de Montichelvo en Valencia.*	*[LOC Montichelvo, Valencia], [DAT 3 nov 2023], [TIM 21:57]*
4	*Allerta incendi in Sicilia, Schifani: ''Richiamati in servizio 1.600 operai forestali e dichiarato lo stato di crisi''*	*[LOC Sicilia]*
5	*Lesné požiare na Sicílii si vyžiadali dva ľudské životy an evakuáciu hotela – 23. septembra 2023 20:57*	*[LOC Sicílii], [DAT 23. septembra 2023], [TIM 20:57]*

Source : (Research Results, 2024)

The results of correct location and time entity recognition are then compared with the number of tweets in each language to obtain model accuracy. The comparison of model accuracy for each language is shown in Table 8.

Table 8. Accuracy comparison of multilingual validation in percent (%)

Language	Model		
	XLM-R	mBERT-cased	mBERT-uncased
Indonesian	**92.32**	71.16	76.36
English	**73.97**	50.10	70.06
Spanish	**77.45**	59.80	75.10
Italian	**78.39**	65.82	75.64
Slovak	**96.50**	75.80	91.40

Source : (Research Results, 2024)

Table 8 shows that the fine-tuned XLM-R model outperformed other models. The XLM-R model achieved the highest accuracy in Indonesian, English, Spanish, Italian, and Slovak, with 92.32%, 73.97%, 77.45%, 78.39%, and 96.50%, respectively.

Accredited Rank 2 (Sinta 2) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti No.225/E/KPT/2022, December 07, 2022. Published by LPPM Universitas Nusa Mandiri

The accuracy comparison of mBERT-cased, mBERT-uncased, and XLM-R on forest fire tweets in Indonesian, English, Spanish, Italian, and Slovak provides critical insights into how these models handle noisy data from social media. For languages where tweet noise levels are high due to abbreviations, and inconsistent grammar, the results reveal that mBERT-cased struggles significantly. In English, mBERT-cased achieves only 50.10% accuracy, reflecting its difficulty with capitalization-sensitive models in noisy data. mBERT-uncased improves in these environments, reaching 70.06% accuracy in English and showing a more robust performance in Spanish and Italian with accuracy of 77.45% and 75.64%, where case sensitivity is less critical. However, XLM-R, pre-trained on a much larger and diverse, multilingual dataset, consistently outperforms the other models, especially in languages with lower noise, such as Indonesian and Slovak, with an accuracy of 92.32% and 96.50%, where it demonstrates superior robustness to linguistic variation.

The pre-training datasets play a pivotal role in the performance of these models across various languages. XLM-R benefits from its extensive pre-training on a wide range of multilingual corpora, allowing it to better generalize across languages and handle the inherent noise in tweets, especially in languages like Indonesian and Slovak. Its ability to adapt to informal social media text, even in high-noise languages like English and Spanish, underscores the value of its diverse pre-training data. In contrast, mBERT-uncased shows better adaptability in noisy environments, particularly in Spanish and Italian, by discarding case distinctions.

Despite achieving strong performance, the proposed approach still has limitations. Fine-tuning on a general-purpose dataset may introduce bias when applied to social media texts, affecting their ability to generalize and capture sentence context, which may result in increased prediction errors when applied to real-world applications. Furthermore, the models may struggle with format differences, as the highly noisy language of social media is often different from formal texts. Future research can explore more diverse datasets and determine appropriate preprocessing techniques to better assess generalization across different languages, contexts, and formats. These findings contribute valuable insights for future research and practical implementation in forest fire monitoring and mitigation globally.

Error Analysis

Referring to the results, we examine the validation results of tweets to highlight the strengths and weaknesses of the proposed model in each language. The correct values are shown in underlined text and the predicted results are shown in bold text.

1) **Indonesian tweet**: "*... Kebakaran Hutan Dan Lahan(Karhutla) DiDesa Banjar Sari Kec.Enggano Kab.Bkl Utara*". The following are the prediction results for each model:
   a) **XLM-R**: "*... Kebakaran Hutan Dan Lahan(Karhutla) Di[LOC **Desa Banjar Sari Kec.Enggano Kab.Bkl Utara**]*"
   b) **mBERT-cased**: "*... Kebakaran Hutan Dan Lahan(Karhutla) DiDesa Banjar Sari Kec.Enggano Kab.Bkl Utara*"
   c) **mBERT-uncased**: "*... kebakaran hutan dan lahan(karhutla) didesa banjar sari kec.enggano kab.bkl utara*"

   The observation results show that the XLM-R model is able to detect location references in Indonesian tweets accurately compared to the other two models, even though there are abbreviations and inconsistent grammar.

2) **English tweet**: "*... forest fire in north carolina*". The following are the prediction results for each model:
   a) **XLM-R**: "*... forest fire in [LOC **north carolina**]*"
   b) **mBERT-cased**: "*... forest fire in [LOC **north**] carolina*"
   c) **mBERT-uncased**: "*... forest fire in [LOC **north carolina**]*"

   XLM-R and mBERT-uncased are able to detect consistently with the reference location, without segmentation or formatting errors. In contrast, mBERT-cased shows inconsistency by only detecting the phrase "north" reflecting its difficulty with capitalization-sensitive models in noisy data.

3) **Spanish tweet**: "*#URGENTE #CHILE Se declara ALERTA ROJA para la comuna de Isla de Pascua por incendio forestal*". The following are the prediction results for each model:
   a) **XLM-R**: "*#URGENTE #[LOC **CHILE**] Se declara ALERTA ROJA para la comuna de **Isla de Pascua** por incendio forestal*"
   b) **mBERT-cased**: "*#[LOC **URGENTE**] #[LOC **CHILE**] Se declara ALERTA ROJA para la comuna de [LOC **Isla de Pascua**] por incendio forestal*"
   c) **mBERT-uncased**: "*#urgente #[LOC **chile**] se declara alerta roja para la comuna de [LOC **isla de pascua**] por incendio forestal*"

   XLM-R provides predictions that are completely consistent with the reference locations while maintaining the original capitalization and

structure, mBERT-uncased is also able to detect location but the entire text is converted to lowercase. Meanwhile, mBERT-cased shows an error by detecting the word "URGENTE" as a location, thus creating an inconsistency.

4) **Italian tweet**: "*En directe: MINUT A MINUT | Incendi forestal de Montitxelvo*". The following are the prediction results for each model:
   a) **XLM-R**: "*En directe: MINUT A MINUT | Incendi forestal de [LOC **Montitxelvo**]*"
   b) **mBERT-cased**: "*En directe: [LOC **MINUT**] A [LOC **MINUT**] | Incendi forestal de [LOC **Montitxelvo**]*"
   c) **mBERT-uncased**: "*en directe: minut a minut | incendi forestal de [LOC **montitxelvo**]*"
   d) XLM-R and mBERT-uncased correctly detect the reference location, although mBERT-uncased converts all of its text to lowercase. Meanwhile, mBERT-cased shows an error by detecting the word "MINUT" as a location, thus creating an inconsistency.
   e) **Slovak tweet**: "*Grécko stále sužujú lesné požiare - 27. augusta 2023 14:44*". The following are the prediction results for each model:
   f) **XLM-R**: "*[LOC **Grécko**] stále sužujú lesné požiare - [DAT **27. augusta 2023**] [TIM **14:44**]*"
   g) **mBERT-cased**: ""*[LOC **G**]récko stále sužujú lesné požiare - [DAT **27. augusta 2023**] [TIM **14:44**]*"
   h) **mBERT-uncased**: ""*[LOC **grecko**] stále sužujú lesné požiare - [DAT **27. augusta 2023**] [TIM **14:44**]*"

In Slovak tweets, XLM-R has successfully detected location and time correctly and consistently as truth values, where mBERT-uncased has also detected such references but the model converts all text to lowercase and normalizes diacritics (accents), for example the letter "é" to "e" in the word "Grécko". As a result, the meaning or identity of the word can change, especially in the context of languages that use diacritics to distinguish words or names semantically. Meanwhile, mBERT-cased has difficulty detecting locations in languages that use diacritics.

CONCLUSION

This study successfully demonstrates the development of a multilingual named entity recognition (NER) model using XLM-RoBERTa and a transfer learning approach to identify critical entities such as location, date, and time in social media texts. By leveraging transfer learning, the model effectively utilizes pre-trained knowledge from a diverse dataset covering over 100 languages, enabling accurate generalization across languages and noisy social media data. The model achieves an impressive fine-tuning accuracy of 98.59%, precision of 91.89%, recall of 92.72%, and F1 score of 92.30%. Multilingual validation across languages such as Indonesian, English, Spanish, Italian, and Slovak achieve respective accuracy scores of 92.32%, 73.97%, 77.45%, 78.39%, and 96.50%, demonstrating the model's robust cross-lingual capabilities for recognizing location and time entities. The results highlight the potential of XLM-RoBERTa for enhancing real-time multilingual disaster response systems. Further, this study provides a scalable and globally relevant solution that complements existing remote sensing technologies, enabling a hybrid approach for forest fire detection. Future research can explore integrating advanced transfer learning models, leveraging annotated multilingual datasets, and augmenting data to further improve performance.

ACKNOWLEDGMENT

This research was conducted as part of the SILVANUS Project, funded by the European Commission under the Horizon 2020 program, call number H2020-LC-GD-2020, project number 101037247. Additionally, this research is supported by funding from the Ministry of Education, Culture, Research, and Technology of the Republic of Indonesia, under the master's thesis research grant, number 107/E5/PG.02.00.PL/2024.

REFERENCE

[1] A. Tyukavina *et al.*, "Global Trends of Forest Loss Due to Fire From 2001 to 2019," *Frontiers in Remote Sensing*, vol. 3, 2022, doi: 10.3389/frsen.2022.825190.
[2] A. Setyanto *et al.*, "Ecological Impact Assessment Framework for Areas Affected by Natural Disasters," in *Proceedings of the 19th International Conference on Content-Based Multimedia Indexing*, in CBMI '22. New York, NY, USA: Association for Computing Machinery, 2022, pp. 155–161. doi: 10.1145/3549555.3549596.
[3] M. Weisse, L. Goldman, and S. Carter, "Tropical Forest Loss Drops Steeply in Brazil and Colombia, but High Rates Persist Overall," Apr. 2024. [Online]. Available:

https://research.wri.org/gfr/latest-analysis-deforestation-trends

[4] Z. Zheng *et al.*, "Wildfire Detection Based on the Spatiotemporal and Spectral Features of Himawari-8 Data," *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

[5] K. A. Yuana *et al.*, "GIS data support technique for forest fire management and decision support system: A Sebangau National Park, Kalimantan case," in *2023 6th International Conference on Information and Communications Technology (ICOIACT)*, 2023, pp. 286–291. doi: 10.1109/ICOIACT59844.2023.10455935.

[6] J. Zulkarnain, M. R. Pahlevi, Y. Astica, W. Pangestuti, and Kusrini, "Fire Detection based on Smoke Image using Convolutional Neural Network (CNN)," in *2022 4th International Conference on Cybernetics and Intelligent System (ICORIS)*, 2022, pp. 1–5. doi: 10.1109/ICORIS56080.2022.10031410.

[7] A. D. Laksito *et al.*, "Machine Learning and Social Media Harvesting for Wildfire Prevention," in *2023 IEEE 13th International Conference on Pattern Recognition Systems (ICPRS)*, 2023, pp. 1–6. doi: 10.1109/ICPRS58416.2023.10179001.

[8] J. Phengsuwan *et al.*, "Use of Social Media Data in Disaster Management: A Survey," *Future Internet*, vol. 13, no. 2, 2021, doi: 10.3390/fi13020046.

[9] X. X. Zhu *et al.*, "Geoinformation harvesting from social media data: A community remote sensing approach," *IEEE Geosci Remote Sens Mag*, vol. 10, no. 4, pp. 150–180, 2023.

[10] L. Li, Z. Ma, and T. Cao, "Data-driven investigations of using social media to aid evacuations amid Western United States wildfire season," *Fire Saf J*, vol. 126, p. 103480, Dec. 2021, doi: 10.1016/J.FIRESAF.2021.103480.

[11] J. Phengsuwan *et al.*, "Use of social media data in disaster management: a survey," *Future Internet*, vol. 13, no. 2, p. 46, 2021.

[12] S. A. Shah, S. Ben Yahia, K. McBride, A. Jamil, and D. Draheim, "Twitter Streaming Data Analytics for Disaster Alerts," in *2021 2nd International Informatics and Software Engineering Conference (IISEC)*, 2021, pp. 1–6. doi: 10.1109/IISEC54230.2021.9672370.

[13] H. Sanjaya, K. Kusrini, K. A. Yuana, and J. R. M. Salio, "Multilingual Named Entity Recognition Model for Location and Time Extraction of Forest Fire," in *2024 4th International Conference of Science and Information Technology in Smart Administration (ICSINTESA)*, 2024, pp. 611–615. doi: 10.1109/ICSINTESA62455.2024.10747844.

[14] E. H. Muktafin and P. Kusrini, "Sentiments analysis of customer satisfaction in public services using K-nearest neighbors algorithm and natural language processing approach," *TELKOMNIKA (Telecommunication Computing Electronics and Control)*, vol. 19, no. 1, pp. 146–154, Feb. 2021, doi: 10.12928/TELKOMNIKA.V19I1.17417.

[15] A. Deshpande, P. Talukdar, and K. Narasimhan, "When is BERT multilingual? isolating crucial ingredients for cross-lingual transfer," *arXiv preprint arXiv:2110.14782*, 2021.

[16] R. Koshy and S. Elango, "Applying social media in emergency response: an attention-based bidirectional deep learning system for location reference recognition in disaster tweets," *Applied Intelligence*, 2024, doi: 10.1007/s10489-024-05462-6.

[17] N. Eligüzel, C. Çetinkaya, and T. Dereli, "Application of named entity recognition on tweets during earthquake disaster: a deep learning-based approach," *Soft comput*, vol. 26, no. 1, pp. 395–421, 2022, doi: 10.1007/s00500-021-06370-4.

[18] R. Suwaileh, T. Elsayed, M. Imran, and H. Sajjad, "When a disaster happens, we are ready: Location mention recognition from crisis tweets," *International Journal of Disaster Risk Reduction*, vol. 78, p. 103107, Aug. 2022, doi: 10.1016/J.IJDRR.2022.103107.

[19] C. Berragan, A. Singleton, A. Calafiore, and J. Morley, "Transformer based named entity recognition for place name extraction from unstructured text," *International Journal of Geographical Information Science*, vol. 37, no. 4, pp. 747–766, 2023, doi: 10.1080/13658816.2022.2133125.

[20] D. Rothman, *Transformers for Natural Language Processing: Build, train, and fine-tune deep neural network architectures for NLP with Python, Hugging Face, and OpenAI's GPT-3, ChatGPT, and GPT-4*. Packt Publishing Ltd, 2022.

[21] S. Akpatsa *et al.*, "Online news sentiment classification using distilbert," *Journal of Quantum Computing*, vol. 4, no. 1, p. 1, 2022.

[22] K. L. Tan, C. P. Lee, K. S. M. Anbananthen, and K. M. Lim, "RoBERTa-LSTM: a hybrid model

for sentiment analysis with transformer and recurrent neural network," *IEEE Access*, vol. 10, pp. 21517–21525, 2022.

[23] N. Deng, H. Fu, and X. Chen, "Named Entity Recognition of Traditional Chinese Medicine Patents Based on BiLSTM-CRF," *Wirel Commun Mob Comput*, vol. 2021, no. 1, p. 6696205, 2021.

[24] G. Girsang and B. Noveta, "Six classes named entity recognition for mapping location of Indonesia natural disasters from twitter data," *International Journal of Intelligent Computing and Cybernetics*, May 2024, doi: 10.1108/IJICC-09-2023-0251.

[25] PT Gria Inovasi Teknologi, "Nergrit Corpus," Huggingface. Accessed: Jan. 15, 2024. [Online]. Available: https://huggingface.co/datasets/id_nergrit_corpus

[26] A. Gaikwad, P. Belhekar, and V. Kottawar, "Advancing Multilingual Sentiment Understanding with XGBoost, SVM, and XLM-RoBERTa," in *International Conference on Data Science, Machine Learning and Applications*, 2023, pp. 990–1000.

[27] P. Azunre, *Transfer Learning for Natural Language Processing*. Manning, 2021. [Online]. Available: https://books.google.co.id/books?id=bGI7EAAAQBAJ

Accredited Rank 2 (Sinta 2) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti No.225/E/KPT/2022, December 07, 2022. Published by LPPM Universitas Nusa Mandiri