

Monitoring Indonesian online news for COVID-19 event detection using deep learning

Purnomo Husnul Khotimah¹, Andria Arisal¹, Andri Fachrur Rozie¹, Ekasari Nugraheni¹, Dianadewi Riswantini¹, Wiwin Suwarningsih¹, Devi Munandar¹, Ayu Purwarianti²

¹Research Center for Data and Information Sciences, National Research and Innovation Agency, Bandung, Indonesia

²School of Electrical Engineering and Informatics, U-CoE AI VLB, Institut Teknologi Bandung, Bandung, Indonesia

Article Info

Article history:

Received Nov 24, 2021

Revised Aug 9, 2022

Accepted Sep 10, 2022

Keywords:

BERT

Coronavirus disease 2019

Deep learning

Event detection

News monitoring

Online news

System framework

ABSTRACT

Even though coronavirus disease 2019 (COVID-19) vaccination has been done, preparedness for the possibility of the next outbreak wave is still needed with new mutations and virus variants. A near real-time surveillance system is required to provide the stakeholders, especially the public, to act in a timely response. Due to the hierarchical structure, epidemic reporting is usually slow particularly when passing jurisdictional borders. This condition could lead to time gaps for public awareness of new and emerging events of infectious diseases. Online news is a potential source for COVID-19 monitoring because it reports almost every infectious disease incident globally. However, the news does not report only about COVID-19 events, but also various information related to COVID-19 topics such as the economic impact, health tips, and others. We developed a framework for online news monitoring and applied sentence classification for news titles using deep learning to distinguish between COVID-19 events and non-event news. The classification results showed that the fine-tuned bidirectional encoder representations from transformers (BERT) trained with Bahasa Indonesia achieved the highest performance (accuracy: 95.16%, precision: 94.71%, recall: 94.32%, F1-score: 94.51%). Interestingly, our framework was able to identify news that reports the new COVID strain from the United Kingdom (UK) as an event news, 13 days before the Indonesian officials closed the border for foreigners.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Purnomo Husnul Khotimah

Research Center for Data and Information Sciences, National Research and Innovation Agency

Bandung, Indonesia

Email: hkhhotimah@mail.informatika.lipi.go.id, purn005@brin.go.id

1. INTRODUCTION

The rapid movement of people and goods globally increases the latent danger and potential spread of infectious diseases beyond international borders [1]. Coronavirus disease 2019 (COVID-19) pandemic gives us a lesson that preparedness towards emerging infectious diseases (EIDs) is urgently needed. It could be in the form of a monitoring system for early warning that allows stakeholders to conduct appropriate assessments, rapid responses, and collaborations at regional, national, and global scales [1]. In particular, the Association of Southeast Asian Nations (ASEAN) region is at risk for infectious diseases threat due to its high population density, mobility, and socio-economic development with inadequate public health services [2]. Rapid outbreak detection ensures timely reactions to minimize morbidity and mortality, as well as social and economic mitigation disruption.

The conventional system for monitoring public health depends on a hierarchical structure in which health care providers, laboratories, and health institutions report infectious diseases on the list of monitored

diseases. Local authorities forward the report to regional authorities, who then forward it to national officials who subsequently report to international agencies world health organization (WHO) [3]. Even in the digital era, where information from the official can be integrated and distributed faster, it still needs to go through several layers of bureaucracy and causes time gaps for public awareness of new and emerging events of EIDs. One of the possible solutions is to adapt informal sources, such as from social media or online news.

Online news usually reports various events globally in a timely manner. The ability of timely reporting has made the online news media potential and reliable informal data source for conducting surveillance for infectious diseases [4]. Several previous studies have been leveraging online news-stream data as an informal data source for monitoring infectious diseases [5]–[15]. Global public health intelligence network (GPHIN) is one of the earliest systems that based its EIDs surveillance system on the internet [8]. It is managed by health Canada's center for emergency preparedness and response (CEPR) [5]. GPHIN looks through the Internet and extract news sources worldwide in nine languages (Arabic, English, Farsi, French, Portuguese, Russian, Spanish, simplified Chinese, and traditional Chinese) [15]. GPHIN uses search keywords that are updated regularly to retrieve relevant articles. The articles are then screened and classified based on GPHIN taxonomies. The relevant articles (published on the GPHIN database) are detected by calculating a relevancy score. It is derived by utilizing weight attributed to terms and keywords within the taxonomy or taxonomies where they belong. Additional human analysis is required to evaluate articles whose relevancy score is between automatic publish and automatic trash.

HealthMap [7] is also harnessing data from the Internet. It is a web-based system that integrates data from various electronic sources. The sources include news from the Google News aggregator, ProMED Mail, and validated official alerts such as WHO announcements. It scans for articles in Arabic, Chinese, English, French, Portuguese, Russian and Spanish [12]. The input data is then converted into standardized alerts and categorized by their location and diseases. Further, the alerts are overlaid on an interactive geographic map [10]. For the categorization, a parser module, and a dictionary of a known pattern of place and disease patterns are developed. The parser module uses a word-level N-gram approach to match the input towards the dictionary mapping text patterns to the database.

Another EIDs surveillance system is semantic processing and integration of distributed electronic resources for epidemiology (EpiSPIDER). It is a system that maps emerging infectious disease information on Google Maps [6]. EpiSPIDER uses information from ProMed Mail, central intelligence agency (CIA) Factbook and PubMed. Natural language processing (NLP) is used to extract location and plots then on Google Map. It searches only for English articles. In a later report [10], EpiSPIDER started to outsource some of its preprocessing, including NLP tasks to external service providers such as OpenCalais and the unified medical language system (UMLS) web service for concept annotation.

Until this point, the previously mentioned studies focus on international languages online news. BioCaster is a non-governmental public health surveillance system on health hazards and used an ontology-based text mining system [9]. The event recognition automatic classification of the reports uses a conventional machine learning method, a naive Bayes algorithm. BioCaster covers Asia-Pacific languages (Chinese, Japanese, Korean, Thai, and Vietnamese). However, even though it is given the priority for the Asia-Pacific languages, BioCaster found few articles, other than Chinese [12]. This insight shows that there are still gaps in tapping online news media resources in other local languages, such as Indonesian or Bahasa Indonesia. Bahasa Indonesia is the fourth largest language used over the Internet yet it still has limited NLP resources [16], especially for infectious disease online news monitoring. Therefore, it would be beneficial to develop a deep learning-based framework for utilizing Indonesian online news to conduct surveillance towards COVID-19 events.

From the technical viewpoint, the applications mentioned above had been using a weight-based strategy (GPHIN), natural language processing (Health Map, EpiSPIDER) that includes word distribution, n-gram, parsing, and naive-Bayes (BioCaster) to conduct the classification procedure. However, with the development of machine learning, deep learning models have gained success in many fields. By using deep learning, the system can avoid the possible complexity of the classification procedure. For example, the weight-based method must determine the weight values for each word used to represent a category. Another complexity is in the iterative procedure for information extraction, which is required for categorization in the parser-based system. The deep learning approach hides the complexity of the procedure, such as the classifiers could automatically extract the determinant features and predict the results based on the data fed in the training process. Hence, a deep learning approach should provide a simpler solution for text classification to detect news that reports COVID-19 infectious disease events.

Further studies in deep learning have shown impressive results for text classification. Kim [17] showed that convolutional neural network (CNN) could achieve excellent results for sentence classification. Li *et al.* [18] used a hybrid model of CNN and long short-term memory (LSTM) that is Bi-CNN-LSTM to address the NLP problem in news text classification, and Luan and Lin [19] added a combination of CNN and LSTM in text

categorization problem. Another approach is creating a hybrid CNN, LSTM, and multi-layer perceptron (MLP) in text classification, such as in sentiment analysis [20]. Compared to conventional NLP, deep learning models can avoid the possible complexity of the classification procedure by hiding the complexity of the procedure, such as the classifiers could automatically extract the determinant features and predict the results based on the data fed in the training process. Bangyal *et al.* [21] got similar results that deep learning models, especially bidirectional LSTM (BiLSTM) and CNN, gave the best performance compared to the non-deep learning classification algorithms in the text classification task. The introduction of transformers model in NLP, such as generalized pre-trained transformer (GPT) [22], and bidirectional encoder representations from transformers (BERT) [23] improved the performance in NLP tasks because the transformer adopted the encoder-decoder architecture with an attention mechanism. Furthermore, the training parallelization in transformer architecture makes building a pre-trained model with a large dataset possible. After that, the model can be retrained and fine-tuned for specific tasks. Qasim *et al.* [24] showed that fine-tuned BERT-based transfer learning can improve text classification tasks. Thus, transformer architecture has become state of the art in natural language processing tasks. As one of the architecture implementations, BERT has been adopted and retrained in various languages datasets. For example, in Bahasa Indonesia, a newly pre-trained BERT with a large dataset in Bahasa Indonesia is called IndoBERT [16]. However, the model has yet to be fine-tuned for a specific task, such as text classification with an additional labeled dataset.

This paper presents a framework for monitoring an infectious disease using online news. In this case, we use COVID-19 as it is the latest infectious and attracts the public's attention. Sentence classification was explored to perform news titles classification related to COVID-19 events or non-events using machine learning methods (from classical to deep learning models). This paper extends the previously proposed classifiers [25] for the COVID-19 dataset. In addition, it implements BERT as the state-of-the-art language model for classification that has been trained using Bahasa Indonesia corpora (IndoBERT), which is the latest benchmark for Indonesian NLP [16].

This paper has three-fold contributions. First, it shows that a framework for monitoring COVID-19 infection event from Indonesian online news portal can be used to filter COVID-19 event news. To the best of our knowledge, our framework might be one of its kind frameworks where Indonesian online news articles are used for text classification and monitoring events especially related with infectious disease. The framework could be easily scaled up for other EIDs, such as dengue fever as shown in [25]. Further, the current framework is also possible to be cloned for other local languages given the appropriate data for the model training. Second, this paper gives a valuable insight on Indonesian text classification using conventional machine learning, deep learning and the latest language model BERT that is to acquire higher performance using IndoBERT fine tuning with the respective dataset is needed. Third, it shows the use of the framework to provide a geo-information visualization of COVID-19 related event news coverage from Indonesian online news portals.

The remainder of this paper is organized as follows. In section 2, we present the proposed framework for monitoring COVID-19 events using online news and deep learning. In section 3, we describe our experimental results and discuss the significance of our study. Finally, we conclude the paper in section 4.

2. METHOD

We build a framework for monitoring Indonesian online news as shown in Figure 1. First, online news is collected by crawling Indonesian online news portals and stored in a database. Vote-based data labeling is used in the data preparation. After the data preparation, we transform the input text into a feature matrix that will be fed to the classifier models for training. This trained classifier model is then used to label news articles into event and non-event news. The labeled data are stored in a database and REST API is developed for applications to retrieve the data. The following description provides a detailed explanation for each procedure.

2.1. Data acquisition

The news titles were acquired from seven Indonesian news portals, including the following: *Tirto*, *Tempo*, *Republika*, *Merdeka*, *Kompas*, *Detik*, and *Antara*. These news portals were selected because of their reliability, speed of news delivery and are verified by the Indonesian press council. We obtained 20,431 news titles by crawling the articles using several keywords: corona, covid, and COVID-19. The articles were collected from January 26th until May 24th, 2020. In the crawling process, we gathered the timestamp, title, contents, and metadata.

As opposed to using the entire article, we use the news title only to conduct the classification. This approach is chosen because a title should commonly summarize and represent the article's content. Smaller text size also reduces the computing cost during the model training. Preceding the data preparation, the title is extracted, and deduplication is conducted. Deduplication eliminates title duplicates that happen when online news is divided into several parts or pages.

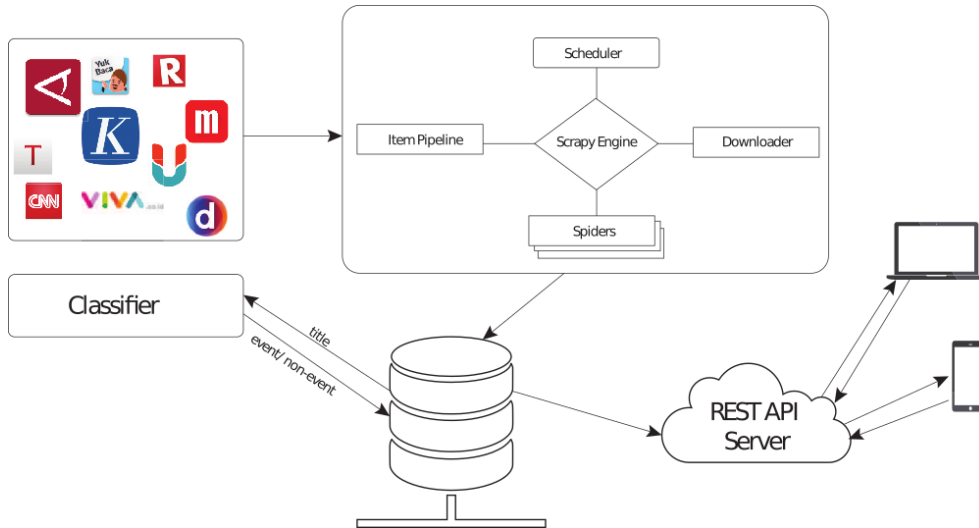


Figure 1. The system framework

2.2. Data preparation

Online news writes various information from COVID-19 infection status to economic impacts, from health tips to celebrities’ activities during the pandemic. For an example, two news titles from *Merdeka*, which is one of Indonesian news portal, are ‘*Konser amal satu Indonesia kumpulkan donasi miliaran rupiah untuk lawan COVID-19*’ (English: One Indonesia charity concert collects billions of rupiah donations to fight COVID-19) and ‘*Camat Tambora gelar tes swab massal setelah 30 warga terpapar COVID-19*’ (English: Tambora Sub-district head holds mass swab test after 30 residents were exposed to COVID-19). The first title could be considered non-COVID-19 event, while the second one is related to the COVID-19 event. Hence, in data preparation, we categorize online news titles that report COVID-19 events and those that do not relate to COVID-19 events.

The labeling process is done with a vote-based labeling approach. Three respondents performed the labeling based on the news titles criteria. The title’s label was then determined upon the most votes. We divided the classes into two categories that represent the relevant news towards the COVID-19 event and non-event news as:

- 0 (negative): non-event class—news surrounding COVID-19 information that report non-infection incidents, for example, health recommendation to avoid COVID-19, donations for COVID and others.
- +1 (positive): event class—news that report COVID-19 infection events, such as about a +/- case(s) happening or increasing/decreasing, color changes for the infected area zone, area lockdown and others.

Table 1 shows examples of news titles for each data class and Figure 2 shows the distribution of online news classes extracted from each portal. Negative and positive data numbers are 12,292 and 4,549, respectively.

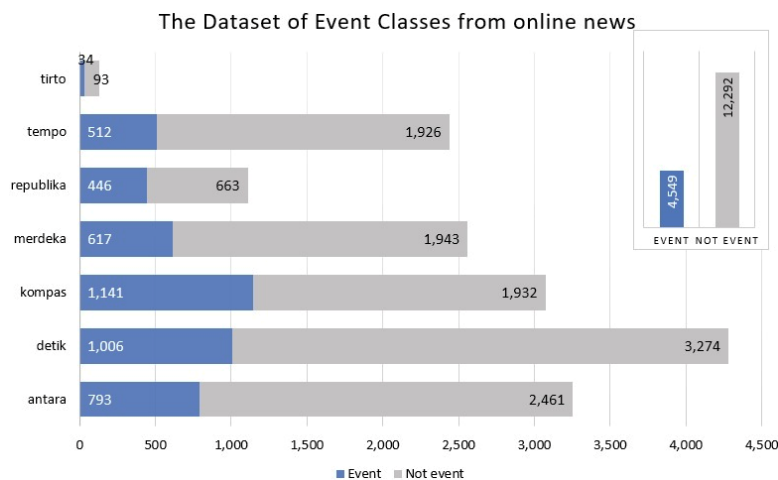


Figure 2. Online news dataset

Table 1. Examples of news titles

Label	Title
0	<i>1 juta KK terdampak COVID-19 di Jabar dapat bansos tunai, Emil apresiasi Kemensos</i> (1 million families affected COVID-19 in West Java get social cash, Emil extend appreciation to Ministry of Social Affairs)
0	<i>10 tanaman herbal bisa perkuat imunitas tubuh buat hindari virus corona COVID-19</i> (10 herbal plants can strengthen body immunity to avoid corona COVID-19 virus)
0	<i>100 negara ikut AS dorong investigasi China terkait COVID-19</i> (100 countries join the US to push for Chinese investigations related to COVID-19)
1	<i>“Lockdown” dilonggarkan, kasus corona kembali meningkat di Jerman</i> (“Lockdown” is loosened, corona cases are increasing in Germany)
1	<i>Rapid test 700 warga Kota Jambi 14 terinfeksi positif COVID-19</i> (In the rapid test, 700 residents of Jambi City 14 were positively infected with COVID-19)
1	<i>10 orang PDP yang dirawat di Rembang dinyatakan negatif corona</i> (10 suspects who were treated in Rembang were declared negative for Coronavirus)

2.3. Event detection classifiers

For the classifiers, we used three deep learning models (MLP, CNN and LSTM) and conventional machine learning (naive Bayes, logistic regression, decision tree, support vector machine (SVM), AdaBoost, neural network) as comparisons. Further, we also investigated the latest Indonesian benchmark for NLP that is IndoBert ([16], which is a BERT model ([23] pre-trained with Indonesian corpora).

Textual data features are created by transforming text into a vector of term frequency-inverse document frequency [26]. Another feature is word embedding [27], [28], which transforms text into a fixed-size dense vector representation. We use various deep learning classifiers to classify textual data in those vector representations. Those classifiers are MLP, CNN, LSTM, and BERT.

MLP is a neural-network model with its architecture containing multiple layers consists of an input layer, hidden layer(s), and an output layer. The nodes in the hidden layer(s) and the output layers operate on nonlinear activation functions. Those layers construct a feed-forward structure that is outputs from one layer become inputs for the successive layer. MLP built its model using backpropagation technique [29], [30]. Although the model’s architecture is simple, it is surprisingly able to perform with good results. The model benefits from the hierarchical network structure and its ability to learn the representation of the given data to consider every given feature in the classification problems. The architecture is shown in Figure 3.

CNN is an extension of MLP by adding a regularizer, which is a convolution layer to avoid overfitting [31]. In text analysis, a series of window filters serves to create a hierarchical and simpler pattern from a complex input pattern. At the end of the process, a classification function is given by adding a fully connected layer and output layer similar to MLP. The architecture is shown in Figure 4.

LSTM [32] is specialized recurrent neural network (RNN) model having capabilities to learn dependencies among data in the dataset. LSTM’s main component is a cell that regulate the information flow from an input gate to an output gate or a forget gate for unnecessary information. On the inter-related inputs, a repetitive process occurs that implies a temporal sequence. This process benefits for tasks such as data-related predictions and classification. The architecture is shown in Figure 5.

BERT is a transformer-based machine learning technique used for various natural language tasks. In this experiment, we utilize the model for BERT, pre-trained with Bahasa Indonesia, and retrain the model for our dataset classification for fine-tuning. The architecture is shown in Figure 6.

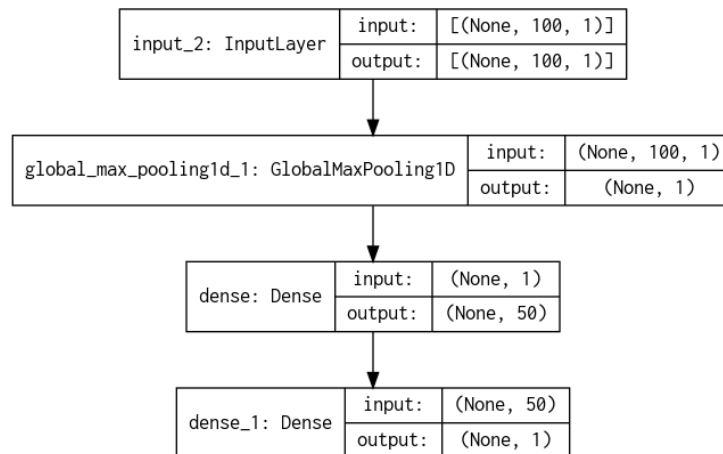


Figure 3. MLP model architecture

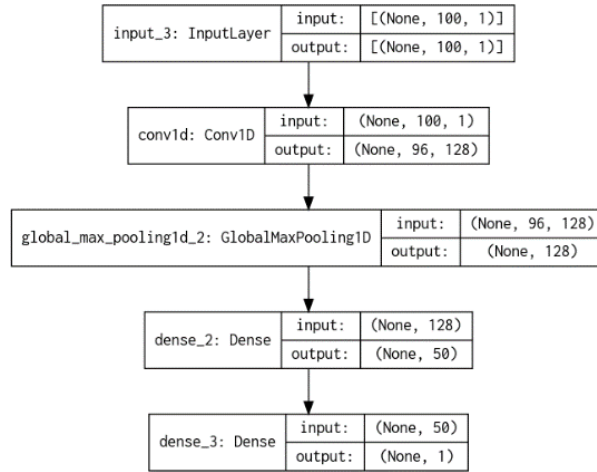


Figure 4. CNN model architecture

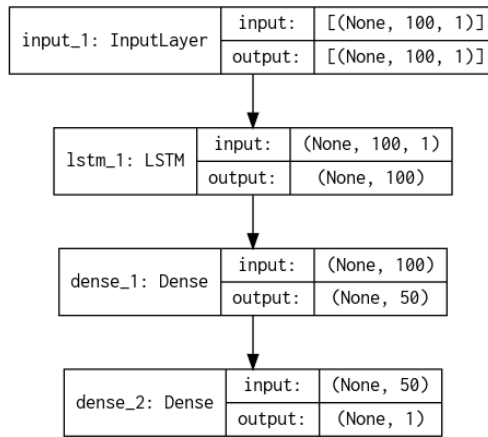


Figure 5. LSTM model architecture

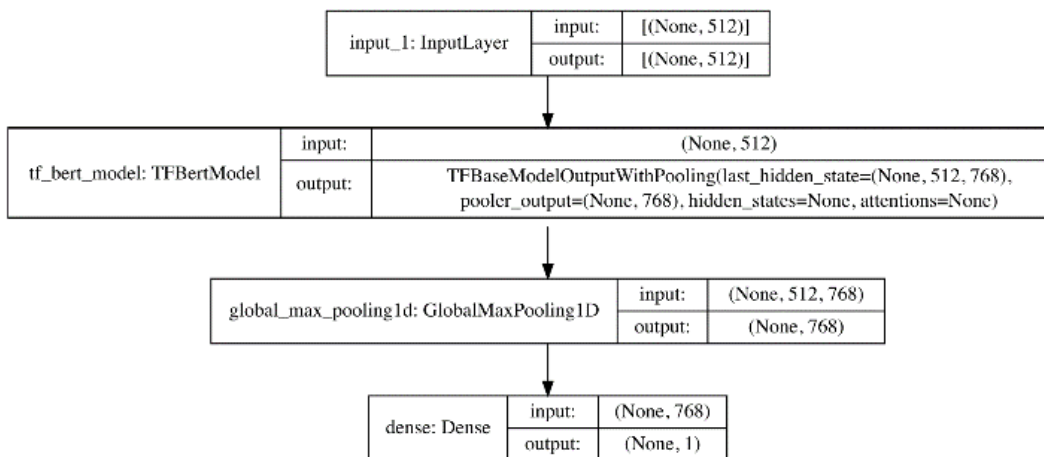


Figure 6. BERT model architecture

Within our experiment, we used term frequency/inverse document frequency (TF/IDF) as input features and a combination of sigmoid with rectified linear unit (ReLU) activation functions as well as Adam [33] as optimizer formula. For evaluation purposes, we also used conventional machine learning classifiers.

- Naive Bayes is a supervised learning algorithm that applies Bayes' theorem and the "naive" assumption of conditional independence between each pair of features given the value of the class variable. Maximum a posteriori (MAP) estimation is used to estimate the relative probability of a given class from the training set [34].
- Logistic regression is a learning algorithm based on statistical model using logistic function. The logistic function is derived from defined constraints to evaluate the probability of data to exist in the given class [35].
- Decision tree is a learning model that falls into a non-parametric supervised method category. The method learns simple decision rules inferred from the data features to create a model that predicts the value of a target variable [36].
- Support vector machine it is an algorithm that perform classification by constructing a hyperplane or set of hyperplanes in a high-dimensional space. The data separation within the hyperplane is used for classifying the data into the given classes [37].
- AdaBoost is a classifier based on meta-estimator. The method tries to fit the classifier to the original dataset to the classifier and then add weight-adjusted copies of the same classifier. The procedure is done repeatedly to give better classification results [38].
- Neural network is inspired by the biological brain. Input is represented by many features where each feature is involved in all possible inputs. Backpropagation is used to regulate the weight of the network. The weight regulation is done repeatedly to decrease the difference between actual output and desired output [29].

2.4. Mobile application implementation

As current EIDs event information is scattered on the Internet, users are required to find the information from the Internet. With current technology trend that is the data that should come to the user, our framework contains an API system to enable other applications/system developers to use the data/information generated by the framework. A mobile application named COVID-19 SIAPP is shown in Figures 7 to 10. SIAPP is an abbreviation for *sistem informasi aplikasi pemantauan dan peringatan dini* (English: information system of monitoring and early warning application). The mobile application is developed to show a prospective application of our approach.



Figure 7. Initial home screen: event and non-event news are displayed

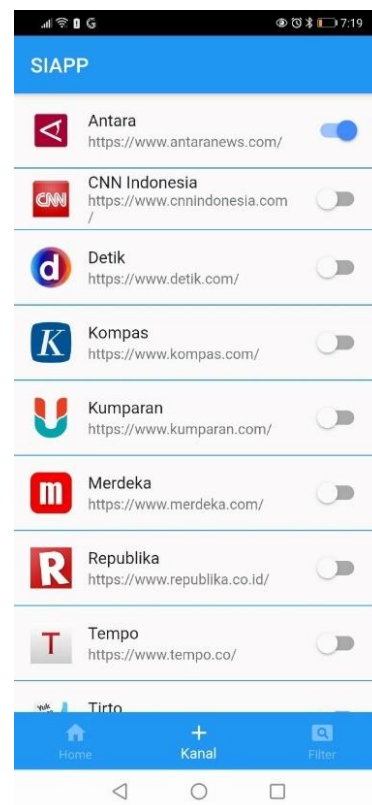


Figure 8. Media portal setting: user portal selection



Figure 9. Event and non-event filter menu

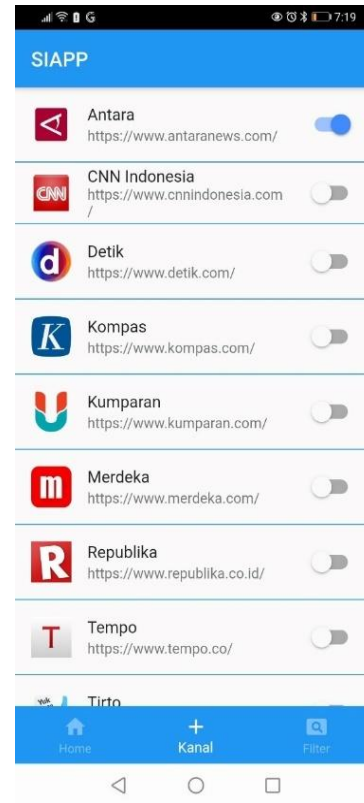


Figure 10. Home screen after event filter selected: non-event news is not displayed

Figure 7 shows the home screen for COVID-19 SIAPP application. The primary function of SIAPP is to display news collections from various portals. Currently, the news portals that can be accessed are *Antara*, *CNN Indonesia*, *Detik*, *Kompas*, *Kumparan*, *Merdeka*, *Republika*, *Tempo*, *Tirto*, and *Viva*. SIAPP COVID-19 accesses online news articles regarding COVID-19 using REST API.

As seen in Figure 8, the main display shows a list of news that are still mixed between event news (orange bounded news: “*Pasien COVID-19 di NTT bertambah 34 menjadi 1.246 orang*”, English: “COVID-19 patient in NTT increased by 34 to 1,246 people”) and non-event news (blue bounded news: “*Bisnis hotel bangkit dari keterpurukan dampak pandemi COVID-19*”, English: “The hotel business rises from the downturn due to the COVID-19 pandemic”). Users can use the filter menu in Figure 9. If the event filter menu is selected, then when the user returns to the home display, the information news (non-event news) will not be displayed. Only event news will appear in the layer as shown in Figure 10.

The application specification contains mobile/front end and backend. In the mobile/front end, the specification is IDE: Android Studio, programming language: Dart, and mobile framework: Flutter. In the backend, the specification is database: MongoDB, Elastic; communication schema: Rest API, programming language: Python, and framework: Flask

3. RESULTS AND DISCUSSION

3.1. Assessment

3.1.1. Performance

In this study, we assessed deep learning models and compared them with conventional machine learning models. Firstly, we used the TF/IDF feature on both models. Secondly, we assessed the performance of deep learning models using word embedding because word embedding could increase deep learning performance. The evaluation was done by k-fold cross-validation. To avoid the small number of test data for each fold, we used $k=5$ instead of 10. Additionally, we used early stopping based on the loss value to avoid over-fitting. Further, we also investigated the BERT model for our problem, both original and fine-tuned. The experiments are carried out using python 3 with Scikit-learn for machine learning library, Keras, TensorFlow, and Transformers for deep learning library.

3.1.2. News coverage

In addition to classifiers performance, we also looked at the news coverage area in our dataset. We used named entity recognition [39], [40], that is provided by Prosa.ai to identify geo-information from the news titles. Outputs which were labeled as geo-political entity (GPC), facilities (FAC), location (LOC), and product (PRO), then transformed into a map to visualize the news coverage. Visualization is useful to identify the areas where the events are reported in the news.

3.2. Discussion

The results of our models' performance assessment are shown in Table 2. The following insights were gathered. In conventional machine learning, logistic regression excels compared to other models (accuracy: 92.38%, precision: 92.30%, recall: 87.27%, F1-score: 89.33%), except in precision (naive Bayes achieved 92.43%).

Deep learning models (MLP, CNN, and LSTM) could not outperform logistic regression when the same feature matrix (TF/IDF) was used as the input. However, when word embedding was used in deep learning models, their performance can compete with the best classical machine learning model's performance. The improved performance is because the word embedding technique is more suitable for neural network models due to its ability to keep the order and interaction of the words within sentences and the probability functions of each word sequence. All deep learning models acquired accuracy 92%. CNN excelled compared to other deep learning models in all metrics (accuracy: 92.87%, precision: 91.30%, recall: 89.95%, F1-score: 90.57%). Additionally, CNN with word embedding outperformed logistic regression in three metrics except precision ($\delta_{\text{precision}}$: 1%). However, the dataset is not a balanced dataset, and F1-score is frequently used as an indicator in that case. In F1-score, CNN outperformed logistic regression by $\delta_{\text{F1-score}}$: 1.24%.

Using BERT, the performance dropped significantly. We believe that this low performance was caused by the model that has not been trained with online news corpora. However, fine-tuning increased the performance significantly in all four metrics compared to deep learning models with word embedding. The differences between *BERT+fine-tune* and CNN with word embedding are as follows: δ_{accuracy} : 2.29%, $\delta_{\text{precision}}$: 2.87%, δ_{recall} : 4.37%, $\delta_{\text{F1-score}}$: 3.94%.

Based on the results, we find that fine-tuned BERT for news title classification achieved the best performance in every aspect of the evaluation. As for the news area coverage, we acquired 574 unique geo-information. The geo-information consists of countries (e.g., China, German, and United State), regions (e.g., Europe, Latin America, and Southeast Asia), global cities (e.g., New York, Tokyo, and Wuhan), provinces (i.e., Indonesian provinces, such as Aceh, Bali, and DKI Jakarta), local cities (e.g., Bojonegoro, Depok, and Makassar) and facilities (e.g., hospital, market, and boarding facility). The number of each geo-information category is shown in Table 3. Various written variations that indicate the exact same location are eliminated. The number of unique geo-information is shown in the "After Deduplication" column.

Table 2. Classifiers Performance

Classification	Accuracy	Precision	Recall	F1-Score
TF/IDF+Naive Bayes	0.9134	0.9243	0.8461	0.8751
TF/IDF+Logistic Regression	0.9238	0.9230	0.8727	0.8933
TF/IDF+Decision tree	0.9012	0.8991	0.8372	0.8613
TF/IDF+SVM	0.8542	0.9033	0.7276	0.7642
TF/IDF+AdaBoost	0.9118	0.9016	0.8616	0.8784
TF/IDF+Neural Net	0.9139	0.9218	0.8488	0.8763
TF/IDF+MLP	0.7324	0.7361	0.5064	0.4379
TF/IDF+CNN	0.7476	0.6944	0.5642	0.5509
TF/IDF+LSTM	0.8242	0.8070	0.7153	0.7294
WE+MLP	0.9225	0.9030	0.8927	0.8971
WE+CNN	0.9287	0.9130	0.8995	0.9057
WE+LSTM	0.9206	0.8950	0.8972	0.8960
BERT (IndoBert)	0.3720	0.5488	0.5163	0.3301
BERT (IndoBert+fine-tune)	0.9516	0.9471	0.9432	0.9451

Table 3. Number of geo-information collected

Geoinfo	Country	Region	Global city	Province	Local city	Facilities
Unique	59	21	29	49	255	161
After Deduplication	51	18	26	36	227	142

We then mapped global cities into countries, counted the number of event-news articles found for each country and plotted the number into color codes of the global map, as shown in Figure 11. The colors in Figure 11 are graded from light yellow to red that show low to higher number. For comparison, we used the timeline of COVID-19 pandemic from Coronavirus Resource Center of Johns Hopkins University’s website at the date of May 24th, 2020, as shown in Figure 12. The colors in Figure 12 are graded from light orange to dark orange that show low to higher number. We could observe that our mapping in Figure 11 is almost similar to the global COVID-19 timeline as shown in Figure 12. Based on Figure 11, some regions such as South America, Africa, and Europe seem to have no news coverage, even though there are a small number of event-news covered within those regions. This condition is because we did not map event-news articles having geo-information of region categories (such as, Africa, Latin America, and East Europe) into countries.



Figure 11. News area coverage-world wide

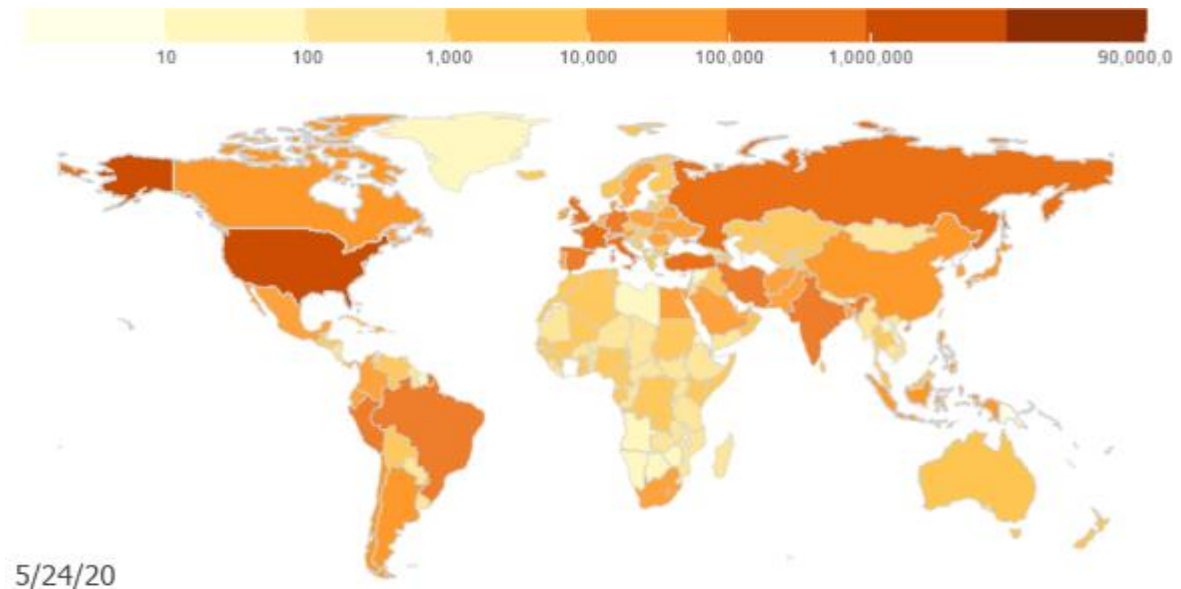


Figure 12. COVID-19 case based on the timeline of Coronavirus Resource Center of John Hopkins University

Furthermore, the results show that even though we use Indonesian online news, the news reports almost all important global COVID-19 events. On an additional note, Indonesian areas are in red color that indicates a significantly high number of events presents in the online news compared to the China area. This condition is due to the dataset used is Indonesian online news. Hence, it is expected that the news reporting COVID-19 event mainly focuses on Indonesia area.

With a similar method, we mapped local cities into provinces, counted the number of event-news articles found for each province, and plotted the number into color codes of Indonesian maps, as shown in Figure 13. In comparison, we acquired timeline data from COVID-19 Indonesia Dataset on Kaggle and plotted the data, as shown in Figure 14. The colors in Figure 13 and Figure 14 are graded from light yellow to red that show low to higher number. As shown in both figures the pattern is similar. Java Island is the hot spot (red and orange colored area) in Indonesia for the COVID-19 event. Additionally, we could observe that some provinces with a relatively high number of COVID-19 events in the online news correlate with the number of cases in those provinces. They are North and South Sumatra in Sumatra Island; North and South Sulawesi in Sulawesi; Bali and West Nusa Tenggara; also, Papua in Papua Island. However, there are exceptions in Riau and Yogyakarta provinces. In these two provinces, the number of news was significantly higher compared to the actual case number. Based on this result, we could think that there are some relevant correlations between the number of event-related news in online news and the actual number of COVID-19 cases. However, further study should be done before using the online count number for further analyses, such as for early warning purposes.



Figure 13. News area coverage-national

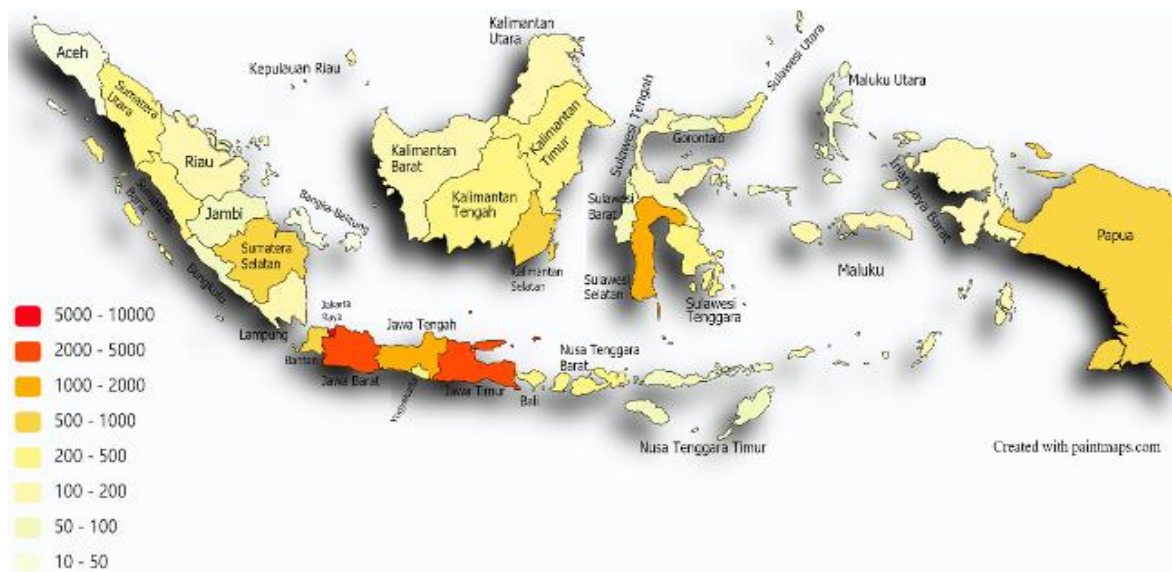


Figure 14. COVID-19 case based on COVID-19 Indonesia dataset on Kaggle

As an additional remark, we have implemented our framework and monitored the COVID-19 event until December 31, 2020. From the prototype, our framework can identify the new variant from the United Kingdom (UK). For example, the news with title '*Varian baru virus corona teridentifikasi di Inggris*' (Eng.: New coronavirus variant identified in England) published by *Merdeka* portal on December 15, 2020, is labeled as event news by our classifier. As pandemic takes longer than predicted earlier and with the new-normal beginning, the public may be unaware of the information. An online news-based monitoring application may help the public to get the information at hand.

4. CONCLUSION

This paper presented a framework for online news monitoring and classification. It is used for COVID-19 event detection from collected news titles using deep learning. In this study, we investigated various deep learning models and compared them to conventional machine learning. Common deep learning models: MLP, CNN, and LSTM, excel compared to conventional machine learning when word embedding is used, and CNN acquired the highest performance (accuracy: 92.87%, precision: 91.30%, recall: 89.95%, F1-score: 90.57%). However, fine-tuned Indonesian pre-trained BERT model gives significantly better performance in all evaluation metrics than other classification models with accuracy: 95.15%, precision: 94.71%, recall: 94.32%, and F1-score: 94.51%.

From the application viewpoint, our event detection framework is able to detect COVID-19 events from online news not only from Indonesia but also from 51 countries. One of our highlights is that our implementation is able to identify the UK strain event 13 days before the local authority announcement for international entry restriction (December 28th, 2020). With this kind of insight at the hands of the public, it could provide public awareness prior to a formal announcement and drive risk reduction towards the next wave of the pandemic.

ACKNOWLEDGEMENTS

The authors declare that there is no conflict of interest regarding the publication of this article, and the authors confirmed that the paper was free of plagiarism. PHK, AA, AFR contributed equally as the main contributor and EN, DR, WS, DM, AP contributed equally as the member contributors of this paper. All authors read and approved the final paper.




REFERENCES

- [1] W. Yang, *Early warning for infectious disease outbreak theory and practice editorial board*. Academic Press, 2017.
- [2] S. Pitsuwan, "Challenges in infection in ASEAN," *The Lancet*, vol. 377, no. 9766, pp. 619–621, Feb. 2011, doi: 10.1016/S0140-6736(10)62143-5.
- [3] K. M. Cordes *et al.*, "Real-time surveillance in emergencies using the early warning alert and response network," *Emerging Infectious Diseases*, vol. 23, no. 13, pp. 131–137, Nov. 2017, doi: 10.3201/eid2313.170446.
- [4] P. Kostkova, "A roadmap to integrated digital public health surveillance," in *Proceedings of the 22nd International Conference on World Wide Web*, May 2013, pp. 687–694, doi: 10.1145/2487788.2488024.
- [5] A. M. and M. Blench, "Global public health intelligence network (GPHIN)," in *Association for Machine Translation in the Americas*, 2006.
- [6] H. Tolentino, R. Kamadjeu, M. Matters, M. Pollack, and L. Madoff, "Scanning the emerging infectious diseases horizon-visualizing ProMED emails using EpiSPIDER," *Adv Dis Surveil*, vol. 2, 2007.
- [7] J. S. Brownstein, C. C. Freifeld, B. Y. Reis, and K. D. Mandl, "Surveillance Sans Frontières: Internet-based emerging infectious disease intelligence and the healthmap project," *PLoS Medicine*, vol. 5, no. 7, pp. 1019–1024, Jul. 2008, doi: 10.1371/journal.pmed.0050151.
- [8] C. C. Freifeld, K. D. Mandl, B. Y. Reis, and J. S. Brownstein, "HealthMap: global infectious disease monitoring through automated classification and visualization of Internet media reports," *Journal of the American Medical Informatics Association*, vol. 15, no. 2, pp. 150–157, Mar. 2008, doi: 10.1197/jamia.M2544.
- [9] N. Collier *et al.*, "BioCaster: detecting public health rumors with a web-based text mining system," *Bioinformatics*, vol. 24, no. 24, pp. 2940–2941, Oct. 2008, doi: 10.1093/bioinformatics/btn534.
- [10] M. Keller *et al.*, "Use of unstructured event-based reports for global infectious disease surveillance," *Emerging Infectious Diseases*, vol. 15, no. 5, pp. 689–695, May 2009, doi: 10.3201/eid1505.081114.
- [11] L. Madoff, "Infectious diseases surveillance and alert systems," *International Journal of Infectious Diseases*, vol. 16, Jun. 2012, doi: 10.1016/j.ijid.2012.05.113.
- [12] A. Lyon, M. Nunn, G. Grossel, and M. Burgman, "Comparison of web-based biosecurity intelligence systems: BioCaster, EpiSPIDER and HealthMap," *Transboundary and Emerging Diseases*, vol. 59, no. 3, pp. 223–232, Dec. 2012, doi: 10.1111/j.1865-1682.2011.01258.x.
- [13] Y. T. Yang, M. Horneffer, and N. DiLisio, "Mining social media and web searches for disease detection," *Journal of Public Health Research*, vol. 2, no. 1, Art. no. jphr.2013.e4, Mar. 2013, doi: 10.4081/jphr.2013.e4.
- [14] A. G. Huff, N. Breit, T. Allen, K. Whiting, and C. Kiley, "Evaluation and verification of the global rapid identification of threats system for infectious diseases in textual data sources," *Interdisciplinary Perspectives on Infectious Diseases*, vol. 2016, pp. 1–5, 2016, doi: 10.1155/2016/5080746.




- [15] M. Dion, P. AbdelMalik, and A. Mawudeku, "Big data and the global public health intelligence network (GPHIN)," *Canada Communicable Disease Report*, vol. 41, no. 9, pp. 209–214, Sep. 2015, doi: 10.14745/ccdr.v41i09a02.
- [16] B. Wilie *et al.*, "IndoNLU: benchmark and resources for evaluating Indonesian natural language understanding," in *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, 2020.
- [17] Y. Kim, "Convolutional neural networks for sentence classification," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1746–1751, doi: 10.3115/v1/D14-1181.
- [18] C. Li, G. Zhan, and Z. Li, "News text classification based on improved Bi-LSTM-CNN," in *9th International Conference on Information Technology in Medicine and Education*, Oct. 2018, pp. 890–893, doi: 10.1109/ITME.2018.00199.
- [19] Y. Luan and S. Lin, "Research on text classification based on CNN and LSTM," in *Proceedings of 2019 IEEE International Conference on Artificial Intelligence and Computer Applications*, Mar. 2019, pp. 352–355, doi: 10.1109/ICAICA.2019.8873454.
- [20] D. Munandar, A. F. Rozie, and A. Arisal, "A multi domains short message sentiment classification using hybrid neural network architecture," *Bulletin of Electrical Engineering and Informatics (BEEI)*, vol. 10, no. 4, pp. 2181–2191, Aug. 2021, doi: 10.11591/EEI.V10I4.2790.
- [21] W. H. Bangyal *et al.*, "Detection of fake news text classification on COVID-19 using deep learning approaches," *Computational and Mathematical Methods in Medicine*, pp. 1–14, Nov. 2021, doi: 10.1155/2021/5514220.
- [22] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving language understanding by generative pre-training," *Homology, Homotopy and Applications*, 2018.
- [23] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," in *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Oct. 2018, vol. 1, pp. 4171–4186, doi: 10.48550/arxiv.1810.04805.
- [24] R. Qasim, W. H. Bangyal, M. A. Alqarni, and A. Ali Almazroi, "A fine-tuned BERT-based transfer learning approach for text classification," *Journal of Healthcare Engineering*, vol. 2022, pp. 1–17, Jan. 2022, doi: 10.1155/2022/3498123.
- [25] P. H. Khotimah, A. Fachrur Rozie, E. Nugraheni, A. Arisal, W. Suwarningsih, and A. Purwarianti, "Deep learning for dengue fever event detection using online news," in *International Conference on Radar, Antenna, Microwave, Electronics and Telecommunications*, Nov. 2020, pp. 261–266, doi: 10.1109/ICRAMET51080.2020.9298630.
- [26] R. Baeza-Yates and B. Ribeiro-Neto, *Modern information retrieval: the concepts and technology behind search*, 2nd ed., Addison-Wesley Publishing Company, 2011, doi: book/10.5555/1796408.
- [27] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *1st International Conference on Learning Representations, ICLR 2013*, Aug. 2013.
- [28] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances Neural information processing systems*, 2006, vol. 1, pp. 1–9.
- [29] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, no. 6088, pp. 533–536, Oct. 1986, doi: 10.1038/323533a0.
- [30] S. K. Pal and S. Mitra, "Multilayer perceptron, fuzzy sets, classification," *IEEE Transactions on Neural Networks*, vol. 3, no. 5, pp. 683–697, 1992.
- [31] R. Collobert and J. Weston, "A unified architecture for natural language processing: Deep neural networks with multitask learning," *Proceedings of the 25th International Conference on Machine Learning*, pp. 160–167, 2008.
- [32] S. Hochreiter and J. Schmidhuber, "Long short term memory. Neural computation," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
- [33] D. P. Kingma and J. L. Ba, "ADAM: a method for stochastic optimization," in *3rd International Conference on Learning Representations*, Dec. 2014.
- [34] H. Zhang, "The optimality of Naive Bayes," *Proceedings of the Seventeenth International Florida Artificial Intelligence Research Society Conference*, vol. 2, pp. 562–567, 2004.
- [35] R. H. Byrd, P. Lu, and J. Nocedal, "A limited memory algorithm for bound constrained optimization," *SIAM Journal on scientific computing*, vol. 16, no. 5, 1995.
- [36] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and regression trees*. Routledge, 2017.
- [37] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, Sep. 1995, doi: 10.1007/BF00994018.
- [38] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *Journal of Computer and System Sciences*, vol. 55, no. 1, pp. 119–139, Aug. 1997, doi: 10.1006/jcss.1997.1504.
- [39] D. Nadeau and S. Sakine, "A survey on named entity recognition," *Linguisticae Investigationes*, vol. 30, no. 1, pp. 3–26, 2007, doi: 10.1007/978-981-13-9409-6_218.
- [40] A. S. Wibawa and A. Purwarianti, "Indonesian named-entity recognition for 15 classes using ensemble supervised learning," *Procedia Computer Science*, vol. 81, pp. 221–228, 2016, doi: 10.1016/j.procs.2016.04.053.

BIOGRAPHIES OF AUTHORS






Purnomo Husnul Khotimah    received her bachelor's degree (S.T.) from Universitas Indonesia in 2005 and her master's degree (M.T.) from Institut Teknologi Bandung, Indonesia in 2009. Later, she obtained her Ph.D. degree in Informatics from Kyoto University, Japan in 2018. She attained a postdoctoral position in Kyoto University Hospital, Japan, from 2018 to 2020. Currently, she is working as a researcher at the Research Center for Data and Information Sciences of National Research and Innovation Agency (BRIN). Her research interests are in data mining, information retrieval, data integration, information systems, distributed information systems, web-based implementations, and open source. Currently, she works on the topics of, but not limited to, health informatics, smart and precision farming, autonomous vehicles, and small and media enterprise. She can be contacted at purn005@brin.go.id.






Andria Arisal    (Member, IEEE) received his Bachelor of Engineering in Informatics Engineering from Institut Teknologi Bandung and Master of Engineering in distributed computing from University of Melbourne. Currently he is working as a researcher at the Research Center for Data and Information Sciences of National Research and Innovation Agency (BRIN). He is involved in various multi modal data management and analytics. He also operates institutions' high performance computing infrastructure and supports other researchers in using the infrastructure. His research interests are in data management and analytics in distributed and parallel computing environments. He can be contacted at: andria.arisal@brin.go.id.






Andri Fachrur Rozie    received his undergraduate majoring in Informatics Engineering and obtained master's degree from Department of Computer and Radio Communications Engineering from Korea University. His research area is in the field of data science, machine learning, natural language processing, and autonomous vehicles. Currently, he is working at Research Center for Data and Information Sciences, National Agency of Research and Innovation (BRIN, Indonesia). He was previously involved in conducting research for the development of data and information center systems for weather and air quality to support natural resource management and in NLP research for text classification. Since 2021, he has been involved in the research of autonomous vehicles. His responsibility includes creating environment simulation for autonomous vehicle testing using RoadRunner and Carla and the teleoperation dashboard. He can be contacted via email at andri.fachrur.rozie@brin.go.id.






Ekasari Nugraheni    obtained a bachelor's degree in information management from the AKRIND Institute of Technology Yogyakarta in 1996. She received a scholarship from the Indonesian Ministry of Research and Technology for her master's degree in informatics from the Bandung Institute of Technology, Indonesia, completed in 2016. Currently, she is a researcher of the Information Retrieval Research Group at the Research Center for Data and Information Sciences, National Research and Innovation Agency (BRIN, Indonesia). Her research interests include data analysis, data mining, deep learning, and natural language processing. She has been involved in many research areas, such as asset management, tide monitoring, semantic data warehouse, emotion recognition, human activities recognition and others. Her latest research project is mining social media for public perception toward halal food industry. She can be contacted at ekasari.nugraheni@brin.go.id.






Dianadewi Riswantini    received a scholarship from the Overseas Fellowship Program, a collaboration between the Indonesian Government and World Bank, for a bachelor's and master's degree in computer science from the Delft University of Technology, the Netherlands that are completed in 1994. She has her second master's degree in Business Administration and is currently a Ph.D. candidate in the School of Business and Management, Bandung Institute Technology, Indonesia. She engaged in big data analytics for business and management. She is a member of the Information Retrieval Research Group at the Research Center for Data and Information Sciences, National Agency of Research and Innovation (BRIN, Indonesia). Her research interests include data analytics, text mining, natural language processing, and machine learning in the fields of social and medical informatics. She can be contacted at dianadewi.riswantini@brin.go.id.






Wiwin Suwarningsih    graduated from her bachelor's degree at Informatics Program, Adityawarman Institute of Technology Bandung in 1996. She got her master's degree in 2000 at the Informatics Study Program and doctoral degree in 2017 at the School of Electrical and Informatics Engineering, Bandung Institute of Technology. Currently, she works at the Research Center for Data and Information Sciences, National Agency of Research and Innovation (BRIN, Indonesia). Her research interests are artificial intelligence, computational linguistics, Indonesian natural language processing and text mining, information retrieval and question answering systems. Since 2020, she has led a project about the identification of chili varieties using deep learning to maintain seed purity and supporting the certification of high-quality chili seeds. One of the study's objectives is to provide early detection system for chili varieties. She can be contacted at wiwi005@brin.go.id.



Devi Munandar    received his undergraduate degree majoring in Informatics Engineering. Obtained master's degree from Department of Mathematics, Universitas Padjadjaran, Bandung, Indonesia. His research is in the fields of data science, machine learning, natural language processing, multivariate statistics, and prediction analysis. Currently, he is working at the Research Center for Data and Information Sciences, National Agency of Research and Innovation (BRIN, Indonesia). He previously involved in conducting research for the development of data and information center systems for weather and air quality to support natural resource management and in NLP research for text classification and POS-Tagging on Indonesian-language tweets. In 2020-2022, he has been involved in statistical modeling research and stochastic processes, using principal component analysis (PCA) integrated with vector autoregressive (VAR) using a datamining approach for forecasting climate effects. He can be contacted at devi.munandar@brin.go.id.



Ayu Purwarianti    (Member, IEEE) received the Ph.D. degree from the Toyohashi University of Technology, in December 2007, with dissertation title of “Cross Lingual Question Answering System (Indonesian Monolingual QA, Indonesian English CLQA, Indonesian-Japanese CLQA)”. The dissertation was in the area of natural language processing or also known as computational linguistics, which is a part of artificial intelligence knowledge domain. Since then, she has been working as a lecturer with the Bandung Institute of Technology (ITB). Other than teaching and doing research, her other activity is in Indonesian Association for Computational Linguistics, where she was elected as the chair from 2016 to 2018. She was also the Chair of IEEE Education Chapter of Indonesian Section, from 2017 to 2019. She has been with IABEE since 2015. She has been founded a start-up named Prosa.ai since 2018. She has been the Chair of Artificial Intelligence Center, ITB, since August 2019. She has written several publications in conferences and journals related with computational linguistics for Indonesian language. She also provides Indonesian natural language processing tools to be used by other researchers. She can be contacted at ayu@stei.itb.ac.id.