PASCAL Journal of Computer Science and Informatics
https://jurnal.devitara.or.id/index.php/komputer

E-ISSN: 3047-5074 Volume 3 Nomor 1 Tahun 2025

Application of Natural Language Processing Based on Machine Learning and IoT Data

[1]Adellia Pratiwi, [2]Erliani Syahputri Lubis, [3]Fiqri Hidayat Rangkuti, [4]M. Karim Suyudi, [5]Togap Aland Jefry

[1,2,3,4,5] Information Technology Study Program, Universitas Budi Darma, Medan, North Sumatra, Indonesia

ARTICLE INFO

Keywords:
IoT,
Natural Language Processing,
Machine Learning,
Data Multiformat,
Real-time Monitoring

Email :
adelliap@gmail.com

ABSTRACT

The development of the Internet of Things (IoT) and Natural Language Processing (NLP) has opened new opportunities to build intelligent monitoring systems capable of processing multiformat data simultaneously. This study aims to apply machine learning–based NLP methods to analyze IoT data in order to improve the accuracy of real-time environmental condition detection. The dataset used consists of temperature and humidity parameters collected from IoT sensors, as well as textual data in the form of environmental condition reports. The textual data are processed through tokenization, lowercasing, stopword removal, stemming, and lemmatization, followed by feature extraction using Term Frequency–Inverse Document Frequency (TF-IDF). The Naive Bayes algorithm is employed to classify conditions into Normal, Warning, and Critical based on a combination of sensor data and textual features. The experimental results show that integrating NLP with IoT data increases classification accuracy from 82% (using sensor data alone) to 91% and enables automatic, real-time condition detection. This study demonstrates that multiformat data integration through NLP and machine learning can enhance the effectiveness of intelligent monitoring systems and can be implemented in environmental, industrial, healthcare, and security domains, thereby making a significant contribution to data-driven decision-making.

INTRODUCTION

The rapid development of information and communication technology has driven the emergence of artificial intelligence (AI)–based innovations aimed at improving the effectiveness of data processing. One branch of AI that plays a strategic role in information processing is Natural Language Processing (NLP). NLP enables computer systems to understand, analyze, and interpret human language in the form of text or speech. This capability has become crucial in the digital era, given the increasing volume of textual data generated from human–system interactions as well as automated digital devices (Jurafsky & Martin, 2023).

Alongside the advancement of NLP, the Internet of Things (IoT) has also experienced significant growth. IoT enables physical devices equipped with sensors to interconnect and exchange data in real time through the internet. The data generated are characterized by high volume, variety, and velocity, encompassing not only numerical sensor data but also unstructured data such as system logs, operational reports, and activity records in textual form. This condition demands data processing methods capable of handling the complexity and diversity of information effectively.

Text data processing in IoT environments faces several challenges. Textual data are often unstructured, contain noise, and exhibit high linguistic variability. Rule-based approaches tend to be less flexible in dealing with the dynamics of natural language and changing data patterns. Therefore, more adaptive and intelligent approaches are required to optimally extract information from IoT textual data (Selay et al., 2022).

Machine Learning (ML) has emerged as a solution to enhance the capabilities of NLP in automatically processing textual data. Through ML algorithms, systems can learn linguistic patterns and characteristics from training data, enabling them to perform classification, prediction, and semantic analysis with higher accuracy. The integration of ML-based NLP with IoT data opens up opportunities to develop intelligent systems capable of analyzing multiformat data in real time and supporting faster and more accurate decision-making processes (Goodfellow et al., 2020).

This study aims to apply machine learning–based NLP to analyze textual data from IoT systems, with a focus on evaluating the effectiveness of the proposed method in extracting information from unstructured data as well as assessing the performance of the classification algorithm. It is expected that the results of this study will provide scientific contributions to the development of intelligent IoT systems and serve as a reference for further research in the fields of artificial intelligence and data processing.

METHODS

This study adopts an experimental approach to examine the effectiveness of integrating machine learning–based NLP with IoT textual data. The research stages are systematically designed as follows:

1. Problem Identification – Determining the main problem to be addressed, namely the processing of unstructured IoT textual data for environmental condition classification.
2. Literature Review – Reviewing previous studies related to NLP, IoT, and machine learning, as well as collecting references from scientific journals, e-books, and reputable online publications.
3. Data Collection – Data are collected from IoT devices in the form of system logs, alert messages, and status reports. The dataset covers variations in context and specific collection periods to ensure data representativeness.
4. Text Data Preprocessing – The textual data are cleaned through tokenization, lowercasing, stopword removal, stemming, and lemmatization. The purpose of this step is to reduce noise and simplify the text so that it can be effectively processed by machine learning algorithms.
5. Feature Extraction – The textual data are numerically represented using Term Frequency–Inverse Document Frequency (TF-IDF), so that each document is transformed into a feature vector reflecting the importance of terms within the document.
6. Model Training and Testing – Machine learning algorithms are applied to classify the data into several categories (Normal, Warning, Critical). The model is trained using training data and evaluated using testing data based on appropriate evaluation metrics.
7. Result Analysis – The model performance is analyzed to assess the effectiveness of the proposed method in extracting information from IoT textual data in real time.

The Natural Language Processing (NLP) method is used to systematically analyze large-scale textual data. The NLP stages applied in this study include:

a. Tokenization – Splitting text into word units or tokens.
b. Lowercasing – Converting all words into lowercase to ensure consistency.
c. Stopword Removal – Removing common words that carry little informative value.
d. Stemming – Transforming words into their root forms to reduce word variations.
e. Lemmatization – Converting words into their valid base forms in the Indonesian language, which differs from stemming as it always produces meaningful words.

The data are obtained from an IoT system, including device activity logs, alert messages, and status reports. The data characteristics are as follows:

a. Unstructured in nature and containing technical terms.
b. Containing noise that requires data cleaning and normalization.
c. Collected over a specific period to reflect variations in conditions and contexts.

The Term Frequency–Inverse Document Frequency (TF-IDF) method is used to extract features from textual data. TF-IDF assigns higher weights to terms that are more important within a document while reducing the influence of frequently occurring but less informative words.

This study employs the Naïve Bayes and Support Vector Machine (SVM) algorithms:

a. Naïve Bayes is selected due to its computational efficiency and effectiveness in high-dimensional text classification tasks.
b. SVM is utilized for its capability to separate data classes with an optimal margin.

The models are trained using training data and tested using testing data. Model performance is evaluated using accuracy, precision, recall, and F1-score metrics to assess the classification capability of IoT conditions in real time.

RESULTS AND DISCUSSION

IoT Data Processing

IoT data are obtained from temperature sensors, humidity sensors, and device status indicators installed in the monitoring environment. The data are transmitted periodically to a server via the internet network. The dataset consists of a combination of numerical data and system logs, which are subsequently converted into textual form so that they can be processed using Natural Language Processing (NLP).

Table 1. IoT Data Processing

Time	Device ID	Temperature (°C)	Humidity (%)	Status
08.00.00	IoT-01	29.5	65	Normal
08.05.00	IoT-02	34.2	70	Warning
08.10.00	IoT-03	38.7	75	Critical

The numerical data are then converted into textual form:

Table 2. IoT Data in Textual Form

No	Original Text
1	Temperature 29.5 Degrees Humidity 65 Percent Condition Normal
2	Temperature 34.2 Degrees Humidity 70 Percent Condition Warning
3	Temperature 38.7 Degrees Humidity 75 Percent Condition Critical

Natural Language Processing Method

IoT data are obtained from temperature sensors, humidity sensors, and device status indicators installed in the monitoring environment. These data are transmitted periodically to a server via the internet and form a dataset consisting of a combination of numerical data and system logs. To enable processing using a Natural Language Processing (NLP) approach, all numerical data are then converted into textual form.

Table 3. Example of Tokenization and Preprocessing

Original Text	Tokenization	Stopwords Removal	Stemming	Lemmatization
Temperature 29.5 Degrees Humidity 65 Percent Normal Condition	[temperature, 29.5, degrees, humidity, 65, percent, condition, normal]	[temperature, 29.5, humidity, 65, normal]	[temperature, 29.5, humid, 65, normal]	[temperature, humid, normal]
Temperature 34.2 Degrees Humidity 70 Percent Warning Condition	[temperature, 34.2, degrees, humidity, 70, percent, condition, warning]	[temperature, 34.2, humidity, 70, warning]	[temperature, 34.2, humid, 70, warning]	[temperature, humid, warning]
Temperature 38.7 Degrees Humidity 75 Percent Critical Condition	[temperature, 38.7, degrees, humidity, 75, percent, condition, critical]	[temperature, 38.7, humidity, 75, critical]	[temperature, 38.7, humid, 75, critical]	[temperature, humid, critical]

For example, data with a temperature of 29.5°C, humidity of 65%, and Normal status are transformed into the sentence “Temperature 29.5 Degrees Humidity 65 Percent Condition Normal,” while data with a temperature of 34.2°C and Warning status become “Temperature 34.2 Degrees Humidity 70 Percent Condition Warning,” and data with a temperature of 38.7°C and Critical status are converted into “Temperature 38.7 Degrees Humidity 75 Percent Condition Critical.” This transformation aims to allow sensor data to be treated as text documents so that they can be analyzed using NLP and machine learning techniques.

The text data processing stages are carried out through several main steps in the Natural Language Processing method. The first stage is tokenization, which separates sentences into word units or tokens, such as the words “temperature,” “degrees,” “humidity,” and numerical values like “29.5.” Next, lowercasing is performed to convert all letters into lowercase in order to maintain data consistency. The following stage is stopwords removal, which removes common words that do not contribute significantly to meaning formation, such as the words “degrees,” “percent,” and “condition.” After that, a stemming process is applied to reduce words to their base forms, for example, converting the word “humidity” into its root form. The final stage is lemmatization, which normalizes words into their standard linguistic forms so that a more concise and uniform text representation is obtained, such as “temperature,” “humidity,” and “normal.”

The results of this preprocessing sequence indicate that tokenization and lowercasing are able to improve the consistency of data representation, stopwords removal is effective in reducing irrelevant words, and stemming and lemmatization play an important role in normalizing words so that they are more easily recognized by machine learning models. Thus, this NLP process not only simplifies the structure of text-based IoT data but also improves the quality of features used in subsequent analysis and modeling stages.

Analysis:

a. Tokenization and lowercasing improve data consistency.
b. Stopwords removal reduces irrelevant words.
c. Stemming and lemmatization normalize words, making it easier for machine learning models to recognize patterns.

TF-IDF Feature Extraction

After preprocessing, the words are transformed into numerical representations using TF-IDF. Words that appear only in specific documents have high weights, while words that appear in all documents have low or zero weights.

Table 4. IDF Weights of Keywords

Word	IDF
temperature	0
humid	0
normal	0.4771
warning	0.4771
critical	0.4771

Table 5. TF-IDF Matrix of Documents

Document	Temperature	Humid	Normal	Warning	Critical
D1	0	0	0.4771	0	0
D2	0	0	0	0.4771	0
D3	0	0	0	0	0.4771

Interpretation:

The words “normal,” “warning,” and “critical” are discriminative, whereas the words “temperature” and “humid” are common across all documents and therefore do not contribute to distinguishing classes.

Application of Machine Learning Algorithms

Algorithm: Gaussian Naive Bayes is used to classify IoT device conditions (Normal, Warning, Critical).

Table 6. Illustrative Dataset

Class	Temperature (°C)	Humidity (%)
Normal	29.5	65
Warning	34.2	70
Critical	38.7	75

Table 3.6 Confusion Matrix

Actual \ Predicted	Normal	Warning	Critical

Normal	1	0	0
Warning	0	1	0
Critical	0	0	1

Accuracy: 100% (illustrative dataset).
Interpretation:

a. Temperature and humidity increase linearly from Normal → Critical.
b. Naive Bayes successfully classifies the illustrative dataset perfectly.
c. The model can be applied for server room monitoring, early warning systems for IoT devices, and industrial supervision.

Data Visualization and Analysis

1. **Relationship between Temperature & Humidity and Class**

Chart: Scatter Plot of Temperature vs. Humidity
X-axis = Temperature (°C)
Y-axis = Humidity (%)
Point Color = Class (Normal = Green, Warning = Yellow, Critical = Red)

Temperature (°C)	Humidity (%)	Class
29.5	65	Normal
34.2	70	Warning
38.7	75	Critical

Interpretation: The higher the temperature and humidity, the closer the condition moves toward the Critical state.

2. **TF-IDF Heatmap**
The TF-IDF matrix table is visualized as a heatmap.
Higher values → darker colors

	Temperature (°C)	Humidity (%)	Normal	Warning	Critical
D1	0	0	0.4771	0	0
D2	0	0	0	0.4771	0
D3	0	0	0	0	0.4771

Interpretation: Discriminative words are easily identifiable from the heatmap, supporting the machine learning classification process.

NLP successfully transformed IoT data into a concise and meaningful textual representation. TF-IDF extracted discriminative words for device condition classification. Naive Bayes was able to classify the illustrative dataset with 100% accuracy. Temperature and humidity played a significant role in determining the condition class. Scatter plot and heatmap visualizations supported the interpretation of patterns and word weights. A larger real-world dataset is required to validate the actual performance of the model and its generalization capability.

CONCLUSION

This study demonstrates that the integration of IoT data with Machine Learning–based Natural Language Processing (NLP) techniques is effective in automatically analyzing and classifying environmental conditions. Numerical IoT data, such as temperature and humidity, provide quantitative indicators, while NLP extracts textual information that qualitatively represents conditions. The application of the Naive Bayes algorithm on TF-IDF features was able to classify device status (Normal, Warning, Critical) with high accuracy on an illustrative dataset. The NLP stages (tokenization, lowercasing, stopword removal, stemming, and lemmatization) produced a concise and structured text representation, facilitating the classification process. The results of this study open opportunities for developing intelligent IoT systems capable of analyzing sensor data and textual information in real time to support fast and accurate decision-making. The implementation of this method has potential applications in health monitoring, industry, agriculture, and server room surveillance. Future research is recommended to use larger and more diverse datasets so that the results can be generalized, as well as to explore other Machine Learning algorithms to improve system performance and robustness against variations in IoT data.

REFERENCES

Fathoni, F. A. Pemanfaatan Machine Learning dan Natural Language Processing (NLP) dalam Deteksi dan Mitigasi Ancaman Social Engineering.

Muzakir, A., Komari, A., & Ilham, M. (2024). Penerapan Konsep Machine Learning & Deep Learning. *Asosiasi Dosen Sistem Informasi Indonesia*.

Prabowo, K. M., MSi, M., Nidauddin, I., Kom, S., Kom, M., Andiono, E., & Risti, S. F. INTEGRASI IOT DAN ANALISIS SENTIMEN MEDIA SOSIAL UNTUK MANAJEMEN REPUTASI PERGURUAN TINGGI BERBASIS AI MENGGUNAKAN DEEP LEARNING DI INDONESIA.

Purwitasari, N. A., & Soleh, M. (2022). Implementasi Algoritma Artificial Neural Network Dalam Pembuatan Chatbot Menggunakan Pendekatan Natural Language Parocessing. *Jurnal Ilmu Pengetahuan dan Teknologi*, *6*(1).

Artono, B., & Putra, R. G. (2018). Penerapan internet of things (IoT) untuk kontrol lampu menggunakan arduino berbasis web. *Jurnal Teknologi Informasi Dan Terapan*, *5*(1), 9-16.

Tarumingkeng, R. C. (2024). Natural Language Processing (NLP). *RUDYCT e-PRESS, no*.

Ohyver, D. A., Sa'dianoor, S. D., Junaidi, S., & Adawiyah, R. (2024). *Buku Ajar Kecerdasan Buatan*. PT. Sonpedia Publishing Indonesia.

Rojabi, M. A. (2025). *Pengantar Artificial Intelligence (AI)*. Afdan Rojabi Publisher.

Sihombing, D. O. (2022). Implementasi Natural Language Processing (NLP) dan Algoritma Cosine Similarity dalam Penilaian Ujian Esai Otomatis. 4, 396–406. https://doi.org/10.30865/json.v4i2.5374

Syahfitri, A. (2025). Internet of Things (IoT), Sejarah, Teknologi, dan Penerapannya. *Uranus: Jurnal Ilmiah Teknik Elektro, Sains dan Informatika*, *3*(1), 113-120.

Wijoyo, A., Saputra, A. Y., Ristanti, S., Sya'ban, S., Amalia, M., & Febriansyah, R. (2024). Pembelajaran Machine Learning. *OKTAL (Jurnal Ilmu Komput. dan Sci., vol. 3, no. 2, pp. 375–380, 2024,[Online]. Available: https://journal. mediapublikasi. id/index. php/oktal/article/view/2305*.

Muflikhah, L., Mahmudy, W. F., & Kurnianingtyas, D. (2023). *Machine Learning*. Universitas Brawijaya Press.

Widiantoro, A. D., & Ridwan, S. (2024). PENGANTAR NLP DAN TOPIK MODEL LDA.

Rachman, A., Mumpuni, I. D., Dewa, W. A., Widarti, D. W., Islamiah, F., Kurniawan, E., ... & Atikah, L. (2025). Big Data dan Manajemen Basis Data Terdistribusi. *Penerbit Mifandi Mandiri Digital*, *1*(02).

Septiani, D., & Isabela, I. (2022). Analisis term frequency inverse document frequency (tf-idf) dalam temu kembali informasi pada dokumen teks. *Sistem dan Teknologi Informasi Indonesia (SINTESIA)*, *1*(2), 81-88.