Jurnal Teknik Informatika (JUTIF)
P-ISSN: 2723-3863
E-ISSN: 2723-3871

Vol. 6, No. 6, December 2025, Page. 5865-5881
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.6.5482

Enhancing Classification of Self-Reported Monkeypox Symptoms on Social Media Using Term Frequency-Inverse Document Frequency Features and Graph Attention Networks

Rizailo Akfa Rizian[1], Irwan Budiman[*2], Mohammad Reza Faisal[3], Dwi Kartini[4], Fatma Indriani[5] Umar Ali Ahmad[6]

[1,2,3,4,5]Department of Computer Science, Lambung Mangkurat University, Indonesia
[6]Collaborative Researcher, Kanazawa University, Kanazawa, Ishikawa, Japan

Email: [2]irwan.budiman@ulm.ac.id

Received : Nov 26, 2025; Revised : Nov 29, 2025; Accepted : Dec 2, 2025; Published : Dec 25, 2025

Abstract

Early detection of infectious diseases plays a crucial role in minimizing their spread and enabling timely intervention. In the digital era, social media has emerged as a valuable source of real-time health information, where individuals often share self-reported symptoms that can serve as early warning signals for disease outbreaks. However, textual data from social media is typically unstructured, noisy, and contextually diverse, posing challenges for conventional text classification methods. This study proposes a hybrid model combining Term Frequency–Inverse Document Frequency (TF-IDF) feature representation with a Graph Attention Network (GAT) to enhance the early detection of Monkeypox-related self-reported symptoms on Indonesian social media. A dataset of 3,200 tweets was collected through Tweet-Harvest and subsequently preprocessed and manually labeled, producing a balanced distribution between positive (51%) and negative (49%) samples. TF-IDF vectors were used to construct a document similarity graph via the k-Nearest Neighbors (k-NN) method with cosine similarity, enabling GAT to leverage both textual and relational information across posts. The model's performance was evaluated using accuracy, precision, recall, and macro-F1, with macro-F1 serving as the primary indicator. The proposed TF-IDF + GAT model achieved 93.07% accuracy and a macro-F1 score of 93.06%, outperforming baseline classifiers such as CNN (92.16% macro-F1), SVM (85.73%), Logistic Regression (84.89%). These findings demonstrate the effectiveness of integrating classical text representations with graph-based neural architectures for improving social media based disease surveillance and supporting early epidemic response strategies.

Keywords : *Graph Attention Network, Monkeypox, Social Media, Text Classification, TF-IDF*

1. INTRODUCTION

Infectious diseases represent a significant global health challenge that can result in profound social, economic, and psychological ramifications within a brief temporal framework. In recent years, there has been a notable resurgence of Monkeypox cases, raising global concern due to their extensive transcontinental dissemination and the resemblance of their clinical manifestations to those of chickenpox [1]. The Monkeypox virus is classified within the genus Orthopoxvirus and was first recognized in human populations in the Democratic Republic of the Congo in 1970. Following several decades of relative containment, the disease re-emerged in 2022, marked by a substantial rise in cases across various non-endemic regions [2], prompting the World Health Organization (WHO) to classify it as a Public Health Emergency of International Concern (PHEIC) [3]. The initial symptoms of Monkeypox, including fever, myalgia, asthenia, and dermal eruptions, frequently overlap with those of other illnesses, rendering the process of early detection exceedingly difficult yet vital for the efficacy of medical interventions[4][5].

Vol. 6, No. 6, December 2025, Page. 5865-5881
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.6.5482

Early detection constitutes a fundamental strategy in the management of outbreaks, as it enables healthcare professionals and public health authorities to administer prompt treatment prior to the widespread propagation of the infection [6]. In contemporary settings, conventional clinical report-based methodologies frequently lag behind in relation to the rapidity with which information circulates in the digital realm. With the growing prevalence of social media, platforms such as Twitter, Facebook, and Instagram are emerging as potential reservoirs of data that reflect public perceptions, individual experiences, and even preliminary symptoms of illness reported voluntarily by users [7]. This phenomenon is referred to as social sensing or digital epidemiology, wherein user-generated content functions as collective sensors that can elucidate the dynamics of public health in real-time [8].

Numerous empirical investigations have demonstrated that social media platforms can be effectively utilized for monitoring infectious diseases such as COVID-19, influenza, and Monkeypox, owing to their rapid and continuous information dissemination capabilities [9]. For instance, Shi et al. [10] indicated that the analysis of social media data can serve as a potent tool for real-time detection and forecasting of public health trends, achieving a commendable level of accuracy in discerning patterns of disease distribution. Nonetheless, textual data derived from social media is predominantly unstructured, often encompasses considerable noise, and exhibits a diverse array of linguistic styles and contextual meanings [11][12]. This necessitates the formulation of text representation methodologies that are proficient in extracting salient information from succinct, non-raw sentences.

One representation technique that remains prevalent due to its straightforwardness and effectiveness is Term Frequency—Inverse Document Frequency (TF-IDF). This technique computes the significance of a term based on its occurrence within a specific document in relation to the entire corpus, thereby elucidating the most pertinent vocabulary [13]. Despite its simplicity, TF-IDF has demonstrated competitiveness across various domains of text analysis research, including hoax detection, sentiment evaluation, and disease classification [14][15]. Research conducted by Das et al. [16] and Sutriawan et al. [17] revealed that TF-IDF is capable of yielding robust classification outcomes when integrated with machine learning algorithms such as Support Vector Machines (SVM) or XGBoost. In the realm of medical text analysis, the TF-IDF methodology is particularly advantageous as it does not necessitate extensive computational resources akin to those required by transformer-based models, such as BERT or GPT [18].

However, the TF-IDF approach is not without its limitations, as it fails to adequately capture semantic interrelations between documents or between words. To mitigate these deficiencies, a Graph Neural Network (GNN)-based strategy has been proposed, which leverages the relational structure of data represented in graph form [19]. Within the GNN architecture, each document or word is conceptualized as a node, while the semantic connections among them are represented by edges. The learning process is facilitated through a message passing mechanism, wherein each node refines its representation by integrating information from its proximal neighbors [20]. This methodology is regarded as capable of encapsulating relational contexts that remain inaccessible to sequence-based methods such as Long Short-Term Memory (LSTM) networks [21].

One notable variant of Graph Neural Networks (GNNs) is the Graph Attention Network (GAT), which incorporates an attention mechanism designed to dynamically evaluate the relative significance of each neighboring node. This mechanism enables the model to prioritize pertinent nodes while disregarding less critical relationships, thereby enhancing the informativeness of the resultant graph representation [22]. A plethora of contemporary research underscores the superiority of GAT across various tasks in text classification and anomaly detection. Li et al. [23] proposed the incorporation of adversarial training within the GAT framework to augment the model's robustness against noise, while Malik et al. [24] implemented the Ensemble Graph Neural Network for the purpose of detecting fake news, achieving a notable enhancement in accuracy. Furthermore, a study conducted by Zikrina and

Vol. 6, No. 6, December 2025, Page. 5865-5881
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.6.5482

Fitriyani [25] demonstrated that the amalgamation of TF-IDF with GNN-oriented graph models significantly bolstered the efficacy of hate speech classification in the Indonesian language, outperforming traditional methodologies.

The synergistic combination of classical text representation (TF-IDF) with graph-based learning (GAT) facilitates a harmonious balance between representation efficiency and complexity. TF-IDF serves a crucial function in producing straightforward yet significant representations of numerical vectors, whereas GAT leverages inter-document relationships to enrich the contextual framework of classification [26]. Consequently, the integration of these two methodologies is anticipated to enhance the system's capability to discern early indicators of disease through the analysis of unstructured social media content [27].

While several prior studies have applied TF-IDF in combination with GNN-based architectures for tasks such as hate speech detection and disaster-related tweet classification [25][27], no existing work has explored the integration of TF-IDF with Graph Attention Networks for classifying self-reported Monkeypox symptoms in social media. This gap indicates a lack of research addressing relational text modeling within the context of digital epidemiology, particularly for emerging diseases. The present study addresses this limitation by proposing a TF-IDF-GAT framework optimized through k-Nearest Neighbors graph construction and extensive hyperparameter tuning to improve early detection performance.

In the realm of early detection of Monkeypox disease, this methodological approach holds considerable promise for fortifying social media-driven monitoring systems. By constructing document similarity graphs utilizing the cosine similarity-based K-Nearest Neighbors (k-NN) methodology, the interrelations among uploads can be modeled as semantic networks that delineate the proximity of topics and contexts [28]. The Attention Network Graph is subsequently employed to extract salient features from this network via an attention layer, which assesses the relevance of each connection in relation to the classification outcomes. Evaluative measures were conducted utilizing accuracy, precision, recall, and macro-F1 metrics, with macro-F1 being designated as the primary indicator due to its greater representativeness of imbalanced data scenarios [29].

This research endeavor is centered on enhancing the efficacy of early detection models by examining the impact of graph parameters, including the quantity of neighbors (K), as well as the arrangement of GAT hyperparameters such as attention heads, hidden units, learning rate, and weight decay. Through a series of comparative experiments conducted against benchmark models including Support Vector Machine (SVM), Logistic Regression, and Convolutional Neural Network (CNN), the investigation seeks to ascertain the most proficient models for classifying user-generated content related to Monkeypox symptoms [30].

The findings of this research are anticipated to yield scholarly contributions across three principal dimensions. First, in pragmatic terms, the formulated model may be employed by health organizations and epidemiological investigators to swiftly monitor disease patterns utilizing publicly available data from social media, thereby enhancing the national early warning framework. Second, from an academic perspective, this study augments the existing literature concerning the amalgamation of Graph Neural Networks with traditional representation techniques such as TF-IDF, particularly within the digital health text sector. Third, methodologically, this inquiry paves the way for additional developmental prospects within the realms of Natural Language Processing (NLP) and Graph Learning, encompassing the utilization of semi-supervised node classification methodologies and the incorporation of contextual embeddings such as BERT within graph-based architectures [31]. Consequently, this methodology is relevant not only for identifying Monkeypox but also for supporting the broader surveillance of emerging infectious diseases. Building on this motivation, the present study proposes and evaluates a TF-IDF, GAT hybrid model, optimized through graph parameter selection and GAT hyperparameter

Vol. 6, No. 6, December 2025, Page. 5865-5881
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.6.5482

tuning, to achieve superior relational classification performance compared to traditional machine learning baselines.

2. METHOD

2.1. Dataset

The datasets employed in this investigation were acquired via a web crawling methodology utilizing the Tweet-Harvest tool, which is designed to extract data from search results on the Twitter (X) platform. The data collection methodology involves targeting the most recent uploads on the “LATEST” tab and implementing lang:id language filters to ensure the inclusion of exclusively Indonesian tweets. The keyword employed in the search methodology is “monkeypox OR mpox lang:id”, specifically intended to identify public posts that encompass topics or personal experiences associated with the Monkeypox disease.

From this process, approximately 3,200 tweets were procured, encapsulating individuals' perceptions, apprehensions, and self-reported manifestations of Monkeypox. The extracted raw data comprises several attributes, including created_at, full_text, username, lang, and retweet_count. Nonetheless, in the context of this study, only the full_text column was utilized as the principal source for textual analysis, while the remaining attributes were disregarded due to their lack of direct relevance to the objectives of the classification, as illustrated in table 1.

Table 1. Raw Dataset

created_at	full_text	username	lang	retweet_count
Fri Oct 04 10:13:47 2024	@txtdarisisange Alhamdullilah negative monkeypox sama covid	virzz	in	6
Sun Oct 20 11:33:46 2024	Kantorku terjangkit monkeypox. Celengg	bagasrivan_	in	11
Mon Jan 06 16:02:28 2025	Aku takut kena monkeypox gejalanya sumpah persis banget anjir plis takut banget. Semoga aku ga kenapa kenapa	rrevales	in	0
Mon Apr 28 12:38:49 2025	kalau monkeypox kan harus demam sakit kepala sakit tenggorokan. aku hanya cewek gatal luar biasa ajah wkwkw	defisparco	in	2

Moreover, a systematic manual labeling procedure was implemented for each tweet to differentiate between posts that contained personal accounts with pertinent indicators of Monkeypox symptoms (label 1) and those that lacked any indication of relevant symptoms (label 0). Both categories provide an equal representation of self-reported content, yet they vary in the degree of alignment of symptoms with the clinical characteristics of Monkeypox. Tweets assigned label 1 (true/positive) encompass personal declarations regarding typical symptoms such as fever, headache, skin rash, or apprehensions about contracting the disease. In contrast, tweets designated with label 0 (false/negative) included assertions that inaccurately linked non-specific symptoms to Monkeypox, referenced other similar illnesses, or contained trivial contexts such as humor and general opinions. The labeling outcomes yielded a balanced distribution of data, specifically approximately 51% of tweets with label 1 (true) and 49% of tweets with label 0 (false). The final dataset is subsequently organized into a two-

Vol. 6, No. 6, December 2025, Page. 5865-5881
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.6.5482

column format, comprising data (tweet text) and label (0 or 1), which will be utilized in the preprocessing and feature extraction phases, shown in table 2.

Table 2. Labeled Dataset

Text	Label
@txtdarisisange Alhamdullilah negative monkeypox sama covid	0
Kantorku terjangkit monkeypox. Celengg	1
Aku takut kena monkeypox gejalanya sumpah persis banget anjir plis takut banget. Semoga aku ga kenapa kenapa	1
kalau monkeypox kan harus demam sakit kepala sakit tenggorokan. aku hanya cewek gatal luar biasa ajah wkwkw	0

2.2. Pre-Processing

The pre-processing phase constitutes a critical component in the preparation of data, facilitating efficient analysis and yielding a pristine representation of textual material [32]. Prior to the ingestion of data into the model, the entirety of the raw tweet acquired via the web crawling methodology is subjected to a comprehensive series of cleansing and structuring procedures [33].

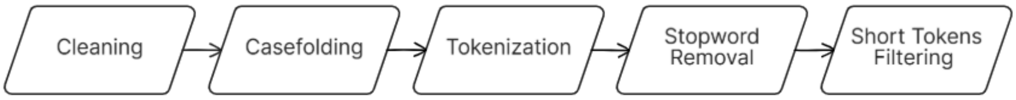

Figure 1. Preprocessing steps

Figure 1 illustrates the preprocessing procedure that is preceded by sanitation, that is, the eradication of hyperlinks, punctuation marks, numerals, symbols, as well as non-alphabetic characters that frequently emerge in social media texts. This phase aims to expunge noise and constituents that are extraneous to the context of the analysis. Subsequently, a case folding operation is executed, in which the entire text is transformed into lowercase letters to uphold the uniformity of the token. The ensuing stage is tokenization, which disaggregates the text into individual words so that it may be processed independently in the subsequent phase. The process advances with stopword elimination, that is, the removal of commonplace words that lack significant meaning, such as auxiliary words or function words. The compilation of stopwords employed encompasses both Indonesian and English to accommodate the phenomenon of code-mixing in tweets. Preprocessing was implemented in Python using the NLTK library for tokenization and stopword removal, combined with custom Indonesian and English stopword lists. The conclusive step is filtering short tokens, specifically the expulsion of words with a length of fewer than three characters to ensure that the final output comprises solely meaningful words.

Then Figure 2 illustrates the outcome of this comprehensive pre-processing operation in the form of purified text that has been normalized and is prepared for utilization at the feature extraction phase employing the Term Frequency-Inverse Document Frequency (TF-IDF) methodology. This methodology converts the text into a quantitative representation that signifies the extent of significance of each lexeme to the entire corpus of data. Moreover, the outcomes of these feature representations are utilized in the procedure of constructing a cosine similarity-based document graph, which subsequently serves as the input for the Graph Attention Network (GAT) model to execute the classification procedure.

Vol. 6, No. 6, December 2025, Page. 5865-5881
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.6.5482

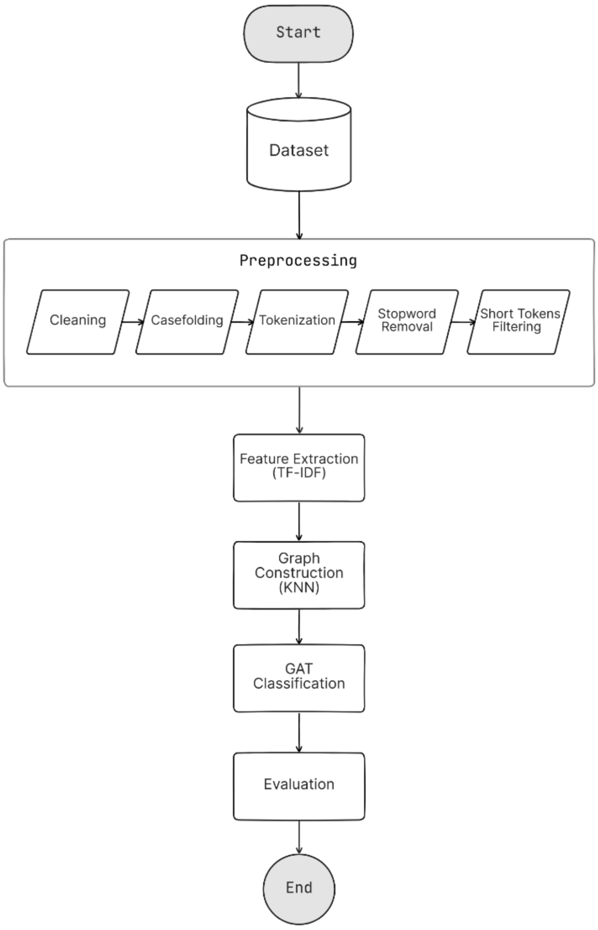

Figure 2. Research flow

The GAT model was formulated to investigate semantic interrelations among documents adaptively through attention mechanisms, thereby it is anticipated to enhance precision in identifying user uploads that encompass self-report indications of Monkeypox symptoms.

2.3. Feature Extraction TF-IDF

TF-IDF (Term Frequency–Inverse Document Frequency) is a statistical method employed to assess the extent to which a term holds significance within a text by considering its frequency of occurrence in the document as well as its rarity across the entire corpus [34]. The metrics pertaining to the TF-IDF scoring generated for each lexeme are subsequently consolidated into a feature vector that represents each tweet during the classification process. This encapsulation facilitates the model in emphasizing salient lexemes while diminishing the influence of less informative common terms, thus proving to be of paramount importance in the domain of feature extraction within textual analysis [34]. The study conducted by Zikrina and Fitriyani utilized a Graph Neural Network (GNN) in conjunction with Term Frequency-Inverse Document Frequency (TF-IDF) to categorize hate speech within a dataset comprising 13,169 tweets in the Indonesian language, resulting in an accuracy of 92.90%, a precision of 95.34%, and a recall of 85.64% [25].

The initial phase in the computation of TF-IDF is to ascertain the Term Frequency (TF), which elucidates the frequency with which a lexeme emerges within an individual manuscript. The TF metric signifies the comparative degree of significance of the lexeme within the framework of the manuscript and is articulated by equation 1 [16]:

Vol. 6, No. 6, December 2025, Page. 5865-5881
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.6.5482

$$TF_{t,d} = \frac{f_{t,d}}{\sum_k f_{k,d}} \quad (1)$$

where $f_{t,d}$ is the prevalence of occurrence of term t in manuscript d, while the denominator signifies the aggregate number of lexemes in the manuscript [16].

Furthermore, Inverse Document Frequency (IDF) is utilized to assess the infrequency of a term throughout the corpus. The less frequently a term occurs within a collection of documents, the greater its IDF value [35]. The IDF equation is illustrated in equation 2:

$$IDF_t = \log\left(\frac{N}{df_t}\right) \quad (2)$$

where N represents the aggregate quantity of documents within the corpus, and df_t denotes the quantity of documents encompassing the term t [35].

The conclusive magnitude of TF-IDF is derived by multiplying the values of TF and IDF. This metric assigns greater significance to terms that manifest frequently in particular documents yet infrequently in alternative documents [36]. The comprehensive scope is delineated in equation 3:

$$TF - IDF_{t,d} = TF_{t,d} \times IDF_t \quad (3)$$

The amalgamation of these two elements enables the system to differentiate a term that possesses a significant contextual connotation from a prevalent word that occurs frequently throughout the manuscript [36].

Feature extraction constitutes a pivotal phase in the methodology of text classification, as at this juncture unstructured textual data is converted into numerical representations that can be comprehended by machine learning algorithms. In this investigation, the feature extraction process utilizing the TF-IDF Vectorizer yielded a fixed-dimensional vector matrix for each tweet, wherein each component signifies the weight of the lexical importance relative to the document and the corpus in its entirety. The dataset that has undergone the pre-processing phase is subsequently partitioned into 80:10:10, consisting 80% allocated for the training dataset, 10% for the validation dataset, and 10% for the testing dataset. This segmentation is implemented to mitigate overfitting and to ensure a robust generalization of the model [37]. After the information dissemination procedure, each textual exemplar was transformed into a numerical format employing the TF-IDF methodology, wherein each tweet was depicted as a weighted feature vector predicated on the relative frequency and scarcity of lexemes in the corpus. This representation is subsequently utilized as an input in the graph construction phase to establish similarity relations among documents based on cosine similarity, prior to ultimately being processed by the Graph Attention Network (GAT) model at the classification juncture.

2.4. Graph Construction

Graph construction is a significant phase in graph-based learning that seeks to depict semantic associations among documents in the form of graph structures. In the domain of text mining, each document (in this instance, a tweet) is portrayed as a vertex (v_i), where the edge (e_{ij}) illustrates the similarity relationship between two documents predicated on the congruence value of the textual content. This representation enables the model to not only comprehend the individual significance of a document but also to leverage the interdependence among documents that possess a comparable context [21][23].

Connections among nodes are formulated utilizing a vector similarity metric, predominantly based on cosine similarity, which assesses the directional nearness of two vectors within a TF-IDF space. The similarity metric between two documents d_i and d_j is determined by equation 4:

Vol. 6, No. 6, December 2025, Page. 5865-5881
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.6.5482

$$CosineSim(d_i, d_j) = \frac{d_i \cdot d_j}{\|d_i\| \, \|d_i\|} \quad (4)$$

with $d_i \cdot d_j$ representing the outcome of point multiplication between two TF-IDF vectors, while $\|d_i\|$ and $\|d_j\|$ denote the magnitude (norm) of each vector. Cosine similarity values fluctuate between 0 and 1, wherein values proximal to 1 signify that both documents exhibit a substantial content resemblance.

To establish an effective graph, the investigation employed the K-Nearest Neighbors (k-NN) methodology, wherein each vertex v_i is linked to another vertex k that possesses the maximum similarity metric [38]. Consequently, the adjacency matrix (Adjacency Matrix) A can be articulated by equation 5:

$$A_{ij} = \begin{cases} 1, d_j \in KNN(d_i) \\ 0, otherwise \end{cases} \quad (5)$$

This configuration generates an undirected graph depicting semantic associations among documents, wherein each vertex possesses a limited degree in accordance with parameter k.

In this study, graph formulation was executed based on the outcomes of the TF-IDF vectorization representation, subsequently the similarity metric between documents was computed utilizing cosine similarity to ascertain connectivity among nodes. The parameter k is ascertained empirically through experimental procedures to achieve a equilibrium between graph density and computational intricacy. The findings of this graph formulation serve as the principal input for the Graph Attention Network (GAT) model to examine the interrelations among nodes adaptively through the attention mechanism, thereby enabling the model to accentuate the pertinent nodes in the classification process of Monkeypox self-reported uploads.

2.5. Graph Attention Network (GAT)

The Graph Attention Network (GAT) constitutes an advancement of the Graph Neural Network (GNN) framework that integrates an attention mechanism to evaluate the relative contributions among nodes in graphs adaptively. In contrast to the Graph Convolutional Network (GCN), which employs fixed-average aggregation, GAT computes dynamic weights for each relation (edge) predicated on its degree of pertinence to the central node [19]. This mechanism enables models to concentrate more on informative neighbors while disregarding less significant connections, thereby yielding more contextual and meaningful representations of nodes [20].

GAT executes a quantitative function that computes the attentional significance on each combination of adjacent vertices, with each vertex v_i possessing an input feature vector $h_i \in \mathbb{R}^F$ being altered via the $W \in \mathbb{R}^{F' \times F}$ learned weight matrix delineated by equation 6:

$$h_i' = Wh_i \quad (6)$$

Subsequently, the significance of the attention coefficient among vertices is computed for each duo of adjacent vertices (i, j) utilize the communal self-attention mechanism $a(\cdot)$ on equation 7:

$$e_{ij} = a(Wh_i, Wh_j) \quad (7)$$

This parameter is then normalized by the softmax function so that each vertex possesses a uniform weight distribution to its adjacent nodes as delineated in equation 8:

$$a_{ij} = softmax_j(e_{ij}) = \frac{exp(e_{ij})}{\sum_{k \in N(i)} exp(e_{ik})} \quad (8)$$

Vol. 6, No. 6, December 2025, Page. 5865-5881
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.6.5482

Lastly, a novel depiction of the node i acquired through the amalgamation of the portrayal of weighted adjacent entities in accordance with the magnitude of the attention coefficient computed by equation 9:

$$h_i'' = \sigma\left(\sum_{i \in N(i)} a_{ij} W h_j\right) \quad (9)$$

Where $\sigma(\cdot)$ is a non-linear activation mechanism, such as *ELU (Exponential Linear Unit)*.

In this study, Graph Attention Network (GAT) was employed to analyze document graphs formulated from similarity associations among tweets based on TF-IDF representations and cosine similarity computations utilizing the K-Nearest Neighbors (kNN) methodology. Each vertex signifies one tweet, whereas the edge interlinks it to other tweets that possess the most analogous semantic context. The attention mechanism in GAT accentuates greater significance on pertinent neighboring vertices, thereby enabling the model to utilize the relationship between tweets more contextually and enhance the system's capacity to identify early self-reported manifestations of Monkeypox.

3. RESULT

3.1 Preprocessing Results

Datasets that have undergone the annotation procedure subsequently progress to the textual pre-processing phase, which encompasses five principal stages: cleansing, case normalization, tokenization, elimination of stopwords, and filtering of brief tokens. This procedure aspires to harmonize textual content and diminish extraneous elements so that the representation of attributes becomes more precise. The phases of such pre-processing are depicted in Table 1, which elucidates the refinement and standardization processes of the data preceding the feature extraction phase, with the resulting cleaned text outcomes are shown in table 3 .

Table 3. Cleaned text

Clean_text
alhamdullilah negative monkeypox covid
kantorku terjangkit monkeypox celengg
takut kena monkeypox gejalanya sumpah persis banget anjir plis takut banget
monkeypox demam sakit kepala sakit tenggorokan hanya cewek gatal ajah

3.2 TF-IDF Feature Configuration Results

Subsequently, a meticulous examination of the TF-IDF parameters was conducted to ascertain the impact of configuration variations on the establishment of vector representations and the quantity of interconnections within the graph. Parameters scrutinized encompass ngram range (1) alongside combinations (1, 2), and min_df values (1, 2). The ngram range value delineates the context of the lexeme employed (singular term or bi-word phrase), whilst min_df excludes infrequently occurring terms to mitigate the introduction of extraneous noise. Modifications in these two parameters influence the number of attributes as well as the intensity of connections among nodes in the graph, as illustrated in table 4.

Table 4. TF-IDF Parameter

max_feature	ngram_range	min_df	max_df	feature
5000	1, 2	1	0,9	1811
5000	1, 2	2	0,9	1475
5000	1	1	0,9	286

Vol. 6, No. 6, December 2025, Page. 5865-5881
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.6.5482

3.3 Graph Construction Parameter Results

The procedure is perpetuated by examining the quantity of neighbors (K) in the Graph Construction methodology to ascertain its influence on graph topology and classification efficacy. Diminutive K values culminate in sparse graphs and jeopardize the retention of semantic knowledge, whereas excessively elevated K values augment complexity and the likelihood of over-smoothing. Furthermore, the magnitude of K also impacts the quantity of edges generated, wherein an increasing number of edges can enhance the dissemination of information among nodes but may concurrently diminish efficiency and result in a reduction in precision if the graph becomes overly dense, as illustrated in table 5.

Table 5. K Value and Number of Edges

K	Edges
7	35188
9	44494
11	53586

3.4 Model Performance

The subsequent procedure concentrates on ascertaining the most advantageous hyperparameter arrangement within the Graph Attention Network (GAT). Predicated on ten experiments of each pair of parameter combinations, a suboptimal configuration was attained with concealed values of 32, heads 2, dropout 0.3, learning rate 8e-3, weight decay 5e-4, patience 35, and epochs 400. This classification was selected due to its provision of training stability in conjunction with the optimal efficacy on evaluative metrics. The amalgamation of TF-IDF parameters as well as K values are subsequently employed in conjunction with the optimal GAT configuration to yield the final model. The assessment was performed utilizing the metrics of accuracy, precision, recall, and F1-score. Comprehensive testing results indicate that this configuration affords the most optimal classification performance on Monkeypox self-reported data.

Table 6. Performance Comparison on Self-Reported Monkeypox Dataset

ngram_range	min_df	K	Accuracy	Precision	Recall	F1-score
1, 2	1	7	85.54%	86.18%	85.54%	85.48%
1, 2	2	7	84.04%	84.32%	84.04%	84.00%
1	1	7	80.72%	83.04%	80.72%	80.38%
1, 2	1	9	93.07%	93.26%	93.07%	93.06%
1, 2	2	9	90.66%	90.66%	90.66%	90.66%
1	1	9	83.43%	86.35%	83.43%	83.09%
1, 2	1	11	90.96%	90.97%	90.96%	90.96%
1, 2	2	11	88.55%	88.92%	88.55%	88.53%
1	1	11	84.34%	85.20%	84.34%	84.24%

The evaluation outcomes delineated in Table 6 elucidate that the amalgamation of the parameters TF-IDF with ngram_range (1, 2), min_df 1, and the quantity of K 9 in the Graph Construction methodology, in conjunction with the optimal GAT hyperparameter configuration (hidden 32, heads 2,

Vol. 6, No. 6, December 2025, Page. 5865-5881
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.6.5482

dropout of 0.3, learning rate 8e-3, weight decay 5e-4, patience 35, epochs 400), culminates in the most favorable performance with an accuracy of 93.07%, accuracy of 93.26%, recall 93.07%, and F1-score 93.06%. This amalgamation illustrates that the parameter configuration yields the most superior results among the diverse variations examined, with consistent performance across all evaluative metrics.

4. DISCUSSIONS

The most effective training paradigms derived from a series of empirical investigations exhibit stable and consistent efficacy across diverse parameter configurations. Based on the evaluative outcomes delineated in the Results section, the amalgamation of TF-IDF (ngram_range = 1, 2; min_df = 1) with K = 9 and the optimal GAT hyperparameter arrangement culminated in the highest accuracy metric of 93.07% and equitable performance in terms of precision, recall, and F1-score indices. This indicates that the model successfully attains a equilibrium point between generalizability and sensitivity to the dataset, thereby demonstrating the capacity to identify Monkeypox symptomatology patterns proficiently.

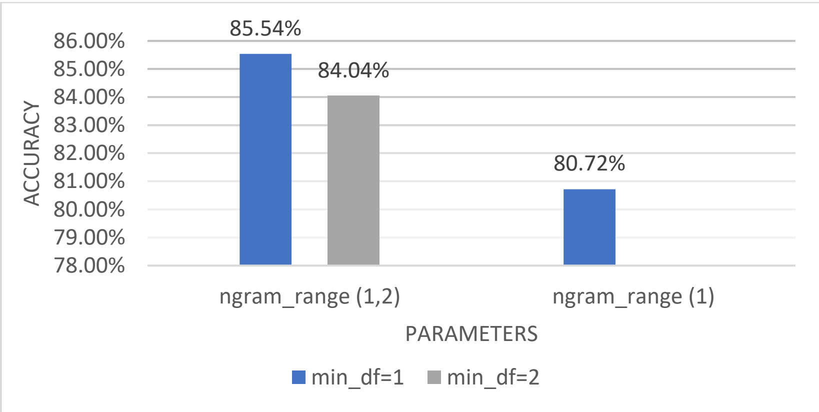

Figure 3. Performance Comparison at K=7

In Figure 3 illustrating the experimental outcomes with K=7, the amalgamation of ngram_range (1, 2) and min_df=1 parameters culminated in the optimal performance with an accuracy of 85.54% and the preeminent macro-F1 among all configurations. This metric signifies that the incorporation of bigrams aids the model in comprehending the contextual relationship between lexemes, whilst the minimal threshold of min_df retains significant attributes that infrequently manifest but are pertinent.

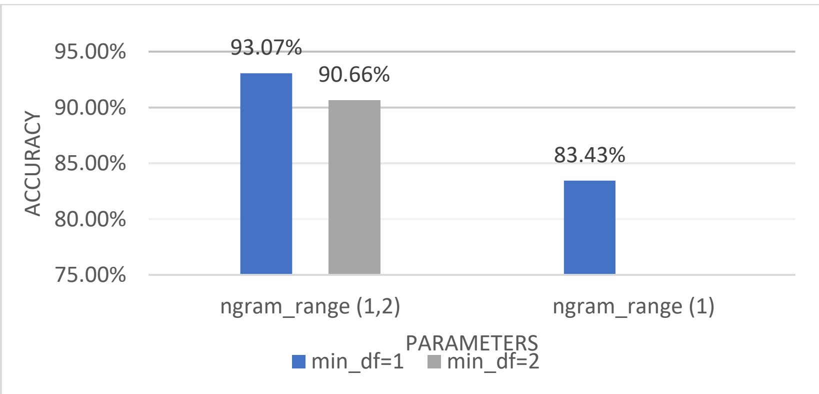

Figure 4. Performance Comparison at K=9

Jurnal Teknik Informatika (JUTIF) Vol. 6, No. 6, December 2025, Page. 5865-5881
P-ISSN: 2723-3863 https://jutif.if.unsoed.ac.id
E-ISSN: 2723-3871 DOI: https://doi.org/10.52436/1.jutif.2025.6.6.5482

Moreover, in Figure 4, with K=9, the model exhibited a substantial enhancement in performance, achieving a peak macro-F1 score of 93.06%, alongside an accuracy measurement of 93.07%. These findings suggest that K=9 represents the most advantageous number of neighbors within the Graph Construction framework, as it effectively equilibrates graph connectivity with computational complexity. With an adequately dense graph configuration that avoids the pitfalls of over-smoothing, the GAT model is capable of analyzing the intricate patterns of inter-node relationships more proficiently and generating stable feature representations throughout the information dissemination process.

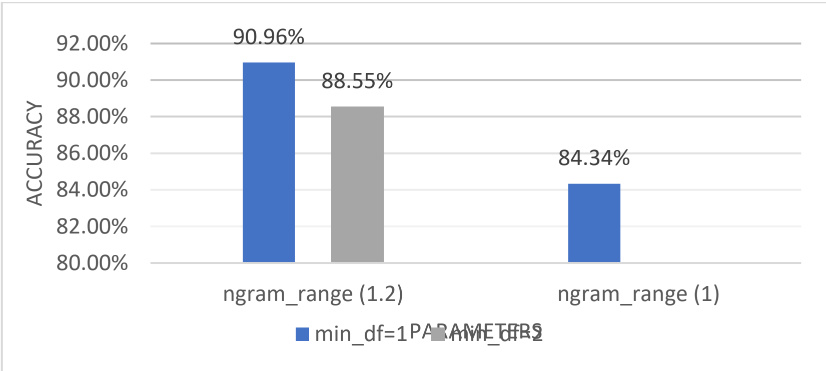

Figure 5. Performance Comparison at K=11

Meanwhile, in Figure 5 with K=11, model efficacy diminished with macro-F1 peaking at 90.95% and accuracy at 90.96%. This decline is attributed to the augmenting number of edges that induce the graph to become excessively dense, thereby rendering the information disseminated among nodes redundant and giving rise to the phenomenon of over-smoothing. This condition culminated in the model encountering challenges in preserving the distinctiveness of each tweet during the classification process.

Overall, the empirical outcomes indicated that the amalgamation of TF-IDF bigram, min_df=1, and K=9 constituted the most effective configuration for identifying self-reported symptoms of Monkeypox on social media. The bigram representation fortifies the linguistic context, whilst the graph structure with nine neighbors facilitates optimal connectivity for the attention mechanism in the GAT to identify the most pertinent relationships among nodes. This configuration produces the highest macro-F1 of 93.06% and accuracy of 93.07%, illustrating the superiority of graph-based methodologies over alternative comparative models.

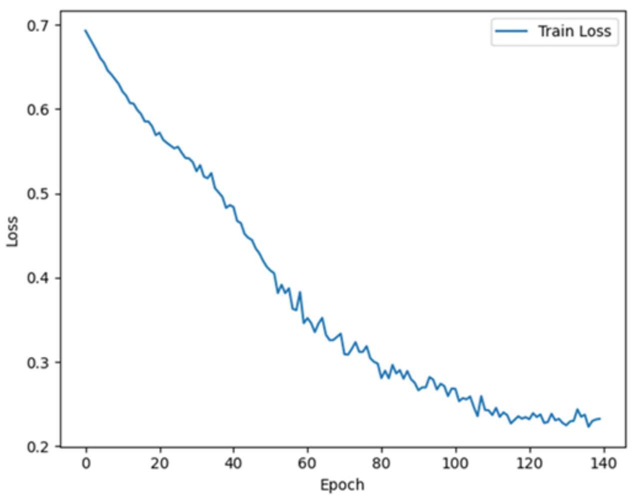

Figure 6. Training Loss

Vol. 6, No. 6, December 2025, Page. 5865-5881
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.6.5482

Figure 6 illustrates the optimal model efficacy produced under TF-IDF bigram configurations, min_df 1, and K=9, further corroborated by the consistency of the model training procedure. This is observable in Figure 6, which delineates the training loss trajectory in relation to the number of epochs throughout the training phase. A sustained trend of loss reduction from the initial value of approximately 0.7029 to achieving 0.2324 at the 140th epoch signifies that the training procedure is progressing proficiently and steadily, devoid of any signs of overfitting. The gradual decrease suggests that the model is capable of assimilating the graph representation incrementally and converging towards optimal conditions. Minor oscillations post the 100th epoch reflect adaptive weight modifications influenced by the attention mechanism, as well as the implementation of weight decay (5e-4), which functions to mitigate overlearning. The amalgamation of a learning rate of 8e-3 and a dropout rate of 0.3 also sustains the equilibrium between generalizability and model precision in the training dataset, thereby facilitating the attainment of optimal performance in these configurations.

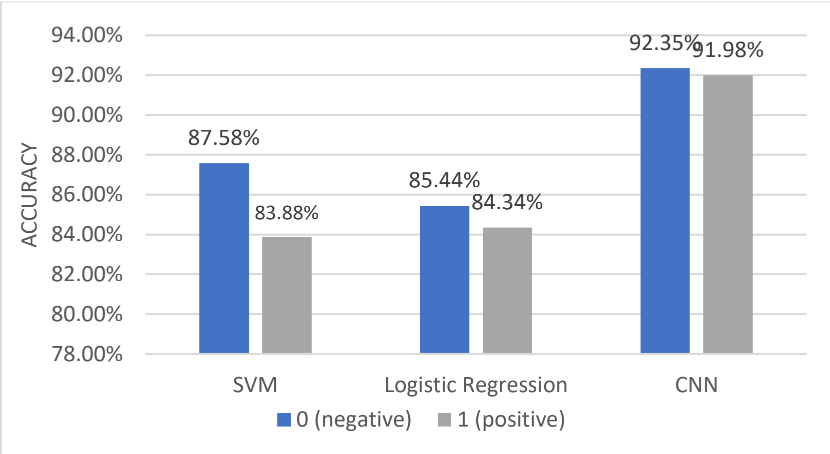

Figure 7. F1-Score Comparison Among Baseline Models

For comparison against the principal GAT-based paradigm, Figure 7 elucidates a juxtaposition of the efficacy of three comparative models, specifically Support Vector Machine (SVM), Logistic Regression, and Convolutional Neural Network (CNN), predicated on F1-score metrics for each category. The CNN model exhibited the most favorable outcomes with an F1-score of 92.35% for the negative category and 91.98% for the positive category, underscoring its proficient capability in discerning spatial patterns in textual representations. In contrast, SVM attained moderate efficacy with F1-scores of 87.58% for negative categories and 83.88% for positive categories, succeeded by Logistic Regression yielding inferior F1-scores of 85.44% and 84.34% correspondingly.

Table 7. Comparison with Related Works

Ref	**Method**	**Performance (Macro-F1)**
[17]	SVM + TF-IDF	85.73%
[30]	Logistic Regression + TF_IDF	84.89%
[30]	CNN + TF-IDF	92.16%
Our Proposed Method	GAT + TF-IDF	93.06%

According to the findings presented in Table 7, the advocated model (GAT + TF-IDF) exhibited the most optimal performance with a macro-F1 value of 93.06%, exceeding all three comparative

Vol. 6, No. 6, December 2025, Page. 5865-5881
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.6.5482

models. The CNN + TF-IDF model secured the second position with a Macro-F1 of 92.16%, followed by Logistic Regression + TF-IDF at 84.89%, and SVM + TF-IDF at 85.73%. These findings substantiate that the integration of Graph Attention Network with TF-IDF representations is capable of capturing semantic correlations among tweets in a more profound manner than traditional vector-based or sequence-based models. The graph-based methodology furnishes a more nuanced context to the interrelations among documents, which substantially enhances the model's capacity to identify self-reported Monkeypox symptoms in social media textual data. Beyond accuracy improvements, this hybrid architecture also offers a computationally efficient alternative to transformer-based approaches, making it suitable for real-time epidemiological surveillance in resource-limited settings. The strong macro-F1 score of 93.06% further demonstrates that GAT's relational attention mechanisms are particularly effective for distinguishing subtle variations in symptom expression, reinforcing its superiority over baseline models in this classification task.

5. CONCLUSION

The investigation successfully established a model for self-reported preliminary identification of Monkeypox symptoms via social media platforms by amalgamating Graph Attention Network (GAT) and Term Frequency-Inverse Document Frequency (TF-IDF) as feature representations. The dataset comprises approximately 3,200 tweets in the Indonesian language, which were procured through web crawling techniques utilizing Tweet-Harvest and have undergone meticulous labeling and pre-processing phases, including data cleansing, case normalization, tokenization, elimination of stopwords, and filtering of short tokens.

The experimental findings indicated that the integration of TF-IDF bigram, with a minimum document frequency (min_df) of 1, and K set to 9 in the Graph Construction yielded the most optimal configuration, achieving a macro-F1 score of 93.06% and an accuracy rate of 93.07%. The model demonstrated its capacity to sustain training stability, exhibiting no signs of overfitting, as evidenced by a loss curve that consistently declined until it converged at the 140th epoch. Furthermore, comparative analyses with alternative models (SVM, Logistic Regression, and CNN) illustrated that GAT outperformed all others, thereby validating the efficacy of graph-based methodologies in capturing the semantic interrelations among tweets.

Moreover, this study contributes to the field of computer science and digital epidemiology by demonstrating that a TF-IDF GAT hybrid model can serve as a computationally efficient alternative to transformer-based architectures. Its ability to leverage graph relational structures enables scalable, realtime surveillance suitable for low-resource settings and diverse linguistic environments, positioning the approach as a practical tool for public health monitoring during rapidly evolving outbreaks.

In summary, this research not only significantly advances the formulation of early detection systems predicated on social media analytics, but also advances the integration of classical NLP representations with graph neural architectures, strengthening the methodological foundation for real-time infectious disease surveillance, which can be employed by health institutions for the prompt surveillance of infectious disease trends. For future investigations, it is advisable to incorporate contextual embeddings such as BERT or IndoBERT during the feature representation phase, as well as to examine semi-supervised node classification techniques to enhance the model's generalizability across broader and more diverse datasets.

ACKNOWLEDGEMENT

This research was supported by funding from the Directorate of Research, Technology, and Community Service, Ministry of Higher Education, Science, and Technology of the Republic of

Vol. 6, No. 6, December 2025, Page. 5865-5881
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.6.5482

Indonesia. Main Contract Number: 075/C3/DT.05.00/PL/2025, Derivative Contract Number: 1396/UN8.2/PG/2025.

REFERENCES

[1] E. M. Bunge *et al.*, "The changing epidemiology of human monkeypox—A potential threat? A systematic review," *PLoS Negl. Trop. Dis.*, vol. 16, no. 2, pp. 1–20, 2022, doi: 10.1371/journal.pntd.0010141.

[2] J. P. Thornhill *et al.*, "Monkeypox Virus Infection in Humans across 16 Countries — April–June 2022," *N. Engl. J. Med.*, vol. 387, no. 8, pp. 679–691, 2022, doi: 10.1056/nejmoa2207323.

[3] WHO, "Multi-country monkeypox outbreak: situation update." Accessed: Nov. 28, 2025. [Online]. Available: https://www.who.int/emergencies/disease-outbreak-news/item/2022-DON396

[4] H. Adler *et al.*, "Clinical features and management of human monkeypox : a retrospective observational study in the UK," vol. 22, no. 8, pp. 1153–1162, 2022, doi: 10.1016/S1473-3099(22)00228-6.

[5] A. Y. Cheema, O. J. Ogedegbe, M. Munir, G. Alugba, and T. K. Ojo, "Monkeypox : A Review of Clinical Features , Diagnosis , and Treatment," vol. 14, no. 7, pp. 14–17, 2022, doi: 10.7759/cureus.26756.

[6] C. Raina MacIntyre *et al.*, "Early detection of emerging infectious diseases - implications for vaccine development," *Vaccine*, vol. 42, no. 7, pp. 1826–1830, 2024, doi: 10.1016/j.vaccine.2023.05.069.

[7] R. Meckawy, D. Stuckler, A. Mehta, T. Al-Ahdal, and B. N. Doebbeling, "Effectiveness of early warning systems in the detection of infectious diseases outbreaks: a systematic review," *BMC Public Health*, vol. 22, no. 1, pp. 1–62, 2022, doi: 10.1186/s12889-022-14625-4.

[8] E. Chen, K. Lerman, and E. Ferrara, "Tracking social media discourse about the COVID-19 pandemic: Development of a public coronavirus Twitter data set," *JMIR Public Heal. Surveill.*, vol. 6, no. 2, Apr. 2020, doi: 10.2196/19273.

[9] E. Du, E. Chen, J. Liu, and C. Zheng, "How do social media and individual behaviors affect epidemic transmission and control," no. January, 2020, doi: 10.1016/j.scitotenv.2020.144114.

[10] B. Shi, W. Huang, Y. Dang, and W. Zhou, "Leveraging social media data for pandemic detection and prediction," *Humanit. Soc. Sci. Commun.*, vol. 11, no. 1, 2024, doi: 10.1057/s41599-024-03589-y.

[11] J. Khan, K. Ahmad, S. K. Jagatheesaperumal, and K. A. Sohn, "Textual variations in social media text processing applications: challenges, solutions, and trends," *Artif. Intell. Rev.*, vol. 58, no. 3, Mar. 2025, doi: 10.1007/s10462-024-11071-z.

[12] M. Rodríguez-Ibánez, A. Casánez-Ventura, F. Castejón-Mateos, and P. M. Cuenca-Jiménez, "A review on sentiment analysis from social media platforms," Aug. 01, 2023, *Elsevier Ltd*. doi: 10.1016/j.eswa.2023.119862.

[13] S. Park, S. Oh, and W. Park, "Automated Classification Model for Elementary Mathematics Diagnostic Assessment Data Based on TF-IDF and XGBoost," *Appl. Sci.*, vol. 15, no. 7, Apr. 2025, doi: 10.3390/app15073764.

[14] K. Li, "Haha at fakedes 2021: A fake news detection method based on tf-idf and ensemble machine learning," *CEUR Workshop Proc.*, vol. 2943, no. September, pp. 630–638, 2021.

[15] P. Pilipiec, I. Samsten, and A. Bota, *Surveillance of communicable diseases using social media: A systematic review*, vol. 18, no. 2 February. 2023. doi: 10.1371/journal.pone.0282101.

[16] M. Das, S. Kamalanathan, and P. Alphonse, "A Comparative Study on TF-IDF feature weighting method and its analysis using unstructured dataset," *CEUR Workshop Proc.*, vol. 2870, pp. 98–107, 2021.

[17] Sutriawan, S. Rustad, G. F. Shidik, and Pujiono, "Performance Evaluation of Text Embedding Models for Ambiguity Classification in Indonesian News Corpus: A Comparative Study of TF-IDF, Word2Vec, FastText BERT, and GPT," *Ing. des Syst. d'Information*, vol. 30, no. 6, pp. 1469–1482, 2025, doi: 10.18280/isi.300606.

[18] S. A. Sazan, M. H. Miraz, and A. B. M. Muntasir Rahman, "Enhancing Depressive Post Detection

Vol. 6, No. 6, December 2025, Page. 5865-5881
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.6.5482

in Bangla: A Comparative Study of TF-IDF, BERT and FastText Embeddings," *Ann. Emerg. Technol. Comput.*, vol. 8, no. 3, pp. 34–49, 2024, doi: 10.33166/AETiC.2024.03.003.

[19] V. Rai and S. Rai, "Attention Mechanisms in Graph Neural Networks for Fake News Detection: A Critical Review and Open Issues," *Researchgate.Net*, no. February 2024, 2025, doi: 2024/IJEASM/5/2024/1996a.

[20] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, and P. S. Yu, "A Comprehensive Survey on Graph Neural Networks," Dec. 2020, doi: 10.1109/TNNLS.2020.2978386.

[21] K. Wang, Y. Ding, and S. C. Han, "Graph neural networks for text classification: a survey," *Artif. Intell. Rev.*, vol. 57, no. 8, Aug. 2024, doi: 10.1007/s10462-024-10808-0.

[22] S. Brody, U. Alon, and E. Yahav, "How Attentive Are Graph Attention Networks?," *ICLR 2022 - 10th Int. Conf. Learn. Represent.*, pp. 1–26, 2022, doi: 10.48550/arXiv.2105.14491.

[23] J. Li, Y. Jian, and Y. Xiong, "Text Classification Model Based on Graph Attention Networks and Adversarial Training," *Appl. Sci.*, vol. 14, no. 11, Jun. 2024, doi: 10.3390/app14114906.

[24] A. Malik, D. K. Behera, J. Hota, and A. R. Swain, "Ensemble graph neural networks for fake news detection using user engagement and text features," *Results Eng.*, vol. 24, Dec. 2024, doi: 10.1016/j.rineng.2024.103081.

[25] S. A. Zikrina and Fitriyani, "Advancing Hate Speech Detection in Indonesian Language Using Graph Neural Networks and TF-IDF," *J. RESTI*, vol. 9, no. 1, pp. 137–145, Feb. 2025, doi: 10.29207/resti.v9i1.6179.

[26] E. Gao, H. Yang, D. Sun, H. Xia, Y. Ma, and Y. Zhu, "Text Classification Optimization Algorithm Based on Graph Neural Network," *2024 IEEE 6th Int. Conf. Power, Intell. Comput. Syst. ICPICS 2024*, pp. 814–822, 2024, doi: 10.1109/ICPICS62053.2024.10796365.

[27] B. Nath, D. Sahoo, and S. S. Patra, "Leveraging Hybrid Model for Classification of Disaster-Related Tweets using TF-IDF and GCN," *Nanotechnol. Perceptions*, vol. 20, no. 12, pp. 52–72, 2024, doi: 10.62441/nano-ntp.v20is12.4.

[28] Y. Liscano, L. A. Anillo Arrieta, J. F. Montenegro, D. Prieto-Alvarado, and J. Ordoñez, "Early Warning of Infectious Disease Outbreaks Using Social Media and Digital Data: A Scoping Review," *Int. J. Environ. Res. Public Health*, vol. 22, no. 7, pp. 1–34, 2025, doi: 10.3390/ijerph22071104.

[29] D. Chicco and G. Jurman, "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation," *BMC Genomics*, vol. 21, no. 1, Jan. 2020, doi: 10.1186/s12864-019-6413-7.

[30] N. A. Rahmi, S. Defit, and Okfalisa, "The Use of Hyperparameter Tuning in Model Classification: A Scientific Work Area Identification," *Int. J. Informatics Vis.*, vol. 8, no. 4, pp. 2181–2188, 2024, doi: 10.62527/joiv.8.4.3092.

[31] M. Giancotti, M. Lopreite, M. Mauro, and M. Puliga, "Innovating health prevention models in detecting infectious disease outbreaks through social media data: an umbrella review of the evidence," *Front. Public Heal.*, vol. 12, no. November, 2024, doi: 10.3389/fpubh.2024.1435724.

[32] P. Kumar and K. Garg, "Data Cleaning of Raw Tweets for Sentiment Analysis," pp. 273–276, 2020, doi: 10.1109/Indo-TaiwanICAN48429.2020.9181326.

[33] S. Shevira, I. M. Agus, D. Suarjaya, and P. Wira, "Pengaruh Kombinasi dan Urutan Pre-Processing pada Tweets Bahasa Indonesia," vol. 3, no. 2, 2022, doi: 10.24843/JTRTI.2022.v03.i02.p06.

[34] P. Prihatini, K. Indah, G. Sukerti, I. Indrayana, and I. Sudiartha, "Feature Extraction Performance on Classified Methods for Text Sentiment Analysis," pp. 1235–1243, 2023, doi: 10.5220/0010962900003260.

[35] Y. Zhang, Y. Zhou, and J. T. Yao, "Feature Extraction with TF-IDF and Game-Theoretic Shadowed Sets," *Commun. Comput. Inf. Sci.*, vol. 1237 CCIS, pp. 722–733, 2020, doi: 10.1007/978-3-030-50146-4_53.

[36] P. Guleria, J. Frnda, and P. N. Srinivasu, "NLP based text classification using TF-IDF enabled fine-tuned long short-term memory: An empirical analysis," *Array*, vol. 27, no. July, 2025, doi: 10.1016/j.array.2025.100467.

[37] A. Nazarkar, H. Kuchulakanti, C. S. Paidimarry, and S. Kulkarni, *Impact of Various Data Splitting Ratios on the Performance of Machine Learning Models in the Classification of Lung*

Vol. 6, No. 6, December 2025, Page. 5865-5881
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.6.5482

Cancer, vol. 1. Atlantis Press International BV, 2023. doi: 10.2991/978-94-6463-252-1_12.

[38] L. Li, W. Yang, S. Bai, and Z. Ma, "KNN-GNN: A powerful graph neural network enhanced by aggregating K-nearest neighbors in common subspace," *Expert Syst. Appl.*, vol. 253, no. May, 2024, doi: 10.1016/j.eswa.2024.124217.