



Sentiment Analysis of X (Twitter) Comments on The Influence of South Korean Culture in Indonesia

Putu Rheya Ananda Savitri¹, I Made Agus Dwi Suarjaya²,
Wayan Oger Vihikan³

^{1,2,3}Departement of Information Technology, Udayana University, Bali, Indonesia
Email: ¹rheya.ananda@gmail.com, ²agussuarjaya@it.unud.ac.id, ³oger_vihikan@unud.ac.id

Abstract

Hallyu or Korean wave refers to the phenomenon of South Korean values and culture spreading to other countries, ultimately influencing global culture. South Korean culture, such as K-pop music, dramas, films, fashion, food, and lifestyle, has gained popularity in Indonesia since 2002. Because South Korean culture influences many aspects of life in Indonesia, responses to this Korean wave are widely discussed in social media, especially through X (Twitter) ranging from positive sentiment to negative sentiment. To gain a more in-depth and detailed understanding of public opinion, a classification process was conducted on the social media platform X (Twitter) using a deep learning algorithm based on the CNN method. The results of this classification provide more accurate and informative insight into the attitudes, opinions, and reactions of the Indonesian people towards the influence of South Korean culture in this country. The research was conducted using 717,998 tweet data resulting in an accuracy of 79%.

Keywords: Korean wave, Convolutional Neural Network, X (Twitter), Sentiment

1. INTRODUCTION

Social media is currently developing very rapidly, making it easier for users to share their thoughts and opinions on online platforms. The easier it is to access social media, can led to the emergence of new trends in interacting and communicating so that social media users have increased, one of the most popular one is X (Twitter), where according to the databoks website, 500 million tweets are posted every day. In 2023 there will be a total of 1.3 billion X accounts and 368.4 million of them will be active users. The number of X users in Indonesia in 2023 will also increase by 30% from the previous year, namely 24 million, placing Indonesia in fifth place as the country with the largest X users in the world [1].

One of X's content that is increasingly in demand is related to South Korea. The influence of South Korean culture in Indonesia has grown rapidly in recent years, which can be said to be a symptom of the spread of certain values and culture of



a country to other parts of the world so that it becomes world culture, this is known as the Korean wave or Hallyu. Some cultural products that are popular in Indonesian society are music (Korean pop), drama, film, fashion, food, and lifestyle [2].

The Korean wave initially arrived in Indonesia in 2002 when private TV stations aired the Korean dramas "Winter Sonata" and "Endless Love." Due to their high ratings, during the first half of the 2000s following the World Cup in South Korea and Japan, nearly all private TV stations began broadcasting numerous other K-dramas. Korean dramas and films have brought new phenomena to influence the Indonesian market. In early 2009, Korean songs used for K-drama and K-movie soundtracks became known to Indonesians as Boy Bands and Girl Bands became increasingly popular [3].

Due to the significant influence of South Korean culture on various aspects of life in Indonesia, it is inevitable that the entry of the Korean wave into the country has generated a range of opinions, both supportive and opposing. This cultural influx has sparked diverse reactions among the Indonesian populace, reflecting differing perspectives on the impact of South Korean culture. Some see the entry of the Korean wave into Indonesia as a positive opportunity to increase the creativity of the Indonesian fashion and entertainment industry. On the other hand, there are those who are worried that this Korean trend will have a big influence on the younger generation's appreciation of local culture [4]. Therefore, this can be used as a data source in analyzing Indonesian public sentiment so that it can reveal various public opinions and attitudes towards the Korean wave.

Sentiment analysis has been developed using machine learning and deep learning methods. However, generally the sentiment analysis carried out still uses machine learning methods where the accuracy is still not optimal. Therefore, deep learning algorithms are used to obtain analysis results with high accuracy [5]. Tweets will be retrieved (crawled) using several keywords related to the topic discussed, then classified using the CNN method.

To obtain a more in-depth and detailed opinion perspective, a process of classifying public opinion was carried out through the X social media platform by applying a deep learning algorithm using the CNN method. This technique allows us to analyze tweets that have been retrieved (crawled) using keywords related to the topic being discussed. The results of this classification will provide more accurate and informative insight into the attitudes, opinions, and reactions of the Indonesian people towards the influence of the Korean wave in this country.

2. METHODS

This research uses 5 pre-processing stages, namely case folding, removing username, removing symbols, normalization, and tokenizing [6]. Here's the process.

- 1) Case Folding: The process of changing all characters to lower case. This is done because capital letters will greatly influence the data analysis process.
- 2) Remove Usernames: Used to remove the @ symbol along with the username in Twitter comments because the analysis only requires the text content of the tweet.
- 3) Remove Symbols: Used to delete all symbols, hashtags, and numbers in tweets so that the text representation becomes uniform [7].
- 4) Normalization: Carried out to clean the text, there are 2 stages, changing non-standard words, slangs, also misspelling and removing unclear words that are not in the KBBI [8].
- 5) Tokenizing: The process of breaking down text into individual words by space and assigning a token to each word [9] .

Figure 1 will explain the process used in this research through a general description of the research.

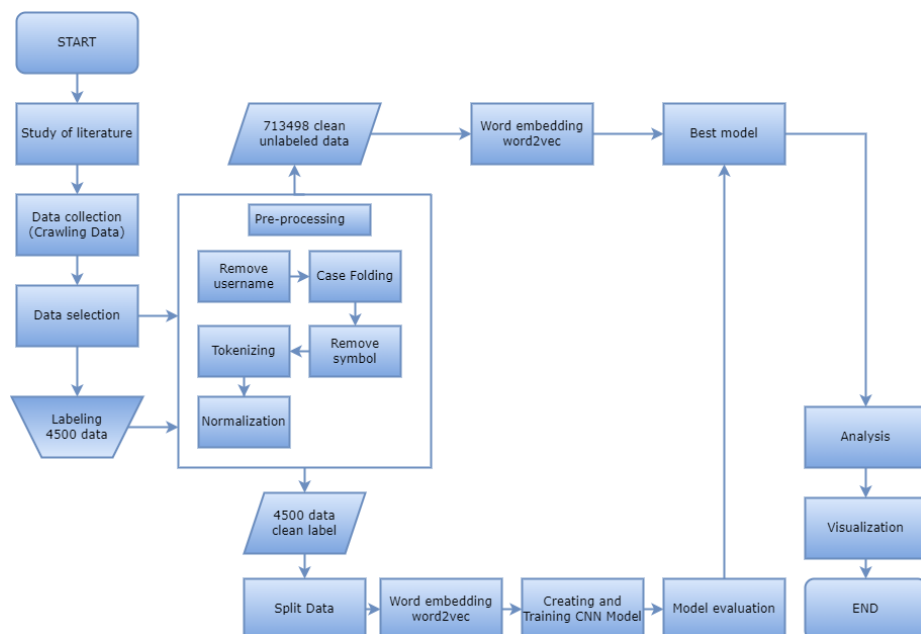


Figure 1. Research Process Flow

Figure 1 explains the flow of the research process carried out, first conducting a literature study through several papers with material related to this research. After that, continue to collect data where this data is taken from social media X (Twitter) which is related to South Korean culture and with the help of tweet harvest. After collecting data from X (Twitter), a selection process is carried out to divide the data. At this stage, manual labeling is performed on 4,500 tweets. Then, both labeled and unlabeled data will proceed to the pre-processing stage where the data will be cleaned using the 5 stages in pre-processing. After that, the data is split only on data that has been given a label. The model will be trained on clean labeled data consisting of 4,500 tweets. The data is split using four different data split scenarios to find the best scenario. Then, word embedding is carried out using pre-trained Word2vec on the data that has been trained with the best scenario. The next stage is creating a CNN model for training and classifying public sentiment. Model evaluation will be conducted to determine which data split scenario yields the highest accuracy. Finally, the analysis results will be visualized using Tableau tools to make them easier to understand.

2.1. Data Collection

The data collection in tweets from social media X (Twitter) obtained by crawling data using several keywords related to South Korean culture. This was carried out using Google Colaboratory. First, install the required libraries, then select the related keywords. Second, when using tweet harvest, we need to enter the auth token from account X (Twitter) which will be used to retrieve data, make sure the account is not private [10]. Then the data will be crawled, and it will get the desired X (Twitter) data. It is amounted to 717,998 data and was collected over 4 years from 2020 to 2023, then stored in a MongoDB database for further analysis.

2.2. Data Pre-processing

The pre-processing phase encompassed multiple stages aimed at refining the tweet data. This involved removing usernames, case folding, removing symbols, normalization, and tokenizing. Each of these steps contributed to the refinement and organization of the tweet dataset, ensuring it was ready for further analysis. Following are the results of each pre-processing stage.

Tabel 1. Data Pre-processing

Process	Text (Indonesia)	Result (Indonesia)
Remove Username	@kyungsooddudu12 Wkkw gak selesai2 masalah KPop artist nih	Wkkw gak selesai2 masalah KPop artist nih
	ÃƒÅ°ÃƒÅ„ÃƒÅ±ÃƒÅ±ÃƒÅ±	ÃƒÅ°ÃƒÅ„ÃƒÅ±ÃƒÅ±ÃƒÅ±

Process	Text (Indonesia)	Result (Indonesia)
Case Folding	Wkkw gak selesai2 masalah KPop artist nih ÃƒÂ°ÃƒÂ,ÃƒÂ,ÃƒÂ,ÃƒÂ,ÃƒÂ	wkkw gak selesai2 masalah kpop artist nih ÃƒÂ°ÃƒÂ,ÃƒÂ,ÃƒÂ,ÃƒÂ,ÃƒÂ
Remove Symbols	wkkw gak selesai2 masalah kpop artist nih ÃƒÂ°ÃƒÂ,ÃƒÂ,ÃƒÂ,ÃƒÂ,ÃƒÂ	wkkw gak selesai2 masalah kpop artist nih
Normalization	wkkw gak selesai2 masalah kpop artist nih	tidak selesai masalah kpop artist nih
Tokennizing	tidak selesai masalah kpop artist nih	[tidak, 'selesai', 'masalah', 'kpop', 'artist', 'nih']

2.3. Data Labeling

This research uses 3 classes, positive, netral, and negative. The amount of data labeled is 4500 data with details of 1500 data with negative labels, 1500 data with positive labels, and 1500 data with neutral labels. It's important to note that these labels were assigned manually, ensuring meticulous attention to each tweet's sentiment classification. This balanced distribution ensures that the dataset maintains uniform representation across all three sentiment classes.

2.4. CNN Model

Convolutional Neural Network (CNN) is a type of neural network designed to process data with a grid-like structure. The term "Convolutional Neural Network" reflects the use of a mathematical operation called convolution. In CNNs, this convolution operation replaces the standard matrix multiplication typically used in neural network layers[11]. CNN was initially only used for image processing, but in 2015 CNN began to be used for text classification. CNN is an application of deep learning models that is different from machine learning methods. Judging from the execution time, deep learning models take longer than conventional machine learning due to the calculation of large weights and the addition of parameters. The main contribution of the CNN algorithm is in the convolutional and pooling layers [12]. Convolutional layers are used to extract semantic information contained in a sentence, convolutional layers in CNN are usually followed by other layers such as max pooling layers or average pooling layers [13]. This layer aims to reduce the dimensions of the features produced by the convolutional layer, thereby reducing model complexity and speeding up the training process [14]. Figure 2 illustrates the CNN architecture developed for this research.

Model: "sequential_2"

Layer (type)	Output Shape	Param #
embedding_2 (Embedding)	(None, 100, 200)	1945600
conv1d_2 (Conv1D)	(None, 99, 600)	240600
max_pooling1d_2 (MaxPooling1D)	(None, 49, 600)	0
dropout_2 (Dropout)	(None, 49, 600)	0
global_max_pooling1d_2 (GlobalMaxPooling1D)	(None, 600)	0
dense_2 (Dense)	(None, 3)	1803
Total params: 2188003 (8.35 MB)		
Trainable params: 2188003 (8.35 MB)		
Non-trainable params: 0 (0.00 Byte)		

Figure 2. CNN Architecture

Figure 2 is the architecture of a CNN model for training the data using TensorFlow and Keras. A pre-trained Word2Vec with an embedding dimension of 200 is included in the model. The data used consists of 4500 labeled samples. The model is built by adding an embedding layer as the input layer, a 1D convolutional layer, a max pooling layer to capture important features representing other features, a global max pooling layer to summarize all information, a dropout layer to deactivate some units during training with a probability of 0.8, and a dense (fully connected) layer used for the classification process, resulting in predictions for three classes, positive, negative, and netral [15].

2.5. Word2vec

Word2Vec is a way of representing text in dense vector form. Word2vec, with its capacity ranging from 50 to 300 dimensions, emerged as a prominent tool in the field of natural language processing around 2013 [16], because it can represent relationships between words better compared to TF-IDF, both semantically and syntactically. This representation of words in the form of mathematical vectors helps many computer learning programs to better process language by grouping words with similar meanings, which has been applied in various natural language processing tasks [17]. This research applies Word2Vec to represent words in the training data. The data used has undergone pre-processing. The Word2Vec utilized is a pre-trained model in the Indonesian language obtained from GitHub with a dimension of 200 and employs the Skip-gram method for training word vectors.

3. RESULTS AND DISCUSSION

Based on the process that started with data crawling and continued through to model training, the results obtained are as follows. Model training is a process for training a model that will be used on tweets data. This research uses 4 data splitting scenarios, the aim is to find data splitting with the best results. The following is a comparison of the data splitting scenarios used.

Table 2. Splitting Data Scenarios

Scenario	Accuracy	Precision	Recall	F1-Score	Validation Accuracy
90:5:5	79%	61%	61%	61%	68%
80:10:10	78%	59%	58%	58%	57%
70:20:10	79%	56%	57%	56%	59%
60:20:20	77%	58%	58%	57%	55%

Table 2 presents a comparison for each data splitting scenario. The objective is to identify the data sharing method that yields the best results. According to the table the optimal scenario is the 90:5:5 split, which achieves the highest accuracy compared to the other scenarios, making it the optimal choice. In contrast, the 60:20:20 split results in the lowest accuracy at 77%. Figure 2 illustrates the confusion matrix results for the best scenario.



Figure 2. Train dan Validation Graph 90:5:5 Scenario

Figure 2 provides a comprehensive overview of the model's training progress. It encapsulates the trends observed in train accuracy, train loss, validation accuracy,

are some groups who do not like the influence of South Korean culture entering Indonesia. 38.82% were neutral towards South Korean culture, reflecting that some groups did not have strong positive or negative feelings.

Trends in Public Satisfaction with the Influence of South Korean Culture in Indonesia

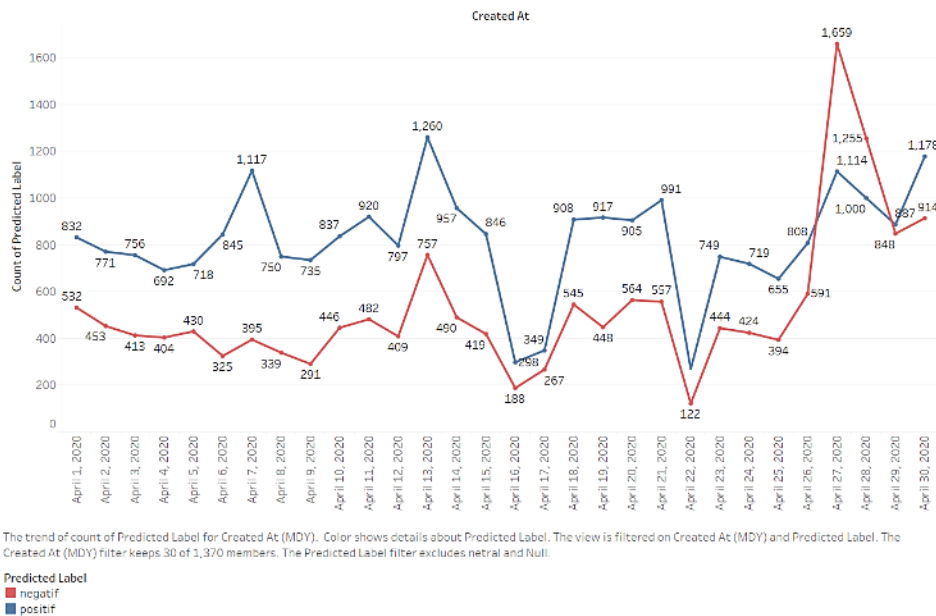


Figure 4. Visualization of Public Satisfaction Trends

Figure 4 is a visualization of trends in public satisfaction with the influence of South Korean culture in Indonesia. It can be seen that in April 2020 there was an increase in negative sentiment, this was because fans of the Korean boy group BTS, ARMY, attacked the account of Filipino TikTok artist Reemar. This was allegedly because Reemar allegedly didn't like the Korean boy group, causing anger from ARMYs who then attacked Reemar's social media accounts, each fan was carrying out mutual account reporting, Reemar's fans reported BTS's account and ARMYs reported Reemar's account causing Reemar had to close his TikTok account. Many negative sentiments were expressed by both fan camps. The visualization in word cloud form is as follows.

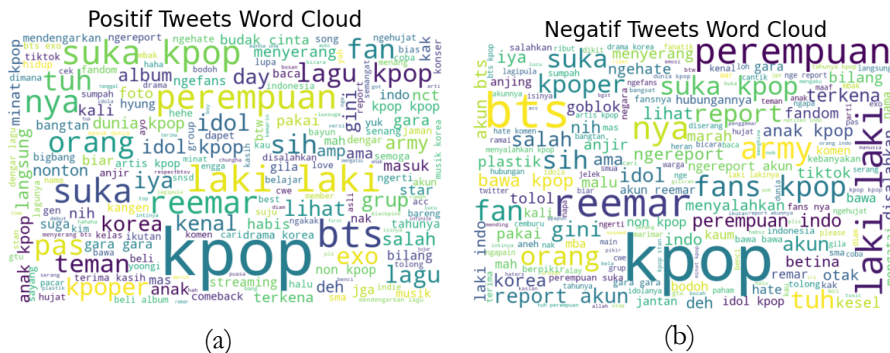


Figure 5. Word Cloud in April 2020 (a) positive tweets (b) negative tweets

Figure 5 is a visualization in the form of a word cloud which displays the words that appear most frequently in April 2023. This word cloud will display an overview of the topics that are the center of public attention at that time. Words like “reemar”, “BTS”, “kpop fans”, “report account” became dominant due to the conflict involving Filipino artist Reemar and Korean boy group BTS. The conflict that caused both fan camps to report each other's accounts caused the word "report account" to also become dominant. The word "kpop fans" appears dominantly indicating that the topic discussed is not only specific to ARMY, but also seems to extend to other Kpop fans.

The findings show that the 90:5:5 data split scenario is the most effective in this research, achieving the highest accuracy and outperforming other scenarios in overall performance metrics. The visualization of sentiment trends over four years provides valuable insights into public opinion towards South Korean culture in Indonesia. Positive sentiment is dominant, but significant negative and neutral sentiments also exist, highlighting a diverse range of opinions. The analysis of the conflict in April 2020 highlights the influence of specific events on public sentiment. The BTS-Reemar incident exemplifies how social media conflicts can significantly influence sentiment trends. This incident emphasizes the importance of understanding the context behind sentiment data, as particular events can affect overall sentiment analysis results.

This research offers several contributions. Methodologically, it demonstrates the effectiveness of different data splitting scenarios in sentiment analysis, providing a basis for selecting optimal data splits in similar research. Culturally, the analysis of sentiments towards South Korean culture in Indonesia provides valuable insights into cultural exchanges and public opinion, which can guide cultural diplomacy and media strategies. Additionally, the study underscores the impact of specific events on public sentiment, highlighting the importance of context-aware sentiment analysis in social media research.

Despite its valuable contributions, this research has certain limitations. The study relies solely on text data from social media, which may not capture the sentiments of individuals who do not actively engage in online discussions or use platforms like Twitter. The analysis covers a specific period (2020-2023), which may not capture longer-term trends or changes in sentiment beyond this timeframe. Future research could address these limitations by expanding the data sources and extending the analysis period to provide a more detailed and comprehensive analysis of public sentiment towards South Korean culture in Indonesia.

4. CONCLUSION

In conclusion, this research shows that Indonesians generally have a positive view of South Korean culture, often called the Korean wave or Hallyu. Specifically, 40.88% of Indonesian Twitter users express positive sentiments towards South Korean culture, while 20.30% hold negative views. Notably, 38.82% remain neutral, showcasing a diverse range of attitudes. However, this sentiment changes depending on what's happening at the time. Social media, especially Twitter, is crucial in spreading people's opinions, giving us valuable insights into how Indonesians feel about cultural globalization. This analysis conducted using a CNN model, achieved a 72% accuracy rate in classifying sentiments into positive, neutral, and negative categories. This study helps us better understand how cultural influences shape Indonesian society and highlights the importance of ongoing exploration in this area for fostering understanding and dialogue in our globalized world.

REFERENCES

- [1] C. M. Annur, "Pengguna Twitter di Indonesia Capai 24 Juta hingga Awal 2023, Peringkat Berapa di Dunia?," *databoks.katadata.co.id*, 2023.
- [2] S. R. Jannah, Z. Khoirunnisa, and A. R. Faristiana, "Pengaruh Korean Wave Dalam Fashion Style Pada Remaja Di Indonesia," *Ekon. Bisnis dan Manaj.*, vol. 1, no. 3, pp. 11–20, 2023, doi: 10.59024/jipa.v1i3.219
- [3] I. P. Putri, F. D. P. Liany, and R. Nuraeni, "K-Drama dan Penyebaran Korean Wave di Indonesia," *ProTVF*, vol. 3, no. 1, p. 68, 2019, doi: 10.24198/ptvf.v3i1.20940.
- [4] L. A. Putri, U. Islam, N. Sultan, S. Kasim, and A. Info, "Dampak Korea Wave Terhadap Prilaku Remaja," *E-Journal uin, Univ. Islam Negeri Sultan Syarif Kasim Riau*, vol. 3, no. 1, pp. 42–48, 2020.
- [5] D. J. M. Pasaribu, K. Kusriani, and S. Sudarmawan, "Peningkatan Akurasi Klasifikasi Sentimen Ulasan Makanan Amazon dengan Bidirectional LSTM dan Bert Embedding," *Inspir. J. Teknol. Inf. dan Komun.*, vol. 10, no. 1, pp. 9–20, 2020, doi: 10.35585/inspir.v10i1.2568.
- [6] P. L. Parameswari and Prihandoko, "Penggunaan Convolutional Neural Network Untuk Analisis Sentimen Opini Lingkungan Hidup Kota Depok

- Di Twitter,” *J. Ilm. Teknol. dan Rekayasa*, vol. 27, no. 1, pp. 29–42, 2022, doi: 10.35760/tr.2022.v27i1.4671.
- [7] A. Srivastava, V. Singh, and G. S. Drall, “Sentiment Analysis of Twitter Data,” *Int. J. Healthc. Inf. Syst. Informatics*, vol. 14, no. 2, pp. 1–16, 2019, doi: 10.4018/ijhisi.2019040101.
- [8] Ni Made Tara Okta Adriana, I Made Agus Dwi Suarjaya, and Dwi Putra Githa, “Analisis Sentimen Publik Terhadap Aksi Demonstrasi di Indonesia Menggunakan Support Vector Machine Dan Random Forest,” *Decod. J. Pendidik. Teknol. Inf.*, vol. 3, no. 2, pp. 257–267, 2023, doi: 10.51454/decode.v3i2.187.
- [9] A. R. Isnain, N. S. Marga, and D. Alita, “Sentiment Analysis Of Government Policy On Corona Case Using Naive Bayes Algorithm,” *IJCCS (Indonesian J. Comput. Cybern. Syst.)*, vol. 15, no. 1, p. 55, 2021, doi: 10.22146/ijccs.60718.
- [10] F. M. J. M. Shamrat *et al.*, “Sentiment analysis on twitter tweets about COVID-19 vaccines using NLP and supervised KNN classification algorithm,” *Indones. J. Electr. Eng. Comput. Sci.*, vol. 23, no. 1, pp. 463–470, 2021, doi: 10.11591/ijeecs.v23.i1.pp463-470.
- [11] I. G. and Y. B. and A. Courville, *Deep Learning*. 2016. [Online]. Available: <https://www.deeplearningbook.org/>
- [12] Y. Yuliska, D. H. Qudsi, J. H. Lubis, K. U. Syaliman, and N. F. Najwa, “Analisis Sentimen pada Data Saran Mahasiswa Terhadap Kinerja Departemen di Perguruan Tinggi Menggunakan Convolutional Neural Network,” *J. Teknol. Inf. dan Ilmu Komput.*, vol. 8, no. 5, p. 1067, 2021, doi: 10.25126/jtiik.2021854842.
- [13] A. K. Jadon and S. Kumar, “Emotion detection using Word2Vec and convolution neural networks,” *Indones. J. Electr. Eng. Comput. Sci.*, vol. 33, no. 3, pp. 1812–1819, 2024, doi: 10.11591/ijeecs.v33.i3.pp1812-1819.
- [14] T. Shoryu, L. Wang, and R. Ma, “A Deep Neural Network Approach using Convolutional Network and Long Short Term Memory for Text Sentimen Classification,” *IEEE Access*, pp. 763–768, 2021.
- [15] T. S. Sajana, S. Jacob, P. Vinod, V. G. Menon, and P. S. Shilpa, “Context-aware gender and age recognition from smartphone sensors,” *Proc. Int. Conf. Comput. Commun. Secur. Intell. Syst. IC3SIS 2022*, no. September, 2022, doi: 10.1109/IC3SIS54991.2022.9885610.
- [16] H. Wang, “Word2Vec and SVM Fusion for Advanced Sentiment Analysis on Amazon Reviews,” *Highlights Sci. Eng. Technol.*, vol. 85, pp. 743–749, 2024, doi: 10.54097/sw4pft19.
- [17] H. Juwiantho *et al.*, “Sentiment Analysis Twitter Bahasa Indonesia Berbasis Word2vec Menggunakan Deep Convolutional Neural Network,” *J. Teknol. Inf. dan Ilmu Komput.*, vol. 7, no. 1, pp. 181–188, 2020, doi: 10.25126/jtiik.202071758.
- [18] F. A. Irawan and D. A. Rochmah, “Penerapan Algoritma CNN Untuk Mengetahui Sentimen Masyarakat Terhadap Kebijakan Vaksin Covid-19,”

- J. Inform.*, vol. 9, no. 2, pp. 148–158, 2022, doi: 10.31294/inf.v9i2.13257.
- [19] Vidya Chandradev, I Made Agus Dwi Suarjaya, and I Putu Agung Bayupati, “Analisis Sentimen Review Hotel Menggunakan Metode Deep Learning BERT,” *J. Buana Inform.*, vol. 14, no. 02, pp. 107–116, 2023, doi: 10.24002/jbi.v14i02.7244.