JITK (JURNAL ILMU PENGETAHUAN DAN TEKNOLOGI KOMPUTER)

VOL. 10. NO. 2 NOVEMBER 2024
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v10i2.5187

SOCIAL MEDIA COMMENTS FOR GOVERNMENT INSTITUTION VIDEO CLASSIFICATION USING MACHINE LEARNING

M. Faris Al Hakim[1*]; Subhan[2]; Prasetyo Listiaji[3]

Informatics Engineering Study Program[1,2]
Science Education Study Program[3]
Universitas Negeri Semarang[1,2,3]
https://unnes.ac.id/[1,2,3]
farisalhakim@mail.unnes.ac.id[1*] , subhan@mail.unnes.ac.id[2], p.listiaji@mail.unnes.ac.id[3]

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

Abstract— *YouTube is a social media site that is quite familiar and is used as a means of disseminating video-based information. With a fairly high number of users, YouTube can become a communication medium for audiences, including government agencies. The user's responses in comments reflect the nuance of the presented video. This research aims to determine the best algorithm for classifying video types based on user comments. Several machine learning algorithms used to carry out classification are Decision Tree, Random Forest, K-Nearest Neighbor, Support Vector Machine, and Logistic Regression. K-Fold Cross Validation was chosen as a method to evaluate the performance of classification algorithms based on the accuracy values. of these algorithms in classifying YouTube videos based on comments. The first experiment with the highest ratio of training and test data for each algorithm was obtained at a ratio of 90:10, with respectively 78.99%, 86.21%, 84.01%, 72.72%, and 79.31%. In the second experiment with k-fold cross validation using a ratio of 90:10, the highest accuracy for each algorithm was obtained at a value of k = 10, which was respectively 74.39%, 81.34%, 78.05%, 85.21%, and 72.15%. From these results, it can be concluded that the most suitable algorithm for classifying YouTube videos based on comments is the Random Forest algorithm with a training and test data ratio of 90:10 and SVM with 10-cross-fold validation. These results show that a larger portion of data for learning has a positive impact on algorithm performance.*

Keywords: *government institution, machine learning, public response, response analysis, video social media.*

Intisari— *Youtube merupakan salah satu sosial media yang cukup familiar digunakan sebagai sarana penyebaran informasi berbasis video. Dengan jumlah pengguna yang cukup tinggi, Youtube dapat menjadi media komunikasi bagi khalayak termasuk instansi pemerintah. Respon pengguna pada kolom komentar menjadi cerminan dari nuansa video yang disajikan. Penelitian ini bertujuan untuk menentukan algoritma terbaik dalam melakukan klasifikasi jenis video berdasarkan komentar pengguna. Beberapa algoritma machine learning yang digunakan untuk melakukan klasifikasi yaitu Decision Tree, Random Forest, K-Nearest Neighbor, Support Vector Machine, dan Logistic Regression. K-Fold Cross Validation dipilih sebagai metode untuk mengevaluasi performa algoritma klasifikasi berdasarkan nilai akurasi. terhadap algoritma-algoritma tersebut, dalam mengklasifikasikan video YouTube berdasarkan komentar. Percobaan pertama dengan rasio data latih dan uji akurasi tertinggi masing-masing algoritma diperoleh pada rasio 90:10 dengan secara berturut-turut sebesar 78.99%, 86.21%, 84.01%, 72.72%, dan 79.31%. Percobaan kedua dengan k-fold cross validation menggunakan rasio 90:10 akurasi tertinggi masing-masing algoritma diperoleh pada nilai k=10 yang secara berturut-turut sebesar 74.39%, 81.34%, 78.05%, 85.21%, dan 72.15%. Dari hasil tersebut, dapat diambil kesimpulan bahwa algoritma yang paling sesuai untuk mengklasifikasikan video YouTube berdasarkan komentar adalah algoritma Random Forest dengan rasio data latih dan uji 90:10 dan SVM dengan 10-cross fold validation. Hasil tersebut menunjukkan bahwa porsi data yang lebih banyak untuk belajar memberikan dampak positif terhadap kinerja algoritma.*

Kata Kunci: *institusi pemerintah, pembelajaran mesin, respon masyarakat, analisis respon, video sosial media.*

INTRODUCTION

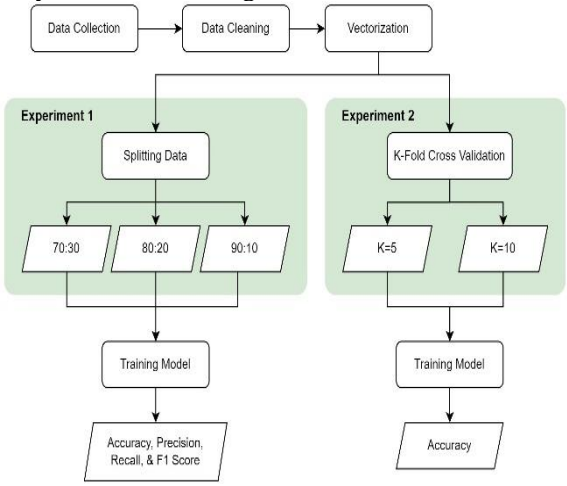
The government is the main element of a country that is responsible for the sustainability of that country. One of the success factors of state institutions is determined by the quality of service to users (society) [1], [2]. Improving the quality of government services is a universal necessity that must be carried out by all state administrators. Various efforts continue to be made to provide the best service to the community. Social media has now become an open space and communication tool for all levels of society. This provides an opportunity for organizations and agencies, including the government, to interact more closely with the community, which is the main object for which services must be provided. The use of social media as a means to disseminate information has a wide impact. YouTube, one of the video-based social media sites, currently has the highest number of users in the world. The number of YouTube users in the world reached 1.9 billion in mid-2019. In Indonesia, the number of YouTube users in 2019 reached 88% of social media users, or around 132 million users. Based on Google Indonesia data, 9 out of 10 people who use the internet in Indonesia watch YouTube every day. YouTube, as a social media platform that is growing in size, has been used in various fields. Several fields such as education [3], [4], tourism [5], health [6], [7], [8], economics [9], [10], government [11], [12], and politics [13] have utilized YouTube social media to strengthen its role well.

Most government agencies in Indonesia have utilized social media in their organizational activities. Ministries and regional governments at provincial and district levels are institutions that are also taking advantage of this potential through YouTube social media. In the health sector, government agencies, namely BPJS, Central General Hospital (RSUP Sardjito Yogyakarta, RSUP Kariadi Semarang, RSUP Soeradji Tirtonegoro Klaten), Regional General Hospital (RSUD), and even health services at the Community Health Center level have also used YouTube. In the education sector, almost all state campuses currently have YouTube accounts as a form of social media. Apart from these fields, most ministry agencies also use YouTube as a medium for socializing with the public. In general, the use of YouTube as a communication channel in an agency, including government, is none other than to optimize services to the community [14].

YouTube, as a social media platform that is currently in demand, also produces large numbers of comments through video uploads. These comments become data that has the potential to produce hidden information (insight) through natural language processing (NLP) technology. NLP refers to the automatic computational processing of human language by taking human-generated text as input and producing output in the form of text [15]. NLP is an artificial intelligence technology that is widely used to improve customer interactions by capturing and analyzing customer voices [16]. Machine learning is an artificial intelligence technology that can be used to support tasks in the NLP field. Several machine learning methods applied in NLP are probabilistic classification, K-Nearest Neighbors (KNN), Neural Network (NN), Support Vector Machine (SVM), Decision Tree (DT), Random Forest (RF), Logistic Regression (LR), LDA (Linear Discriminant Analysis), ensembles, and neural methods [17], [18], [19]. Machine learning is an algorithm that supports a series of processes for completing certain tasks. Based on the description presented in the previous paragraph, this article attempts to answer the research question of how to determine the type of government-owned video based on user comments. This was done because the comments put by the user gave insight into what is contained in the video. This task becomes the initial task to be able to complete subsequent tasks, such as video concept detection. Comments given by users on videos owned by government YouTube channels are the main data source used. The data is then processed using one of the extraction techniques in NLP. The classification process is carried out using several machine learning methods so that the results of each algorithm can be compared.

MATERIAL AND METHODS

The stages of research carried out in this study can be seen in Figure 1.



Source: (Research Results, 2024)

Figure 1. Research Method

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

VOL. 10. NO. 2 NOVEMBER 2024
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v10i2.5187

Base on figure 1, The research stages in this study began with data collection. After the data is collected, the next stage is text processing so that it can be used for the next stage. The next stage is model training, where the processed data will be used to train the model to be able to classify comments from social media.

1. Data Collection

The data used in this research comes from comments on YouTube channels of government agencies and non-government agencies. Data from the government agencies was taken from several ministry youtube channels. The other data was taken from private companies' youtube channels for example Indosat. The categories used for classification are class 1, representing comments aimed at government videos, and class 0, representing comments aimed at non-government videos. The amount of data used is 3000, consisting of 1500 data for class 1 and 2000 data for class 0. The sample data used in this research can be shown in Table I below.

Table 1. A Dataset Sample

Comment	Category
Strongly agree hopefully education in Indonesia will become better and So will the welfare of teachers and non civil servants worker.	1
Big salaries are said to attract the younger generation, the nation's best young men. So that the best generation is not only taken by the private sector because the salaries are higher.	1
Make the G20 Program a success to encourage development and create world peace.	1
Why should analogous tv be eliminated? while so much advancement in nation must be watched by the wider community. They love Indonesia so much with its cultural diversity and beautiful nature.	1
It's funny how our public officials maintain their image. Because they feel they can't get praise, they try not to get criticism.	0
Wow, I am really interested in the appearance of the newest toyota supra, prefer it to the 1993 supra, the newest supra, if I am not mistaken, 2 billion right?.	0
How come there are no oppo stores in lazada.	0
Pull-resistant and durabe, not easily damage, and still good.	0
Amazing, the advertisement is very cool, very inspiring	0

Source: (YouTube Comments, 2024)

2. Data Cleaning

At this stage, the data that has been obtained is filtered based on the conditions in the data. The stage is important to make data in good condition for building the model. There are at least 2 conditions that need to be cleaned, namely meaningless sentences, including foreign terms and non-standard words. Data that contains meaningless sentences or only contains symbols needs to be removed. There is quite a lot of data in this condition, considering that the people who provide comments are people with various characters. People can write various types of comments that are not limited to text in the comments column. Apart from that, data cleaning was also carried out by correcting words that did not match the standard words. These inappropriate words can be caused by the writer using non-standard language, or they can also be caused by a typo. Correction of this condition is carried out by replacing the word with the correct or appropriate word.

3. Word Embedding

Discovered in 2013 by Mikolov, Word2Vec has become a very well-known word embedding technique [20]. Word2vec is often also known as neural word embeddings, which is a feedforward neural network model. Wod2vec relies on local information (the context of each word) to generate vectors that can provide semantic meaning. Wod2vec demonstrates the ability to study linguistic patterns as linear relationships between word vectors [21]. In producing high-dimensional vectors, Word2Vec utilizes shallow neural networks [22]. The architecture used by Word2Vec consists of a hidden layer and a fully connected layer. In the first proposal, Word2Vec had the continuous bag of-words (CBOW) and the continuous skip-gram as the architecture.

In CBOW, the model tries to predict the word in the middle based on the context of the words before and after it [21], [22]. Skip-gram is an architecture that is the opposite of CBOW. Skip-gram tries to predict the word before and after it with the input word itself [21]. Both CBOW and Skip-gram use Windows as kernels to get input and output. The window will shift from start to finish with a predetermined window size. A meaningful mapping for word representation becomes the goal instead of predicting the words correctly [23].

In this research, Word2Vec was used to vectorize for both training and testing data. The parameter setting was done as 100 in vector size, 4 in the window, 5 in the minimal count, and 4 in the workers. The output of a comment after

conversion was 100 vector lengths. The final dataset after pre-processing was in the form of matrix n x 100.

4. Classification Model

Classification is one of the tasks that can be carried out by machine learning. Several studies use classification methods to classify texts. Several methods used in text classification are Decision Tree [24], KNN [24], Naïve-Bayes [24], [25], Random Forest [24], [25], and SVM [24], [25].

It was reasonable that those methods were selected. The decision tree is able to handle non-linear features well in the case of the text. The second method, KNN, does not assume any particular distribution where teks could have ravel distribution. Naive Bayes regards independence among features that influence the minimization or irrelevant features. The other algorithm, Random Forest, could estimate the important features for the final prediction. The last algorithm used was SVM, possibly separating the class optimally using the hyperplane concept.

Decision Tree (DT)

DT is a popular induction algorithm because it has the ability to handle redundant attributes, has high flexibility, low computational costs, and is resistant to noise [26]. However, DT algorithms tend to be unstable. Small changes in the training data can result in large changes in the prediction results [27]. To overcome this, the DT algorithm can be combined with ensemble algorithms such as adaptive boosting [28]. A decision tree (DT) works by creating a decision tree consisting of root, branch, and leaf nodes, where each node represents different features [29].

Random Forest (RF)

RF is an ensemble technique that trains multiple decision trees consistently using bootstrapping, averaging, and bagging [30]. Random Forest works by taking the majority of votes from a number of different decision trees to make a decision. The advantage of this algorithm is that it can achieve a level of precision without overfitting [31]. Random Forest has two tuning parameters, namely ntree and Mtry. The ntree parameter represents the total number of trees, while the Mtry parameter determines the number of features used to divide the nodes [32].

K-Nearest Neighbour (KNN)

KNN is one of the simplest classification methods (a simplified version of the Naïve Bayes method) [33], but it is effective in the field of machine learning. This method groups data by looking at all data points [34]. The KNN method works by classifying data based on the majority of the labels of the k nearest neighbors in the feature space [35]. This means that this method utilizes the concept that data with the same features tends to be in the same class. The k value is a parameter that determines the number of nearest neighbors to consider for data classification. This value is determined based on the data used. A k value that is too small can cause overfitting, while a k value that is too large can cause underfitting [35].

Support Vector Machine (SVM)

SVM is an efficient mathematical computing method for classification [36]. One of the main strengths of SVM is its ability to handle linearly inseparable data [37]. This is done through what is called a kernel trick. Kernels are used to transform data into higher feature spaces, such as linear, polynomial, and radial basis (RBF) kernels. The basic concept of SVM is to perform optimal separation between two classes of data in high-dimensional space. This method will map data to n-dimensional features (n is the number of features). Then, identify the hyperplane or decision boundary that separates one class from another. Determining the hyperplane is done by maximizing the marginal distance between two classes and minimizing errors in classification [37]. The greater the marginal distance, the better the model is at generalizing to data that has never been seen before.

Logistic Regression (LR)

LR is included in the category of supervised learning methods, where the model will learn from data that has been previously labeled [38]. In its implementation, the LR method is often used to group data into categories (classification) [39] and provide binary output [40]. This means that this method is suitable for solving cases in two categories. If the prediction is positive, then the value for the dependent variable is 1, whereas if the prediction is negative, then the value for the dependent variable is 0. The LR method is widely used because it is easy to implement, easy to update, does not make any assumptions about the distribution of the independent variables, and has a good probability interpretation of model parameters [38].

5. Model Evaluation

The resulting model was tested for performance in classification. Measurements of the model are carried out using accuracy, precision, recall, and the F1-score. Accuracy states the number of documents correctly classified by the

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

VOL. 10. NO. 2 NOVEMBER 2024
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v10i2.5187

model [41]. The equation used to calculate the accuracy value of the model is shown in Equation 1.

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \quad (1)$$

where TP, or true positive, shows the number of TP values that correspond to positive predictions, and TN, or true negative, shows the number of TN values that match negative predictions. Likewise, FN, or false negative, shows the number of corresponding FN values from negative predictions, and FP, or false positive, shows the number of corresponding FP values from positive predictions. The resulting accuracy value shows that the model has the ability to predict targets.

Precision measures how many positive predictions are correct out of all positive predictions. The equation used to get the precision value for the model is shown by Equation 2.

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

where TP, or true positive, shows the number of TP values that correspond to positive predictions, while FP, or false positive, is the number of positive prediction values that do not match the target. The resulting precision value shows that the model has the ability to avoid positive value prediction errors.

Recall measures how many correct positive predictions there are from a particular class with all the data in that class. The equation used to get the recall value in the model is shown by Equation 3.

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

where TP, or true positive, shows the number of TP values that correspond to positive predictions, while FN, or false negative, is the number of negative prediction values that do not match the target. The resulting recall value shows that the model has the ability to avoid negative prediction errors.

The last measurement used is the F1-score. The F1-score-scorecombination of precision and recall. The F1-score-scoreed to strengthen the evaluation compared to just precision and recall while also showing the balance of the two. The equation used to get the F1-score value for the model is shown in Equation 4.

$$F1 - score = 2 * \frac{Recall*Precision}{Recall+Precision} \quad (4)$$

RESULTS AND DISCUSSION

In this section, the results of the experiments carried out in the study are described in detail. The experiment was carried out using a dataset that had been cleaned in Section II Point 3 and several classification methods in Section II Point 5.

A. Experimental Results of Data Sharing Variations

The first experiment was carried out by dividing the training and test data into several different percentages, namely 70:30, 80:20, and 90:10. This experiment aims to improve model performance through the amount of data learned. The results of experiments with percentages of 70:30, 80:20, and 90:10 on test and training data are shown in Tables 2, 3, and 4, respectively.

Table 2. Evaluation results on Cross Validation Percentage 90:10 Comparison

Model	*Accuracy*	*Precision*	*Recall*	F_1 *Score*
Decision Tree	78.99%	79.01%	79.13%	78.98%
Random Forest	86.21%	86.63%	86.60%	86.21%
KNN	84.01%	84.20%	84.29%	84.01%
SVM	72.72%	78.37%	74.12%	71.98%
Logistic Regression	79.31%	82.93%	80.38%	79.06%

Source: (Research Results, 2024)

Table 3. Evaluation results on Cross Validation Percentage 80:20 Comparison

Model	*Accuracy*	*Precision*	*Recall*	F_1 *Score*
Decision Tree	76.33%	76.32%	76.33%	76.32%
Random Forest	85.42%	85.77%	85.57%	85.41%
KNN	82.91%	82.97%	82.97%	82.91%
SVM	68.34%	75.17%	69.05%	66.54%
Logistic Regression	79.31%	82.77%	79.76%	78.91%

Source: (Research Results, 2024)

Table 4. Evaluation results on Cross Validation Percentage 70:30 Comparison

Model	*Accuracy*	*Precision*	*Recall*	F_1 *Score*
Decision Tree	75.23%	75.22%	75.22%	75.22%
Random Forest	83.18%	83.42%	83.08%	83.11%
KNN	79.73%	79.81%	79.78%	79.73%
SVM	59.98%	64.92%	59.31%	55.49%
Logistic Regression	67.92%	71.05%	67.48%	66.33%

Source: (Research Results, 2024)

After doing the experiments based on three splitting data ratios and five proposed algorithms,

the results were presented in four evaluation matrices score. Based on Tables 2, 3, and 4, the Random Forest algorithm produces the highest accuracy value, namely 86.21%, with a comparison of training data and test data of 90:10. At a ratio value of 80:20, the Random Forest algorithm produces a lower accuracy value, namely 0.79% and 3.03% at a ratio of 70:30. The enormous features of text dataset became return for Random Forest performance. The dataset in text form was potentially comprised of irrelevant and pretty noisy data. Random forest was also adequately robust against the imbalanced dataset used.

The SVM algorithm is the one that has the lowest performance among the four algorithms used to carry out classification tasks. Not much different from Random Forest, the SVM algorithm achieved the highest accuracy value at a 90:10 data comparison of 72.72%. Meanwhile, other algorithms, namely KNN, Decision Tree, and Logistic Regression, produce the highest accuracy values of 84.01%, 79.31%, and 78.99%, respectively.

B. Results of the K-Fold-Cross Validation Variation Experiment

The second experiment was an attempt to strengthen the experiments carried out in the first experiment. In the first experiment, classification was carried out using several percentage divisions of training data and test data, namely 70:30, 80:20, and 90:10. The second experiment was carried out using the k-fold cross-validation division method with k values of 5 and 10, respectively. The experimental results obtained are shown in Table 5 below.

Table 5. Evaluation results on Cross Validation

Model	*5-fold*	*10-fold*
Decision Tree	70.75%	74.39%
Random Forest	79.91%	81.34%
KNN	75.67%	78.05%
SVM	83.83%	85.21%
Logistic Regression	71.03%	72.15%

Source: (Research Results, 2024)

Based on the results of the second experiment using k-fold cross validation with values k = 5 and k = 10, the results showed that the SVM algorithm had the best performance. The SVM algorithm gets an accuracy value of 83.83% at k = 5 and increases to 85.21% at k = 10. The lowest accuracy value resulting from the second experiment was 70.75% in the Decision Tree algorithm with a value of k = 5. The accuracy of each algorithm produces a greater value at k = 10. The Decision Tree algorithm increased by 3.64%, Random Forest by 1.43%, KNN by 2.38%, SVM by 1.38%, and Logistic Regression by 1.12%.

In the first experiment, a 90:10 comparison of training data and test data resulted in high accuracy values for all algorithms used. This condition may indicate that the algorithm requires a larger portion of data to learn. Meanwhile, in the 70:30 ratio, the algorithm has a learning portion that is less than the total data it has. This condition also occurs at the value k = 10, where the resulting accuracy value is much higher than the value k = 5. At a value of k = 10, the algorithm has a larger learning portion, namely 9 for training data and 1 for test data.

CONCLUSION

Machine learning is the right method for determining the type of government video through comments on the YouTube platform. In this research, machine learning classifies comments into class 1 as government videos and class 0 as non-government videos. The machine learning method used on average has good performance, with a minimum accuracy value of 70.75%. The SVM algorithm is the method with the best performance among the other proposed methods, with an accuracy of 85.21%. The results of this accuracy were achieved in performance evaluation with a value of k = 10 in k-fold cross-validation. The result imposed that machine learning had potential benefits in the field of social media comment classification. Moreover, the research also defined that the classification task in social media comments could be explored more to get other insights. In future research, the use of word embedding and more diverse methods can be used to improve research results. In addition, the amount of data used can be increased to enrich machine learning.

REFERENCE

[1] S. Ashour, "Quality in Higher Education Quality higher education is the foundation of a knowledge society : where does the UAE stand ?," *Quality in Higher Education*, vol. 26, no. 2, pp. 209–223, 2020, doi: 10.1080/13538322.2020.1769263.

[2] R. D. Van Schalkwyk, J. Maritz, and R. J. Steenkamp, "Quality in Higher Education Sociotechnical service quality for students and academics at private higher education institutions in South Africa," *Quality in Higher Education*, vol. 27, no. 1, pp. 77–98, 2021, doi: 10.1080/13538322.2020.1815284.

[3] A. M. Bhandarkar, A. Kumar, and R. Nayak, "Impact of social media on the academic

VOL. 10. NO. 2 NOVEMBER 2024
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v10i2.5187

performance of undergraduate medical students," *Med J Armed Forces India*, vol. 77, pp. S37–S41, 2020, doi: 10.1016/j.mjafi.2020.10.021.

[4] W. Mohammed, T. Alanzi, F. Alanezi, H. Alhodaib, and M. AlShammari, "Usage of social media for health awareness purposes among health educators and students in Saudi Arabia," *Inform Med Unlocked*, vol. 23, p. 100553, 2021, doi: 10.1016/j.imu.2021.100553.

[5] D. Tolkach and S. Pratt, "Travel Professors: A YouTube channel about tourism education & research," *J Hosp Leis Sport Tour Educ*, vol. 28, no. September 2020, 2021, doi: 10.1016/j.jhlste.2021.100307.

[6] T. Ahmad, K. Sattar, and A. Akram, "Medical professionalism videos on YouTube: Content exploration and appraisal of user engagement," *Saudi J Biol Sci*, vol. 27, no. 9, pp. 2287–2292, 2020, doi: 10.1016/j.sjbs.2020.06.007.

[7] M. C. Meacham, E. A. Vogel, J. Thrul, D. E. Ramo, and D. D. Satre, "Addressing cigarette smoking cessation treatment challenges during the COVID-19 pandemic with social media," *J Subst Abuse Treat*, vol. 129, no. February, p. 108379, 2021, doi: 10.1016/j.jsat.2021.108379.

[8] H. A. Fang *et al.*, "An evaluation of social media utilization by general surgery programs in the COVID-19 era," *Am J Surg*, vol. 222, no. 5, pp. 937–943, 2021, doi: 10.1016/j.amjsurg.2021.04.014.

[9] G. Appel, L. Grewal, R. Hadi, and A. T. Stephen, "The future of social media in marketing," *J Acad Mark Sci*, vol. 48, no. 1, pp. 79–95, 2020, doi: 10.1007/s11747-019-00695-1.

[10] W. R. Fitriani, A. B. Mulyono, A. N. Hidayanto, and Q. Munajat, "Reviewer's communication style in YouTube product-review videos: does it affect channel loyalty?," *Heliyon*, vol. 6, no. 9, p. e04880, 2020, doi: 10.1016/j.heliyon.2020.e04880.

[11] G. Yavetz and N. Aharony, "Social media in government offices: usage and strategies," *Aslib Journal of Information Management*, vol. 72, no. 4, pp. 445–462, 2020, doi: 10.1108/AJIM-11-2019-0313.

[12] Z. Zhu, Y. Liu, N. Kapucu, and Z. Peng, "Online media and trust in government during crisis: The moderating role of sense of security," *International Journal of Disaster Risk Reduction*, vol. 50, p. 101717, 2020, doi: 10.1016/j.ijdrr.2020.101717.

[13] Y. Ryoo, H. Yu, and E. Han, "Political YouTube Channel Reputation (PYCR): Development and validation of a multidimensional scale," *Telematics and Informatics*, vol. 61, no. March, 2021, doi: 10.1016/j.tele.2021.101606.

[14] T. Notley, M. Dezuanni, S. Chambers, and S. Park, "Using YouTube to seek answers and make decisions: Implications for Australian adult media and information literacy," *Comunicar*, vol. 31, no. 77, 2023, doi: 10.3916/C77-2023-06.

[15] Y. Goldberg, *Neural network methods for natural language processing*. Springer Nature, 2022.

[16] Y. Piris and A. C. Gay, "Customer satisfaction and natural language processing," *J Bus Res*, vol. 124, no. January 2020, pp. 264–271, 2021, doi: 10.1016/j.jbusres.2020.11.065.

[17] Alamsyah, B. Prasetiyo, M. F. Al Hakim, and F. D. Pradana, "The improvement of COVID-19 prediction accuracy using optimal parameters in reccurent neural network model," in *AIP Conference Proceedings*, Semarang, 2023, p. 040019. doi: 10.1063/5.0125767.

[18] A. Gasparetto, M. Marcuzzo, A. Zangari, and A. Albarelli, "A survey on text classification algorithms: From text to predictions," *Information*, vol. 13, no. 2, p. 83, 2022.

[19] S. Subhan, M. F. Al Hakim, P. Listiaji, and W. Syafrizal, "Modeling news topics on government youtube channels with latent Dirichlet allocation method," in *AIP Conference Proceedings*, 2023, pp. 400091–400095. doi: 10.1063/5.0125954.

[20] D. Srivamsi, O. M. Deepak, M. D. A. Praveena, and A. Christy, "Cosine Similarity Based Word2Vec Model for Biomedical Data Analysis," in *7th International Conference on Trends in Electronics and Informatics, ICOEI 2023 - Proceedings*, 2023. doi: 10.1109/ICOEI56765.2023.10125794.

[21] E. M. Dharma, F. L. Gaol, H. L. H. S. Warnars, and B. Soewito, "the Accuracy Comparison Among Word2Vec, Glove, and Fasttext Towards Convolution Neural Network (Cnn) Text Classification," *J Theor Appl Inf Technol*, vol. 100, no. 2, pp. 349–359, 2022.

[22] A. Gasparetto, M. Marcuzzo, A. Zangari, and A. Albarelli, " Survey on Text Classification Algorithms: From Text to Predictions," *Information (Switzerland)*, vol. 13, no. 2, pp. 1–39, 2022, doi: 10.3390/info13020083.

[23] Q. Jiao and S. Zhang, "A Brief Survey of Word Embedding and Its Recent Development," in

IAEAC 2021 - IEEE 5th Advanced Information Technology, Electronic and Automation Control Conference, 2021. doi: 10.1109/IAEAC50856.2021.9390956.

[24] Derisma, D. Yendri, and M. Silvana, "Comparing the classification methods of sentiment analysis on a public figure on indonesian-language social media," *J Theor Appl Inf Technol*, vol. 98, no. 8, pp. 1214–1220, 2020.

[25] R. Amanda and E. S. Negara, "Analysis and implementation machine learning for youtube data classification by comparing the performance of classification algorithms," *Jurnal Online Informatika*, vol. 5, no. 1, pp. 61–72, 2020.

[26] M. Jena and S. Dehuri, "DecisionTree for Classification and Regression: A State-of-the Art Review," *Informatica*, vol. 44, no. 4, 2020.

[27] A. Plaia, S. Buscemi, J. Fürnkranz, and E. L. Mencía, "Comparing boosting and bagging for decision trees of rankings," *J Classif*, pp. 1–22, 2022.

[28] T. A. Assegie, R. L. Tulasi, and N. K. Kumar, "Breast cancer prediction model with decision tree and adaptive boosting," *IAES International Journal of Artificial Intelligence*, vol. 10, no. 1, p. 184, 2021.

[29] X. Chen, "Research on the Application of Decision Tree Algorithm in Agricultural Economic Development," in *2nd IEEE International Conference on Data Science and Information System, ICDSIS 2024*, Institute of Electrical and Electronics Engineers Inc., 2024. doi: 10.1109/ICDSIS61070.2024.10594692.

[30] A. K. Balyan *et al.*, "A hybrid intrusion detection model using ega-pso and improved random forest method," *Sensors*, vol. 22, no. 16, p. 5986, 2022.

[31] J. B. Awotunde, F. E. Ayo, R. Panigrahi, A. Garg, A. K. Bhoi, and P. Barsocchi, "A Multi-level Random Forest Model-Based Intrusion Detection Using Fuzzy Inference System for Internet of Things Networks," *International Journal of Computational Intelligence Systems*, vol. 16, no. 1, p. 31, 2023.

[32] R. Saini, "Integrating Vegetation Indices and Spectral Features for Vegetation Mapping from Multispectral Satellite Imagery Using AdaBoost and Random Forest Machine Learning Classifiers," *Geomatics and Environmental Engineering*, vol. 17, no. 1, pp. 57–74, 2023.

[33] P. H. Progga, M. J. Rahman, S. Biswas, M. S. Ahmed, and D. M. Farid, "K-Nearest Neighbour Classifier for Big Data Mining based on Informative Instances," in *2023 IEEE 8th International Conference for Convergence in Technology, I2CT 2023*, Institute of Electrical and Electronics Engineers Inc., 2023. doi: 10.1109/I2CT57861.2023.10126147.

[34] K. Shah, H. Patel, D. Sanghvi, and M. Shah, "A comparative analysis of logistic regression, random forest and KNN models for the text classification," *Augmented Human Research*, vol. 5, pp. 1–16, 2020.

[35] S. Uddin, I. Haque, H. Lu, M. A. Moni, and E. Gide, "Comparative performance analysis of K-nearest neighbour (KNN) algorithm and its different variants for disease prediction," *Sci Rep*, vol. 12, no. 1, pp. 1–11, 2022, doi: 10.1038/s41598-022-10358-x.

[36] M. Arya and C. S. S. Bedi, "Survey on SVM and their application in image classification," *International Journal of Information Technology*, vol. 13, no. 5, pp. 1867–1877, 2021, doi: 10.1007/s41870-017-0080-1.

[37] J. Alcaraz, M. Labbé, and M. Landete, "Support Vector Machine with feature selection: A multiobjective approach," *Expert Syst Appl*, vol. 204, 2022, doi: 10.1016/j.eswa.2022.117485.

[38] C. Brito-Pacheco, C. Brito-Loeza, and A. Martin-Gonzalez, "A regularized logistic regression based model for supervised learning," *J Algorithm Comput Technol*, vol. 14, 2020, doi: 10.1177/1748302620971535.

[39] R. Wang, N. Xiu, and C. Zhang, "Greedy Projected Gradient-Newton Method for Sparse Logistic Regression," *IEEE Trans Neural Netw Learn Syst*, vol. 31, no. 2, pp. 527–538, 2020, doi: 10.1109/TNNLS.2019.2905261.

[40] N. Srimaneekarn, A. Hayter, W. Liu, and C. Tantipoj, "Binary Response Analysis Using Logistic Regression in Dentistry," 2022. doi: 10.1155/2022/5358602.

[41] N. Hidayat, M. F. Al Hakim, and J. Jumanto, "Halal Food Restaurant Classification Based on Restaurant Review in Indonesian Language Using Machine Learning," *Scientific Journal of Informatics*, vol. 8, no. 2, pp. 314–319, 2021, doi: 10.15294/sji.v8i2.33395.

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