Jurnal TRANSFORMATIKA
Vol.20, No.2, JANUARI 2023, pp. 42-50
P-ISSN: 1693-3656, E-ISSN: 2460-6731
https://journals.usm.ac.id/index.php/transformatika/

Klasifikasi Sentimen Terhadap Gelaran MotoGP Mandalika 2022 Menggunakan Machine Learning

Doughlas Pardede[1], B. Herawan Hayadi[2]
Universitas Potensi Utama email:[1] doug.pardede@gmail.com,[2] b.herawan.hayadi@gmail.com

ARTICLE INFO

History of the article :
Received 29 Juli 2022
Received in revised form 29 Desember 2022
Accepted 9 Januari 2023
Available online 30 Januari 2023

Keywords:

Analisis sentimen; mandalika; motogp; k-means; svm

*** Correspondece:**

Telepon:
-

E-mail:
doug.pardede@gmail.com

ABSTRACT

MotoGP is a world-class motorcycle racing event, which will be held in the 19th series in 2022 at the Pertamina Mandalika Circuit. This study tries to analyze public sentiment collected from the results of tweeter social media tweets, in the form of sentiment and emotion values. With the features of sentiment and emotion values extracted from the contents of this tweet, k-means clustering is used to generate sentiment clusters as targets for classification using the MLP algorithm. From the results of the evaluation using 10-fold cross validation, the accuracy value is 97%, the precision value is 94.64% and the recall value is 100%. The classification results also show that the public response to the 2022 MotoGP event at the Mandalika circuit is quite balanced, where 53% have a positive response, while the rest have a negative response.

INTRODUCTION

MotoGP is a world-class motorcycle racing event held in several cities on different continents [1]. The competition, which is held in several series, which is broadcast live through television media, uses several different circuits in its implementation [2]. In 2022, precisely from March 18 to March 20, 2022, Indonesia was appointed as one of the host countries for the 19th MotoGP 2022 series which was held at the Pertamina Mandalika Circuit. Regarding the implementation of this prestigious world-class motorcycle racing event, it is interesting to analyze how the opinions or sentiments of the Indonesian people towards its implementation, based on data collected from social media Twitter.

Sentiment analysis can be understood as a process to understand, extract, and process data in the form of text so that sentiment information is obtained in the form of an opinion sentence [3]. By using computational methods, the information in the text in the form of opinions, sentiments, and emotions can be extracted and grouped, so that it can be decided whether the text has positive or negative sentiments. [4]. Sentiment analysis can be used to analyze how public opinion on a problem, such as lecturer evaluations [5], online learning [6], or the Covid-19 pandemic [7]. By using sentiment analysis techniques to extract sentiment and emotion values from Indonesian people's tweets about the 2022 MotoGP event, this study combines the k-means clustering and multilayer perceptron (MLP) algorithms to analyze the response from the public to this problem.

K-Means clustering groups the data into K groups, based on the partition of the data entered [8]. By using the minimum distance of each data to each cluster, k-means can be used to classify data into groups that can be used as targets in other algorithms [9], as in customer personality segmentation research [10], classification of malaria parasite patterns [11], and grouping employee performance levels [12]. The results of grouping tweet sentiment on the dataset using this algorithm are used as targets for the classification process using the MLP algorithm.

MLP is a neural network-based algorithm consisting of an input layer, two or more hidden layers and an output layer, which has a fairly good resistance to noise in the data [13]. MLP architecture which has more than one hidden layer allows this algorithm to solve complex data problems with good performance [14]. This algorithm is quite popularly used in classification research such as research on forecasting the volume of exports and imports of crude oil [15], citrus fruit ripeness texture classification [16], and classification of breast cancer [17]. This algorithm is used to classify public sentiment based on the features of the sentiment analysis results and the target from the results of the k-means clustering grouping, to then be evaluated using 10-fold cross validation, resulting in accuracy, precision and recall values.

By analyzing public sentiment obtained from social media, this study aims to see how the level of response from the community for the 2022 MotoGP event at Pertamina Mandalika Circuit, is it leaning towards a positive response or a negative response.

RESEARCH METHODS

This study aims to analyze sentiment towards the 2022 MotoGP race at the Mandalika circuit, based on netizen comments collected from Twitter social media. The experimental research method is used by building two different models, and then evaluating the results as a reference in sentiment analysis. The stages carried out in this study are shown in Figure 1.

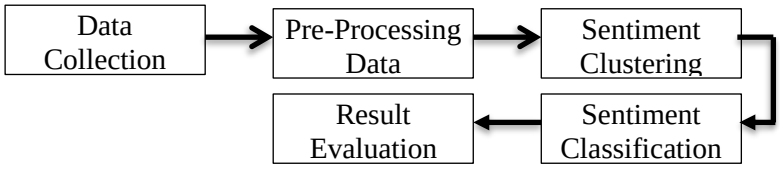

Figure 1. Research Stages

a. Data collection
   The data used in this study is sourced from Tweeter social media, which is crawled using the Orange 3 application, namely the Tweeter widget. In crawling the required tweet data, two keywords are used, namely "mandalika" and "motogp 2022".
b. Pre-processing data
   The data collected is first transformed through the process of removing words in the form of links (http and https) and converting all words in the contents of the tweet into lowercase letters. This whole process is done using the Pre-Processing Text widget in the Transformation section.
   Furthermore, the tokenization process is carried out by utilizing the pre-trained Twitter model provided as an option in the Orange 3 application, namely breaking sentences while maintaining symbols that are often used on Twitter social media such as "#" or "@", as well as emoticons. This whole process is done using the Pre-Processing Text widget in the Tokenization section.
   Furthermore, a filtering process is carried out on the tokens generated by removing words that have no meaning in sentiment analysis (stopwords). The list of words used for this process is ID-Stopwords obtained from the GitHub site [18]. This whole process is done using the Pre-Processing Text widget in the Filtering section.

Then the normalization process is carried out on the tokens generated using the Porter Stemmer algorithm, which is provided as an option in the Orange 3 application. The whole process is carried out using the Pre-Processing Text widget in the Normalization section.

c. Sentiment clustering

In this process, tweet data will be grouped based on the probability of the emotions contained in their contents, such as joy (happy), sadness (sad), anger (angry), fear (fear), disgust (disgusted) and surprise (surprised), using pre -trained multi-class tweeter emotion classification model Ekman [19]. The results of this grouping are then used as the first feature in the classification process.

This whole process is done using the help of the Tweet Profiler widget.

In addition to using the emotions contained in the tweet content, sentiment grouping is carried out using the pre-trained Multilingual Sentiment model provided as an option in the Orange 3 application, to measure whether tweet sentiment is positive or not. The results of this grouping are then used as the second feature in the classification process. This whole process is done using the help of the Tweet Profiler widget.

d. Sentiment classification

Sentiment classification in this study uses a combination of the k-means clustering algorithm and MLP. The k-means clustering algorithm is used to group the dataset into two clusters, namely C1 as a positive sentiment cluster and C2 as a negative sentiment cluster.

The results of this grouping are used as targets in the classification process using the MLP algorithm, which consists of four input neurons, three hidden layers, each of which has 10 neurons and three output neurons. In its architecture, the MLP algorithm uses the ReLu activation function and the Adam . optimization function.

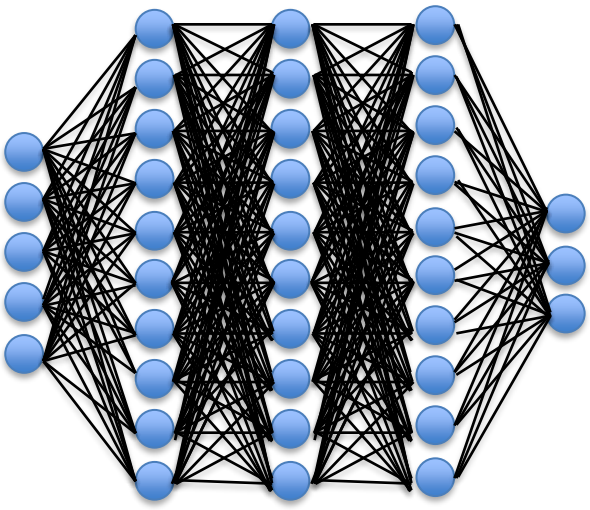

Figure 2. MLP Architecture

The ReLu function is used to optimize the values received from neurons in the previous layer, using equation (1) [20]. While the Adam optimization function is used as an optimizer in the process of changing the weight value and bias value, using equation (2) [21].

$$f_{ReLu} = Max(0, x)_t \quad (1)$$

$$\theta_{t+1} = \theta_t - \frac{\eta}{\sqrt{\widehat{u_t}+\varepsilon}}\widehat{a_t} \quad (2)$$

e. Result evaluation

Sentiment classification results are evaluated using accuracy, precision, and recall values obtained from the calculation of true positive, false positive, false negative, and true negative values from the confusion matrix table. The confusion matrix table is obtained from the evaluation process using 10-fold cross validation, while equation (3) to (5) are used to calculate each value [22]:

$$Accuracy = \frac{TP+TN}{TP+FP+FN+TN} * 100\% \quad (3)$$

$$Precision = \frac{TP}{TP+FP} * 100\% \quad (4)$$

$$Recall = \frac{TP}{TP+FN} * 100\% \quad (5)$$

In this study, the above stages were modeled using the Orange 3 data mining application, using widgets such as Twitter, Preprocess Text, Corpus Viewer, Tweet Profiler, Statistics, Select Rows, Select Columns, Edit Domain, Data Table, Neural Network, Test and Score, and Confusion Matrix.

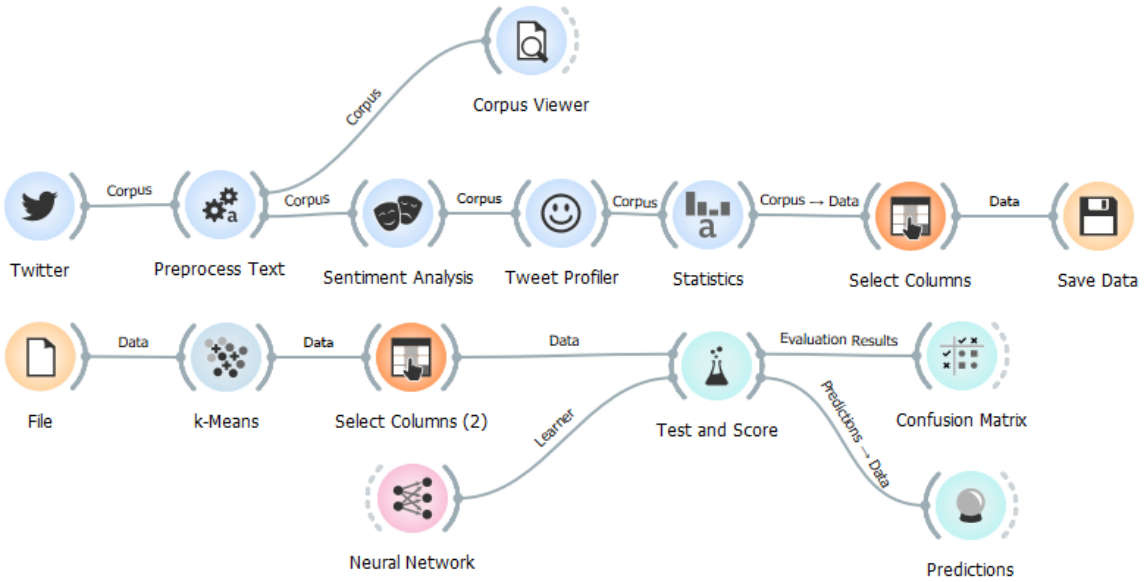

Figure 3. Research Model

RESULTS

In the process of collecting data using the Twitter API, crawling results were obtained in the form of 1116 tweets. Furthermore, pre-processing is carried out on the contents of the crawled tweet by transforming all the words in it into lowercase letters and discarding words in the form of links. The results of this transformation are then processed using Twitter's pre-trained model to break sentences into words that are used as tokens. The tokens obtained are then filtered to remove tokens that have no meaning, words containing the symbol "#" or "@", and emoticons. The results of this filtering are then normalized using Porter Stemmer to produce 1930 normalized tokens that are ready to be analyzed for the sentiment value in it.

Table 1. Sample of Normalized Token Results

Content	Token
Semoga fasilitas Solar Charging Station ini akan menambah kenyamanan Sahabat yang berkunjung ke kawasan The Mandalika.	semoga, fasilitas, solar, charging, station, menambah, kenyamanan, sahabat, berkunjung, kawasan, mandalika

Setelah sukses dilaksanakan Mandalika, Danau Toba dan Sumba, kali ini Program Relawan Bakti BUMN akan dilaksanakan di 6 (enam) lokasi, antara lain: - Bakti BUMN Untuk Way Kambas - Bakti BUMN Untuk Surakarta	sukses, dilaksanakan, mandalika, danau, toba, sumba, kali, program, relawan, bakti, bumn, dilaksanakan, enam, lokasi, bakti, bumn, way, kambas, bakti, bumn, surakarta
@teguhpositif @aniezz479 @yaniahmad695 Semula Mandalika 1,2T trus aspal hancur, di aspal lagi bengkak 2,5 T lgsg diem https://t.co/WG8EriAtiR	mandalika, 1,2, trus, aspal, hancur, aspal, lagi, bengkak 2,5, lgsg diem
@aniezz479 @yaniahmad695 https://t.co/L67EDFP3ld PNS disuruh beli Tiket	pns, disuruh, beli, tiket
Pembalap uji coba Stefan Bradl buka-bukaan soal penurunan yang terjadi di timnya, Repsol Honda selama MotoGP 2022. #StefanBradl https://t.co/l0ZLJ08nN4	pembalap, uji, coba, stefan, bradl, buka-bukaan, penurunan, timnya, repsol, honda, motogp
Peringati Hari Mangrove Sedunia, Traveloka dan SPI Kembali Tanam Ribuan Mangrove Di Mandalika https://t.co/OOI5xednsO	peringati, mangrove, sedunia, traveloka, spi, tanam, ribuan, mangrove, mandalika
Juara MotoGP Mandalika Malu Dibandingkan dengan Ronaldo https://t.co/wm6e3XsiPx	juara, motogp, mandalika, malu, dibandingkan, ronaldo
Beranikah congor kalian bersuara tentang mandalika, kereta cepat, bandara mangkrak, Tol dijual murah, korupsi gila-gilaan, BUMN pada rugi, penegakan hukum kacau, harga-harga kacau, ekonomi rakyat sekarat,,,,,,?? @psi_id https://t.co/OO8lHvHSVp	beranikah, congor, bersuara, mandalika, kereta, cepat, bandara, mangkrak, tol, dijual, murah, korupsi, gila-gilaan, bumn, rugi, penegakan, hukum, kacau, harga-harga, kacau, ekonomi, rakyat, sekarat
@nrwkra Di-carry sama Mandalika. Setelah itu ya anyep 😅	di-carry, mandalika, ya, anyep
@kusuma_loka Ramenya cuma hanya pas Mandalika, kebantu maniak MotoGP di Indonesia. Seri-seri lain rame, tapi kerasa banget kalah dari F1 yang mengglobal.	Ramenya, pas, mandalika, kebantu, maniak, motogp, indonesia, seri-seri, rame, kerasa, banget, kalah, f1, mengglobal

Based on the tokens obtained, each tweet content is analyzed using a pre-trained Multilingual Sentiment model in Indonesian, resulting in positive and negative sentiment values for each tweet content.

Table 2. 10 Samples of Sentiment Analysis Results

Content	Sentiment Value
1	18.1818
2	0
3	0
4	0
5	10
6	-4.34783

7	0
8	-20
9	-11.1111
10	-25

Based on the tokens obtained, each tweet is grouped based on the probability of the emotion contained in it. The results obtained are six types of emotions detected in the dataset, consisting of 2 tweets belonging to the anger emotion group, 9 tweets belonging to the fear emotion group, 60 tweets belonging to the joy emotion group, 1 tweet belonging to the sadness emotion group, and 28 tweets belong to the surprise emotion group.

Table 3. Emotions Clustering in the Dataset

Emotion	Total
Anger	2
Fear	9
Joy	60
Sadness	1
Surprise	28

The results of the sentiment analysis and grouping are then used as a feature in the process of grouping sentiment clusters using k-means clustering, so that the results obtained are clusters C1 (Positive Sentiment) and C2 (Negative Sentiment) clusters.

Table 4. Sentiment Clustering Results

Sentiment	Anger	Disgust	Fear	Joy	Sadness	Surprise	C
18.1818	0.234768	0.128325	0.229532	0.132432	0.108491	0.166451	C2
0	1.03E-08	1.47E-11	0.00109629	0.998904	8.95E-09	1.24E-07	C1
0	0.0740298	0.0483292	0.155999	0.334836	0.123816	0.262989	C2
0	0.0717993	0.0365372	0.201717	0.310706	0.110274	0.268966	C2
10	2.54E-07	5.07E-09	0.00236705	0.997631	1.50E-07	1.30E-06	C1
-4.34783	1.69E-07	3.41E-10	0.00130043	0.998699	1.20E-08	8.11E-07	C1
0	9.70E-08	7.34E-11	0.00252831	0.997471	1.61E-07	6.20E-07	C1
-20	0.0555115	0.135355	0.180116	0.0392376	0.27581	0.31397	C2
-11.1111	8.51E-07	1.42E-08	0.00276776	0.997217	9.09E-07	1.30E-05	C1
-25	0.0545852	0.0752294	0.258741	0.0958554	0.0659583	0.44963	C2

The results of sentiment clustering in Table 7 are used as targets in the classification using the MLP algorithm, which are then evaluated using 10-fold cross validation. The results of 10-fold cross validation in the form of a confusion matrix table that are generated are then processed so that the values of accuracy, precision, and recall are obtained to analyze the performance of the resulting model.

		Predicted		
		Positif	Negatif	∑
Actual	Positif	53	3	56
	Negatif	0	44	46
	∑	53	47	100

Figure 4. Confusion Matrix Results

From Figure 4 above, then the accuracy, precision, and recall values are calculated using equations (3) to (5):

DOI: http://dx.doi.org/10.26623/transformatika.v20i2.5364

$$Accuracy = \frac{53+44}{53+3+0+44} * 100\% = \frac{97}{100} * 100\% = 97\%$$

$$Precision = \frac{53}{53+3} * 100\% = \frac{53}{56} * 100\% = 94{,}64\%$$

$$Recall = \frac{53}{53+0} * 100\% = \frac{53}{53} * 100\% = 100\%$$

DISCUSSION

From the results of sentiment analysis on the 2022 MotoGP event at the Mandalika circuit, Ekman's pre-trained model was able to analyze the emotional sentiments of joy, surprise, sadness, anger, and fear in the contents of tweets. This emotional value can be used as a feature in the training dataset in the clustering and classification process. In addition to emotional features, the results of sentiment analysis can also be used as features in the clustering and classification process. Token frequency statistics in the dataset such as character count, word count, average word length and N-Gram value can be used as features in the classification. The results of the sentiment classification of the MotoGP 2022 show that the resulting model can classify twitter content sentiment very well, as can be seen from the accuracy results obtained, which is 97%, the precision results are 94.64%, and the recall results are 100%. From the classification results, there were 53 positive responses and 47 negative responses.

CONCLUSIONS AND RECOMMENDATIONS

From the results of this study, it can be seen that to analyze sentiment towards certain problems, data collection can be carried out through crawling on Twitter social media. The results of this crawl are then used as a dataset, after pre-processing to get tokens for each tweet. The tokens generated are then used to analyze the emotions that depend on them such as joy, surprise, sadness, anger, and fear. In addition to emotions, another thing that can be analyzed is how the sentiment value of each tweet is, whether it leads to a positive or a negative direction. From the results of these two analyzes, clustering can be done using the k-means clustering algorithm to obtain a target in the form of positive sentiment or negative sentiment from the dataset. The classification is then continued to obtain the category of Positive Sentiment and Negative Sentiment class. The results of the model in this study have excellent performance, as evidenced by the 97% accuracy value, 94.64% precision value, and 100% recall value. Overall, the public response to the 2022 MotoGP event at the Mandalika circuit is quite balanced, where 53% of the public's response is positive and 47% of the other response is negative.

REFERENCES

[1] Pinch P, Reimer S. MotoGP and heterogeneous design. In: Spinney J, Reimer S, Pinch P, editors. Mobilising Des., Abingdon, Oxon, ; New York, NY : Routledge, 2017. | Series: Routledge Studies in Human Geography ; 69: Routledge; 2017, p. 117–29. https://doi.org/10.4324/9781315560113-10.

[2] Purnamasari Y. Competition Frame on MotoGP Articles: A Cognitive Linguistic Approach.

J Bhs Ingg Terap 2018;4:20–33. https://doi.org/10.35313/jbit.v4i1.1355.

[3] Mandasari S, Hayadi BH, Gunawan R. Analisis Sentimen Pengguna Transportasi Online Terhadap Layanan Grab Indonesia Menggunakan Multinomial Naive Bayes Classifier. J Teknol Sist Inf Dan Sist Komput TGD 2022;5:118–26.

[4] Fitri E. Analisis Sentimen Terhadap Aplikasi Ruangguru Menggunakan Algoritma Naive Bayes, Random Forest Dan Support Vector Machine. J Transform 2020;18:71. https://doi.org/10.26623/transformatika.v18i1.2317.

[5] Santoso VI, Virginia G, Lukito Y. Penerapan Sentiment Analysis Pada Hasil Evaluasi Dosen Dengan Metode Support Vector Machine. J Transform 2017;14:72. https://doi.org/10.26623/transformatika.v14i2.439.

[6] Putra RS, Agustin W, Anam MK, Yaakub S. The Application of Naïve Bayes Classifier Based Feature Selection on Analysis of Online Learning Sentiment in Online Media. J Transform 2022;20:44–56.

[7] Furqan M, Mayang Sari S, Ilmu Komputer Fakultas Sains dan Teknologi P. Analisis Sentimen Menggunakan K-Nearest Neighbor Terhadap New Normal Masa Covid-19 Di Indonesia. TechnoCOM 2022;21:52–61. https://doi.org/https://doi.org/10.33633/tc.v21i1.5446.

[8] Selvida D, Zarlis M, Situmorang Z. Analysis of the effect early cluster centre points on the combination of k-means algorithms and sum of squared error on k centroid. IOP Conf Ser Mater Sci Eng 2020;725. https://doi.org/10.1088/1757-899X/725/1/012089.

[9] Hartono, Sitompul OS, Tulus, Nababan EB. Optimization Model of K-Means Clustering Using Artificial Neural Networks to Handle Class Imbalance Problem. IOP Conf Ser Mater Sci Eng 2018;288. https://doi.org/10.1088/1757-899X/288/1/012075.

[10] Harahap M, Lubis Y, Situmorang Z. Analisis Pemasaran Bisnis dengan Data Science : Segmentasi Kepribadian Pelanggan berdasarkan Algoritma K-Means Clustering. Data Sci Indones 2022;1:76–88. https://doi.org/10.47709/dsi.v1i2.1348.

[11] Prasetio A, Rosnelly R. Identification of Malaria Parasite Patterns With Gray Level Co-Occurance. J RESTI (Rekayasa Sist Dan Teknol Informasi) 2022;5:359–69.

[12] Agustina N, Prihandoko P. Perbandingan Algoritma K-Means dengan Fuzzy C-Means Untuk Clustering Tingkat Kedisiplinan Kinerja Karyawan. J RESTI (Rekayasa Sist Dan Teknol Informasi) 2018;2:621–6. https://doi.org/10.29207/resti.v2i3.492.

[13] Pardede D, Hayadi BH, Iskandar. Kajian Literatur Multi Layer Perceptron Seberapa Baik Performa Algoritma Ini. J ICT Apl Syst 2022;1:23–35. https://doi.org/10.56313/jictas.v1i1.127.

[14] Hayadi BH, Sudipa IGI, Windarto AP. Model Peramalan Artificial Neural Network pada Peserta KB Aktif Jalur Pemerintahan menggunakan Artificial Neural Network Back-Propagation. MATRIK J Manajemen, Tek Inform Dan Rekayasa Komput 2021;21:11–20. https://doi.org/10.30812/matrik.v21i1.1273.

[15] Wanto A, Herawan Hayadi B, Subekti P, Sudrajat D, Wikansari R, Bhawika GW, et al. Forecasting the Export and Import Volume of Crude Oil, Oil Products and Gas Using ANN. J Phys Conf Ser 2019;1255. https://doi.org/10.1088/1742-6596/1255/1/012016.

[16] Yanto B, Fimawahib L, Supriyanto A, Hayadi BH, Pratama RR. Klasifikasi Tekstur Kematangan Buah Jeruk Manis Berdasarkan Tingkat Kecerahan Warna dengan Metode Deep Learning Convolutional Neural Network. INOVTEK Polbeng - Seri Inform 2021;6:259. https://doi.org/10.35314/isi.v6i2.2104.

[17] Kusuma J, Hayadi BH, Rosnelly R. Komparasi Metode Multi Layer Perceptron (MLP) dan Support Vector Machine (SVM) untuk Klasifikasi Kanker Payudara. MIND (Multimedia Artif Intell Netw Database) J 2022;7:51–60.

[18] Haryalesmana D, Wieriks M. ID-Stopwords. 2019.

[19] da Silva IRR, Lima ACES, Pasti R, de Castro LN. Classifying emotions in Twitter messages using a deep neural network. Adv Intell Syst Comput 2019;801:283–90. https://doi.org/10.1007/978-3-319-99608-0_32.

[20] Işık G, Ekici S, Şahin G. A neural network model for UAV propulsion system. Aircr Eng Aerosp Technol 2020;92:1177–84. https://doi.org/10.1108/AEAT-04-2020-0064.

[21] Irfan D, Rosnelly R, Wahyuni M, Samudra JT, Rangga A. PERBANDINGAN OPTIMASI SGD, ADADELTA, DAN ADAM DALAM KLASIFIKASI HYDRANGEA MENGGUNAKAN CNN. J Sci Soc Res 2022;4307:244–53.

[22] Samudra JT, Hayadi BH, Ramadhan PS. Komparasi 3 Metode Algoritma Klasifikasi Data Mining Pada Prediksi Kenaikan Jabatan. J Teknol Sist Inf Dan Sist Komput TGD 2022;5:127–33.