Accredited rank 3 (SINTA 3), excerpts from the decision of the Minister of RISTEK-BRIN No. 200/M/KPT/2020

COMPARISON OF CLASSIFICATION ALGORITHMS FOR ANALYSIS SENTIMENT OF FORMULA E IMPLEMENTATION IN INDONESIA

Fachri Amsury[1], Nanang Ruhyana[2], Tati Mardiana[3*)]

Sistem Informasi[1], Sains Data[2,3]
Universitas Nusa Mandiri
fachri.fcy@nusamandiri.ac.id[1], nanang.ngy@nusamandiri.ac.id[2], tati.ttm@nusamandiri.ac.id[3*)]
(*) Corresponding Author

Abstract

The Formula E racing series has become one of the world's most prestigious competitions. In 2022, Indonesia hosted the famous Formula E race. The event possesses the potential for economic benefits for Indonesia worth 78 million euros through the arrival of 35,000 spectators. Indonesians are enthusiastic about Formula E since it allows their nation to encourage tourists and gain international prominence. However, some people do not support this event. Since they regard that amid the COVID-19 pandemic, it is preferable for the government to focus on people affected by the pandemic rather than support a Formula E event. This study compares the Support Vector Machine and Naive Bayes algorithms in classifying public opinion in the Formula E race. This study gets its information from user comments on social media platforms, especially Twitter. The stages start with text preprocessing and include cleaning, case folding, tokenization, filtering, and stemming. Proceed with weighting using the TF-IDF approach. Data testing uses a confusion matrix to evaluate the classification results by testing accuracy, precision, and recall. Categorizing public opinion using the SVM algorithm has an accuracy of 82 percent, a precision of 97.86 percent, and a recall of 77.90 percent. On the other hand, the accuracy of the Naive Bayes technique is more limited, at 87.54 percent. Society's opinion on Twitter shows positive sentiment towards implementing Formula E.

Keywords: Formula E; Tweet; Sentiment Analysis; Support Vector Machine; Naive Bayes

Abstrak

Seri balap Formula E telah menjadi salah satu kompetisi paling bergengsi di dunia. Pada tahun 2022, Indonesia menjadi tuan rumah balapan Formula E yang terkenal. Ajang tersebut memiliki potensi manfaat ekonomi bagi Indonesia senilai 78 juta euro melalui kedatangan 35.000 penonton. Orang Indonesia antusias dengan Formula E karena memberikan kemungkinan bagi bangsa mereka untuk mendorong wisatawan dan mendapatkan keunggulan internasional. Namun, ada pihak yang tidak mendukung acara ini. Karena mereka menilai di tengah pandemi COVID-19, lebih baik pemerintah fokus pada masyarakat yang terdampak pandemi daripada mendukung ajang Formula E. Penelitian ini bertujuan untuk membandingkan algoritma Support Vector Machine dan Naive Bayes dalam klasifikasi opini publik pada balapan Formula E. Penelitian ini mendapatkan informasi dari komentar pengguna pada platform media sosial, khususnya Twitter. Tahapannya dimulai dengan text preprocessing dan meliputi cleaning, case folding, tokenization, filtering, dan stemming. Selanjutnya dilanjutkan dengan pembobotan menggunakan pendekatan TF-IDF. Untuk mengevaluasi hasil klasifikasi, pengujian data menggunakan matriks konfusi dengan menguji akurasi, presisi, dan recall. Dalam hal pengkategorian opini publik, algoritma SVM memiliki akurasi 82 persen, presisi 97,86 persen, dan recall 77,90 persen. Di sisi lain, akurasi teknik Naive Bayes lebih terbatas, yaitu 87,54 persen. Opini masyarakat di Twitter menunjukkan sentimen positif terhadap pelaksanaan Formula E.

Kata kunci: Formula E; Tweet; Analisis Sentimen; Support Vector Machine; Naive Bayes

INTRODUCTION

Formula E, at first glance, looks the same as Formula 1, but Formula E uses completely electric energy as its fuel. The Formula E event delivers a mission to promote environmentally friendly electric vehicles, along with many negative criticisms about motorsports, excessive consumption of fuel oil that contributes to changes global environment from emissions produced (Robeers, 2019)

The work is distributed under the Creative Commons Attribution-NonCommercial 4.0 International License

Formula E is a new racing series dedicated to electric racing vehicles (Fatemi, Ionel, Popescu, & Demerdash, 2016). The Formula E uses an electric battery to operate and regulate the motor, utilizing the electric energy that is pushed at all global racing events (Hall, 2017). In modern motorsport, a crucial problem that often occurs is energy management in racing events, and technical regulations impose limits on the power and energy used in Formula E. The limit is the limitation of output power and various rechargeable energy storage systems (RESS) for cars. Depending on the type of event (Liu, Fotouhi, & Auger, 2020).

The Formula E racing series has become one of the world's most prestigious competitions. In 2022, Indonesia hosted the famous Formula E race. The event possesses the potential for economic benefits for Indonesia worth 78 million euros through the arrival of 35,000 spectators.

The Governor of DKI Jakarta, Mr. Anis Baswedan, issued a regulation through the instructions of the Governor of the Special Capital Region of Jakarta Number 49 of 2021 regarding the settlement of regional priority issues for 2021-2022, which was set on August 4, 2021, by Mr. Anis Baswedan. Formula E is one of the resolutions of priority issues, then the target for completion is implementing the Formula E race (Tanjung, 2021). Contains a list of achievements in resolving regional priority issues in 2021-2022.

The instructions from the Governor of DKI Jakarta regarding the implementation of formula E raise pros and cons among the public because this activity will be held amid the COVID-19 pandemic currently engulfing this country, especially because of the issue of electric cars will cost much money in its implementation. The most crucial problem is the method of funding which is considered to violate existing regulations due to the commitment fee installments, which will last until 2023, which means it will exceed the term of office of the Governor of DKI, which ends in 2022 (Tanjung, 2021)

PT. Jakarta Propertindo (JakPro) was appointed by the Governor of DKI Jakarta, Mr. Anis Baswedan, to hold an international formula E racing event in 2020. However, this implementation has caused pro and contra reactions among politicians and the public regarding the DKI provincial government's plan to use APBD funds with a value of 1.6 Trillion to carry out the Formula E racing event, and there were also protests regarding the place that would become the planned Formula E circuit at Monas. The support and criticism of the Formula E event have attracted attention and warm discussion among the public (Kartinawati & Wiyawan, 2021).

Related research on the topic of analysis of public sentiment for Twitter to analyze public opinion from tweets posted with analysis using the nave Bayes method has an accuracy of 87.34%, a sensitivity of 93.43%, and a specificity of 71.76%, proving a fairly good result in his research (Putu et al., 2021). Research on sentiment analysis of television broadcasts based on public opinion on Twitter with the K-NN method with a total of 400 tweets data produces an accuracy rate using textual weighting is 82.50% with a value of k = 3, giving good results (Berliana, Shaufiah, S.T., & Siti Sa'adah, S.T., 2018). Research on the classification of posting tweets regarding government policies by applying the nave Bayes algorithm by extracting the unigram feature with tweet data of 578 data produces an accuracy rate of 80% (Berliana et al., 2018). Research on complaint classification using the SVM algorithm with a total dataset of 1040 complaints data using rapid miner tools produces an accuracy rate of 95.67%(Fatmawati & Affandes, 2018).

This study aims to see the public response to the Formula E event, whether to support or reject this event based on community comments taken from one of the social media platforms, namely Twitter, by applying data mining techniques to classify tweets that support the formula e event using a technical approach. SMOTE and Support Vector Machine and Naïve Bayes algorithms.

RESEARCH METHODS

Research using the concept of Knowledge Discovery in Database (Tsytsarau & Palpanas, 2012). To classify tweets containing support for organizing a Formula E event in Jakarta.

Types of research

This study uses qualitative data sourced from Twitter in the form of tweet data with the query Formula E

Research Target / Subject

The data population is sourced from Twitter, namely tweet data taken from May 2020 to June 2022 with a Formula E query. The data structure of the results of crawling Twitter data using the RapidMiner application consists of several columns, namely Created-At, From-User, From - User-Id, To-User, To-User-Id, Language, Source, Text, Geo-Location-Latitude, Geo-Location-Longitude, Retweet-Count and ID. The tweet data generated from this process amounted to 917 tweets.

The work is distributed under the Creative Commons Attribution-NonCommercial 4.0 International License

Procedure

The following is a framework that is carried out based on the steps and procedures of this research process

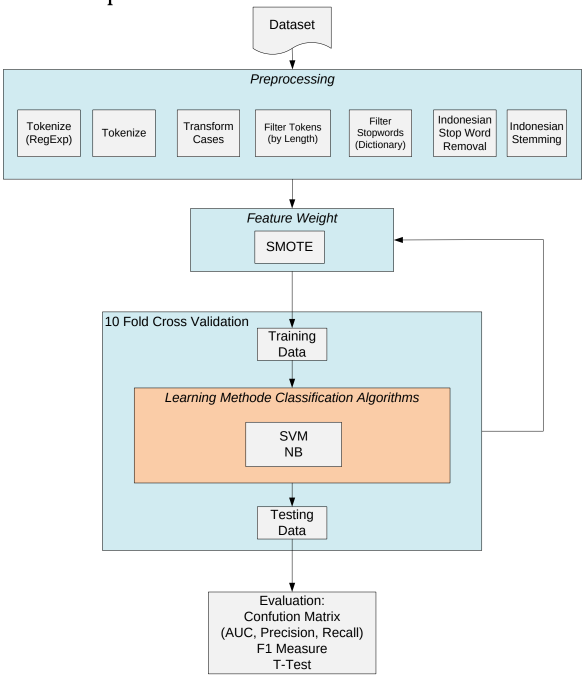


Figure 1. Research Method

Figure 1 explains that the research model carried out is the classification of tweets containing support for implementing the Formula E event in Indonesia, precisely in Jakarta. The text mining technique and the Support Vector Machine (SVM) and Naïve Bayes methods are used by applying the SMOTE technique to balance the data (Gata et al., 2019).

1. Data

The data is sourced from Twitter which is a tweet posted by the general public, tweet data retrieval using the Twitter data crawling technique using the RapidMiner application, then labeling the tweet data to classify tweets that support the Formula E event in Jakarta and tweets that do not support the Formula E event.

2. Data Selection & Cleaning

It aims to select and clean tweet data originating from Twitter, eliminating noise and duplication of data.

3. Data Transformation

Tweet data selected and cleaned from duplication and noise is then transformed into Xls format to facilitate the data mining.

4. Data mining

Tweet data that has been transformed into a dataset, the next stage is to approach the right algorithm by comparing the two algorithms, namely SVM and Naïve Bayes, to see the performance generated from the two algorithms and measure the level of accuracy generated.

5. Evaluation

The dataset modeled using the SVM and Naïve Bayes algorithms are then visualized to simplify the process of comparing the performance and accuracy of the two algorithms.

Data, Instruments, and Data Collection Techniques

The data population is sourced from Twitter, namely tweet data taken from May 2020 to June 2022 with a Formula E query. The data structure of the results of crawling Twitter data using the RapidMiner application consists of several columns, namely Created-At, From-User, From - User-Id, To-User, To-User-Id, Language, Source, Text, Geo-Location-Latitude, Geo-Location-Longitude, Retweet-Count and ID. The tweet data generated from this process amounted to 917 tweets. The attribute used in this data is the Text Column which contains tweets regarding support for the Formula E event in Jakarta.

Data analysis technique

This study uses tweet data taken from Twitter using the rapid miner tweet data tool from May 2020 to June 2022 with a Formula E query that will be used in the text mining process. The data structure of the Twitter data crawl process consists of several columns, namely Created-At, From-User, From-User-Id, To-User, To-User-Id, Language, Source, Text, Geo-Location-Latitude, Geo -Location-Longitude, Retweet-Count and ID. The next phase is adding a new field, namely the label used as a class. Twitter has much information about tweets that support and don't support the implementation of Formula E in Jakarta.

RESULTS AND DISCUSSION

The following is a sample of the data used in the study.

Table 1. Sample twitter data with a label

text	label
@KRMTRoySuryo2 Jegal kiri kanan depan belakang jalan terus formula E harumkan Indonesia pada dunia ??	support
@KRMTRoySuryo2 Mau dijegal sana-sini, dinyinyirin sana-sini, kalo Emang sudah takdirnya acara	support

The work is distributed under the Creative Commons Attribution-NonCommercial 4.0 International License

text	label
sukses, bakal sukses, turut senang yg diundang pada hadir, dari pak presiden, bu ketua MPR, Pak ketua DPR, kepala BPKP, Menpora, sampai Menteri pariwisata, kapolri, dll https://t.co/0wK7TjAb10	
Event Formula E selesai dan tuntas.. terimakaih semua yang sudah menyukseskan… baik dr tenaga kerja sampai presiden @jokowi yg sdh hadir.. semua berjalan dg lancar dan tdk sampai hujan.. lupakan Evoria formula e.. waktunya kita sambut hajatan DKI jkt nanti.. https://t.co/1WzJjthXKS	support
@alisyarief GILU LE NDRO!!!!! Ajuin proposal sponsorship Formula E ke BUMN ala ala ormas preman malak buat 17 agustusan . https://t.co/WsjqipeDdr	notsupport
@msaid_didu hanya orang tolol yang mau sponsorin FORMULA E, makanya ditangan mu BUMN banyak lubang buat lahan korupsi! karna yang ngurus TOLOL	notsupport
@KangMasPrabu165 @erickthohir Makanya kalau mau ngadain event Internasional itu jangan dadakan ngajuin proposal, ini kok kayak jualan tahu bulet... Dadakan????	not support

Table 1 is a sample of research tweet data consisting of a text column containing tweets and a label column is a class of tweet labeling results using the Crowdsourced labeling technique divided into two classes, namely support and not support classes.

Teknik SMOTE

The SMOTE (Synthetic Minority Over-sampling Technique) technique treats class imbalances (Amsury, Ruhyana, Saputra, & Sulistyowati, 2020). This technique synthesizes new sample data from the minority class to balance the dataset by creating new instances from the minority class and forming a convex combination of adjacent instances. The training data set consists of Smin minority data points and Smaj majority data points. For each (Xi,Yi) ∈ Smin , most data points set Smaj. For each (xi, yi) ∈ Smin, SMOTE generates a new minority data point along a line segment joining xi and one of the k closest neighbors chosen randomly.

$$\chi_{new} = \chi_i + (\chi_j - \chi_i) \times \delta, ...(1)$$

Where is δ a random number between [0, 1]. The xnew class label is set to +1 (ynew), the minority class label. The newly generated data point {(xnew, ynew)} is then added to the original set S and used to train the classifier. From the above formulation, it is clear that SMOTE informs oversampling to generalize the minority class by creating a larger and less specific decision region rather than a smaller and more specific original region

Support Vector Machine

Support Vector Machine (SVM) is a supervised learning technique with many desirable qualities making it a popular algorithm (Mathew, Luo, Pang, & Chan, 2015). It has a solid theoretical foundation and performs classification more accurately than most other application algorithms (Windasari, Uzzi, & Satoto, 2018).

Many researchers have reported that SVM may be an accurate method for text classification (Demidova & Klyueva, 2017). It is also widely used in sentiment classification. The SVM formulation for solving binary classification problems is briefly described in this section. It should be noted that binary SVM can easily be extended to multi-class classification using methods such as one-on-many and one-on-one. Figure 2 shows the SVM algorithm modeling validation process to get the accuracy value.

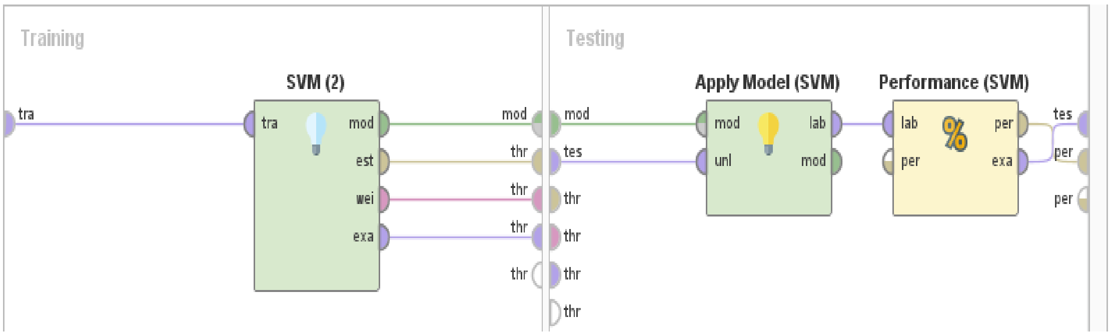
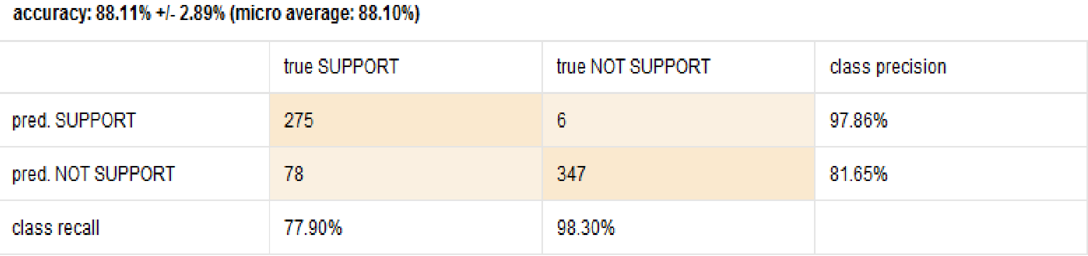


Figure 2. SVM Algorithm Validation Process

The work is distributed under the Creative Commons Attribution-NonCommercial 4.0 International License

P-ISSN: 2656-1743 |E-ISSN: 2656-1735
DOI: https://doi.org/10.34288/jri.v4i3.XXX
Accredited rank 3 (SINTA 3), excerpts from the decision of the Minister of RISTEK-BRIN No. 200/M/KPT/2020


accuracy: 88.11% +/- 2.89% (micro average: 88.10%)

	true SUPPORT	true NOT SUPPORT	class precision
pred. SUPPORT	275	6	97.86%
pred. NOT SUPPORT	78	347	81.65%
class recall	77.90%	98.30%	

Figure 3. The results of the accuracy of the SVM algorithm

Figure 3 explains the results of the SVM algorithm accuracy is 88.11%, with class precision, prediction support results of 97.86%, class precision prediction do not support 81.65%, and class recalls true support results 77.90%, class recall true not support 98.30%. The tweet data results show that the classification for true support tweets with predicted support is 275 data, and tweet data is not supported with 6 data support predictions. True support tweet data by not supporting prediction to much as 78 data, and does not support tweet data to 347 data as not supporting the prediction.

Naïve Bayes

Naive Bayes is a learning algorithm based on Bayes theory using strong assumptions (naive)(Barfian, Iswanto, & Isa, 2017). Bayes's theory is about finding a probability of something based on pre-existing data. This method can also classify opinions based on previously trained data. The essence of naive Bayes is finding the data's highest probability. Bayes formula can be written as follows:

$$Pcd = \frac{P(c)X\ Pdc}{P(d)} \quad \text{...} (2)$$

Information:
Pcd: probability class c after d is entered into class c
P(c): the probability of class c before
Pdc: probability d in class c
Pd: probability d

Figure 4 shows the validation process using the Naïve Bayes algorithm modeling to get the accuracy value.

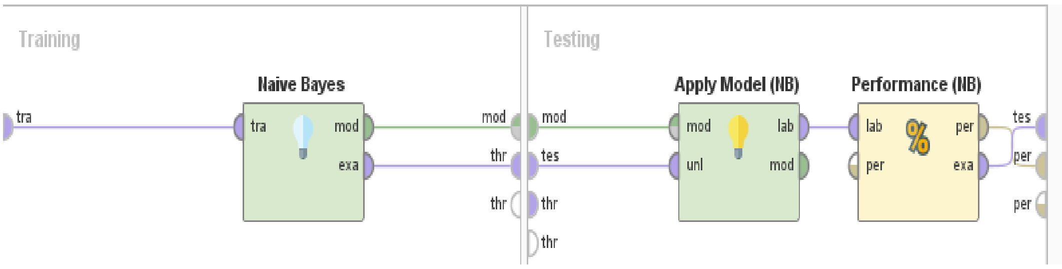
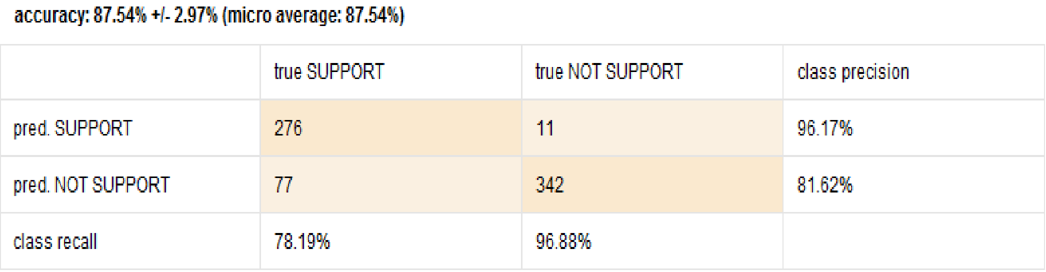


Figure 4. Nave Bayes algorithm validation process


accuracy: 87.54% +/- 2.97% (micro average: 87.54%)

	true SUPPORT	true NOT SUPPORT	class precision
pred. SUPPORT	276	11	96.17%
pred. NOT SUPPORT	77	342	81.62%
class recall	78.19%	96.88%	

Figure 5. The results of the accuracy of the Naïve Bayes algorithm

Figure 5 explains that the accuracy of the Naïve Bayes algorithm results is 87.54%, with

The work is distributed under the Creative Commons Attribution-NonCommercial 4.0 International License

results to support the precision class prediction of 96.17%, class precision prediction does not help 81.62%, and class recalls actual support results 78.19%, class recall true not supporting 96.88%. Based on the results of real support tweet data with support predictions of 276 data and true not support tweet data with 11 data support predictions. Since numerous nations worldwide take part in the Formula E. True support tweet data with predictions of not keeping with as 77 data, and tweet data of not help with forecasts of not investing with 342 data event.

CONCLUSIONS AND SUGGESTIONS

Conclusion

The classification results from this study using the SVM algorithm get an accuracy value of 88.11%, with precision class prediction is support results of 97.86%, class precision prediction does not support 81.65%, and class recalls real support results 77.90%, class recall true not supporting 98.30%. While the results of the accuracy of the Naïve Bayes algorithm are 87.54%, with the results of class precision prediction support of 96.17%, class precision prediction of not supporting 81.62%, and the effects of class recalling true support being 78.19%, class remembering being true not being helping 96.88%. Based on the results obtained, the accuracy level using the SVM algorithm has a higher level of accuracy compared to the accuracy results generated by the Naïve Bayes algorithm. A high level of precision so that you can see how many tweets from the Indonesian people are intended to support implementing an international standard Formula E event in Indonesia.

Suggestion

Suggestions for further research are expected to add features such as N-Gram and apply optimization techniques. Then explore various data, for example, from comments on Instagram, YouTube, or other social media, adding classes with more than two classes, and adding the number of data samples so that research results become more and more accurate, especially in classifying Twitter data.

REFERENCES

Amsury, F., Ruhyana, N., Saputra, I., & Sulistyowati, D. N. (2020). Classification of Customer Complaints on Instagram Comments Using Naïve Bayes Algorithm With N-Gram Feature Extension. *Jurnal Techno Nusa Mandiri*, *17*(2), 109–116. https://doi.org/10.33480/techno.v17i2.1632

Barfian, E., Iswanto, B. H., & Isa, S. M. (2017). Twitter Pornography Multilingual Content Identification Based on Machine Learning. *Procedia Computer Science*, *116*, 129–136. https://doi.org/10.1016/j.procs.2017.10.024

Berliana, G., Shaufiah, S.T., M. T., & Siti Sa'adah, S.T., M. T. (2018). Klasifikasi Posting Tweet mengenai Kebijakan Pemerintah Menggunakan Naive Bayesian Classification. *E-Proceeding of Engineering*, *5*(1), 1562–1569. Bandung: Telkom University. Retrieved from https://openlibrarypublications.telkomuniversity.ac.id/index.php/engineering/article/view/6170

Demidova, L., & Klyueva, I. (2017). SVM classification: Optimization with the SMOTE algorithm for the class imbalance problem. *2017 6th Mediterranean Conference on Embedded Computing, MECO 2017 - Including ECYPS 2017, Proceedings*, (June), 17–20. https://doi.org/10.1109/MECO.2017.7977136

Fatemi, A., Ionel, D. M., Popescu, M., & Demerdash, N. A. O. (2016). Design optimization of spoke-type PM motors for Formula e racing cars. *ECCE 2016 - IEEE Energy Conversion Congress and Exposition, Proceedings*. https://doi.org/10.1109/ECCE.2016.7855032

Fatmawati, F., & Affandes, M. (2018). Klasifikasi Keluhan Menggunakan Metode Support Vector Machine (SVM) Pada Akun Facebook Group iRaise Helpdesk. *Jurnal CoreIT: Jurnal Hasil Penelitian Ilmu Komputer Dan Teknologi Informasi*, *3*(1), 24. https://doi.org/10.24014/coreit.v3i1.3552

Gata, W., Amsury, F., Wardhani, N. K., Sugiyarto, I., Sulistyowati, D. N., & Saputra, I. (2019). Informative Tweet Classification of the Earthquake Disaster Situation in Indonesia. *5th International Conference on Computing Engineering and Design, ICCED 2019*. https://doi.org/10.1109/ICCED46541.2019.9161135

Hall, T. J. (2017). An Analysis of Braking Behavior in Formula-E® Racing. *SAE Technical Papers*, *Part F1301*(September). https://doi.org/10.4271/2017-01-2533

Kartinawati, E., & Wiyawan, H. (2021). ' Penyelenggaraan Formula E Jakarta Pada Program Aiman Kompas Tv. *Jurnal Assosiativ*, *1*(1), 1–10.

Liu, X., Fotouhi, A., & Auger, D. J. (2020). Optimal energy management for formula-E cars with regulatory limits and thermal constraints. *Applied Energy*, *279*(September), 115805.

The work is distributed under the Creative Commons Attribution-NonCommercial 4.0 International License

https://doi.org/10.1016/j.apenergy.2020.115805

Mathew, J., Luo, M., Pang, C. K., & Chan, H. L. (2015). Kernel-based SMOTE for SVM classification of imbalanced datasets. *IECON 2015 - 41st Annual Conference of the IEEE Industrial Electronics Society*, 1127–1132. https://doi.org/10.1109/IECON.2015.7392251

Putu, N., Naraswati, G., Rosmilda, D. C., Desinta, D., Statistika, P. D., & Stis, P. S. (2021). Analisis Sentimen Publik dari Twitter Tentang Kebijakan Penanganan Covid-19 di Indonesia dengan Naive Bayes Classification. *Sistemasi: Jurnal Sistem Informasi*, *10*(1), 228–238. Retrieved from http://sistemasi.ftik.unisi.ac.id/index.php/stmsi/article/view/1179

Robeers, T. (2019). 'We go green in Beijing': situating live television, urban motor sport and environmental sustainability by means of a framing analysis of TV broadcasts of Formula E. *Sport in Society*, *22*(12), 2089–2103. https://doi.org/10.1080/17430437.2018.1558212

Tanjung, R. (2021). Instruksi Gubernur DKI Jakarta Tentang Penyelenggaraan Balap Formula E Dalam Tinjauan Siyasah Islam. *Al Ahkam*, *17*(2), 9–21. https://doi.org/10.37035/ajh.v17i2.5263

Tsytsarau, M., & Palpanas, T. (2012). Survey on mining subjective data on the web. *Data Mining and Knowledge Discovery*, *24*(3), 478–514. https://doi.org/10.1007/s10618-011-0238-6

Windasari, I. P., Uzzi, F. N., & Satoto, K. I. (2018). Sentiment analysis on Twitter posts: An analysis of positive or negative opinion on GoJek. *Proceedings - 2017 4th International Conference on Information Technology, Computer, and Electrical Engineering, ICITACEE 2017*, *2018-Janua*, 266–269. https://doi.org/10.1109/ICITACEE.2017.8257715

The work is distributed under the Creative Commons Attribution-NonCommercial 4.0 International License

The work is distributed under the Creative Commons Attribution-NonCommercial 4.0 International License