JSAI: Journal Scientific and Applied Informatics
Vol. 6, No. 1, Januari 2023, hal. 65~70
E-ISSN: 2614-3054; P-ISSN: 2614-3062, accredited by Kemenristekdikti, Sinta 4
DOI: 10.36085

Model *Extreme Gradient Boosting* Berbasis *Term Frequency* (TFXGBoost) Untuk Pengolahan Laporan Pengaduan Masyarakat

Vina Ayumi[1a], Desi Ramayanti[2b], Handrie Noprisson[1c], Yuwan Jumaryadi[1d], Umniy Salamah[1e]
[1]Fakultas Ilmu Komputer, Universitas Mercu Buana, Jakarta, Indonesia
[2]Fakultas Teknik dan Informatika, Universitas Dian Nusantara, Jakarta, Indonesia
[a]*vina.ayumi@mercubuana.ac.id,* [b]*desi.ramayanti@undira.ac.id,* [c]*handrie.noprisson@mercubuana.ac.id,*
[d]*yuwan.jumaryadi@mercubuana.ac.id,* [e]*umniy.salamah@mercubuana.ac.id*

Article Info

Article history:

Received, xxx xx xxxx
Revised, xxx xx xxxx
Accepted, xxx xx xxx

Kata Kunci:

TFXGBoost
Term Frequency
Pengaduan Masyarakat

ABSTRAK

Berbagai algoritma dan teknik pembelajaran mesin sedang diterapkan untuk meningkatkan efisiensi dan efektivitas proses klasifikasi laporan pengaduan secara otomatis dari masyarakat di Indonesia. Salah satu algoritma pembelajaran mesin yang baru-baru ini memperoleh benchmark dalam state of the art berbagai masalah dalam pemelajaran mesin adalah *eXtreme Gradient Boosting* (XGBoost). Penelitian ini bertujuan mengembangan model *extreme gradient boosting* berbasis *term frequency* (TFXGBoost) untuk memprediksi suatu teks apakah tergolong pengaduan atau bukan pengaduan berdasarkan data yang diteliti. Berdasarkan hasil eksperimen, TFXGBoost mencapai akurasi 92,79% dengan hyperparameter tingkat *eta* / pembelajaran sebesar 0,5, *gamma* sebesar 0, dan max_*depth* sebesar 3 dan waktu komputasi yang diperlukan untuk menyesuaikan *hyperparameter* adalah 13870.012468 detik.

Keywords:

TFXGBoost
Term Frequency
Community Complaint

ABSTRACT

Various algorithms and machine learning techniques are being applied to improve the efficiency and effectiveness of the process of automatically classifying complaint reports from the public in Indonesia. One machine learning algorithm that has recently gained benchmarks in the state of the art of various problems in machine learning is eXtreme Gradient Boosting (XGBoost). This study aims to develop an extreme gradient boosting model based on term frequency (TFXGBoost) to predict whether a text is classified as a complaint or not a complaint based on the data studied. Based on the experimental results, TFXGBoost achieved 92.79% accuracy with eta / learning rate hyperparameters of 0.5, gamma of 0, and max_depth of 3 and the computation time required to adjust the hyperparameters was 13870.012468 seconds.

Penulis Korespondensi:

Vina Ayumi,
Fakultas Ilmu Komputer
Universitas Mercu Buana, Jakarta, Indonesia
Email: vina.ayumi@mercubuana.ac.id

1. PENDAHULUAN

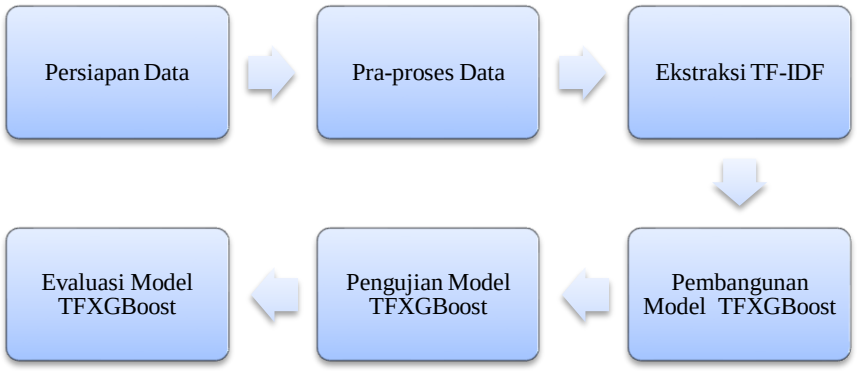
Model machine learning digunakan untuk mengklasifikasikan laporan pengaduan secara otomatis dari masyarakat di Indonesia. Berbagai algoritma dan teknik pembelajaran mesin sedang diterapkan untuk meningkatkan efisiensi dan efektivitas proses klasifikasi. Penelitian sebelumnya menggunakan metode *Recurrent Neural Network* (RNN), seperti Bidirectional *Long Short-Term Memory* (Bi-LSTM) dan *Gated Recurrent Unit* (GRU), bersama dengan teknik word embedding seperti Word2Vec dan FastTexts. Penelitian lain menerapkan algoritma *Naïve Bayes* (NB) untuk mengklasifikasikan keluhan dan laporan masyarakat, mencapai nilai akurasi tinggi 92%. Selain itu, sebuah proyek penelitian membandingkan beberapa model pembelajaran mesin, termasuk *Support Vector Machine* (SVM), *Random Forest* (RF), *Extreme Gradient Boosting* (XGBoost), dan *Adaptive Boosting* (AdaBoost) [1]–[10].

Salah satu algoritma pembelajaran mesin yang baru-baru ini memperoleh *benchmark* dalam *state of the art* berbagai masalah dalam pemelajaran mesin adalah eXtreme Gradient Boosting (XGBoost). XGBoost adalah teknik klasifikasi *supervised* yang menggunakan *ensamble decision trees*. Teknik ensamble boosting digunakan untuk meningkatkan *Taylor expansion* terhadap penurunan *loss function* [11], [12]. Model yang dibangun oleh XGBoost juga *insensitive* terhadap *imbalance data.* Algoritma ini banyak memenangkan di berbagai *data mining* maupun *machine learning challenge*, dan telah diterapkan di berbagai masalah dengan memberikan hasil yang baik [11], [13], [22], [14]–[21].

Dalam pengolahan teks, penggunaan *term frequency-inverse document frequency* (TF-IDF) dalam *text mining* dapat bermanfaat untuk menganalisis keluhan masyarakat. Penelitian yang dilakukan oleh Alamsyah et al. (2022) berfokus pada klasifikasi laporan pengaduan menggunakan ekstraksi fitur TF-IDF yang menghasilkan peningkatan akurasi untuk klasifikasi pengaduan [2]. Selain itu, Karo et al (2023) melakukan studi tentang mendeteksi hoaks dalam tweet Indonesia menggunakan pengklasifikasi *Naïve Bayes* dengan TF-IDF [23]–[25]. Penelitian ini berfokus pada klasifikasi teks berupa laporan teks pengaduan. Penelitian ini akan melakukan klasifikasi data teks menggunakan *XGBoost* yang dioptimasi ekstraksi fitur TF-IDF dengan untuk memprediksi suatu teks apakah tergolong pengaduan atau bukan pengaduan berdasarkan data yang diteliti.

2. METODE PENELITIAN

Tahapan proses eksperimen teks XGBoost dapat diringkas sebagai berikut. Pertama, fitur teks diekstraksi menggunakan pendekatan TF-IDF, dan kata-kata fitur dari laporan pengaduan dipilih dan diubah menjadi vektor kata bobot. Kemudian, sampel data dibagi menjadi set pelatihan dan tes berdasarkan data teks laporan pengaduan. Model TFXGBoost kemudian digunakan untuk klasifikasi teks pengaduan.



Gambar 1 Tahap Penelitian

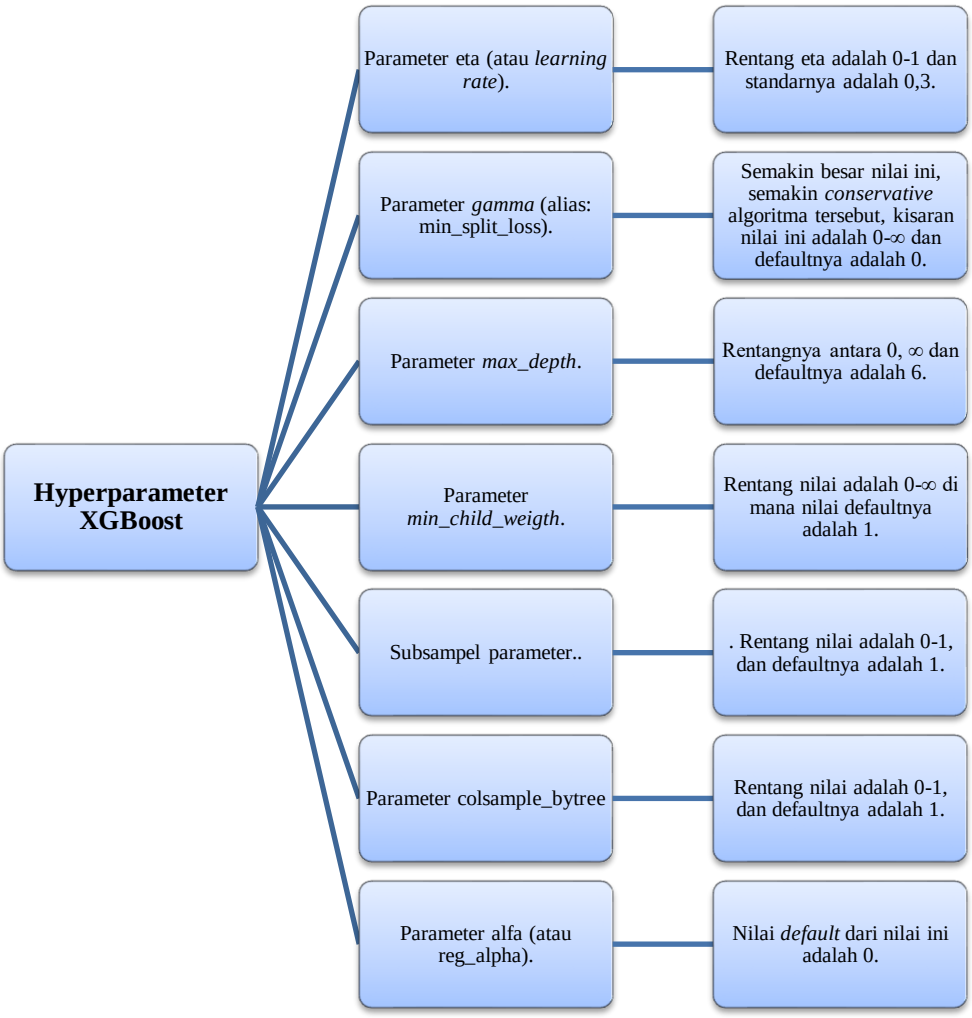
Data yang digunakan digunakan disini adalah hasil *crawling* menggunakan Twitter dan script R yang berisi teks yang bermakna laporan pengaduan. Setelah data diperoleh selanjutnya adalah melakukan pra-proses pada data. Pra-proses meliputi *data cleansing* dengan menghilangkan data duplikat (retweet), repetisi kata dan/atau kalimat, penghilangan '@username' (nama user), URL (alamat web), 'RT' (tanda retweet), tanda baca, dan karakter-karakter spesial, serta penghilangan *stop word*. Setelah itu data dilakukan pelabelan dengan memberikan label untuk setiap data apakah tergolong sentiment positif, negatif, atau netral. Setelah diperoleh data yang telah bersih dan memiliki label selanjutnya adalah mengekstraksi data menjadi fitur, sehingga akan menghasilkan vektor. Metode ekstraksi yang digunakan adalah fitur TF-IDF.

JSAI: Journal Scientific and Applied Informatics
Vol. 6, No. 1, Januari 2023, hal. 65~70
E-ISSN: 2614-3054; P-ISSN: 2614-3062, accredited by Kemenristekdikti, Sinta 4
DOI: 10.36085

Setelah data dalam bentuk vektor selanjutnya dilakukan pemisahan data untuk pelatihan dan pengujian. Pemisahan data dilakukan dengan cross validasi dengan proporsi data 70% untuk pelatihan dan 30% untuk pengujian. Data pelatihan kemudian akan digunakan untuk membangun model. Model klasifikasi akan dilatih menggunakan data pelatihan sehingga dihasilkan model yang paling optimal. Model TFXGBoost yang dihasilkan pada tahap pelatihan selanjutnya dilakukan pengujian menggunakan data uji. Pengujian dilakukan dengan memprediksi label untuk data pengujian. Evaluasi model TFXGBoost dilakukan dengan mengukur hasil akurasi prediksi. Selain itu juga dihitung pengukuran kappa statistic, precision dan recall.

3. HASIL DAN PEMBAHASAN

XGBoost adalah model pembelajaran mesin yang banyak digunakan, dan hyperparameters memainkan peran penting dalam kinerjanya. *GridSearch* adalah metode yang digunakan untuk menemukan hyperparameter optimal untuk XGBoost. Penerapan *GridSearch* untuk mengoptimalkan model XGBoost untuk emperkirakan *reservoir porosity*, mencapai akurasi tinggi dan menghindari *overfitting*. Dalam penelitian model XGBoost yang dioptimalkan dengan ekstraksi fitur TF-IDF diharapkan dapat meningkatkan kinerja XGBoost. Hyperparameter dari model XGBoost yang diatur dalam eksperimen dapat dilihat pada **Gambar 2**.



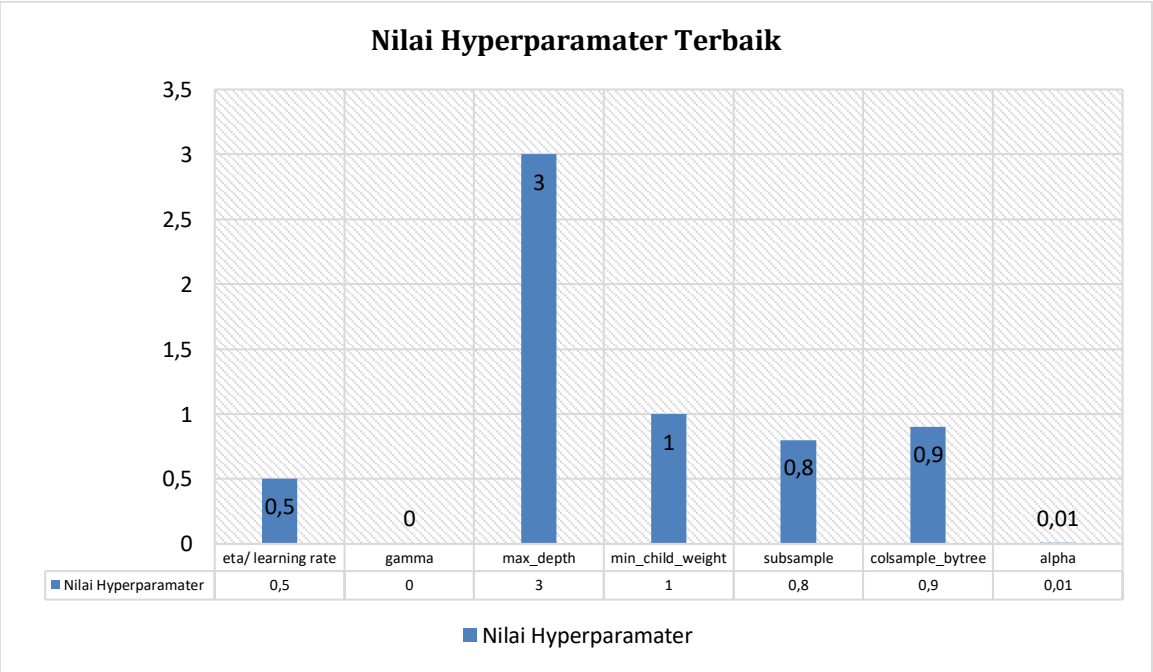
Gambar 2 Hyperparameter XGBoost

XGBoost adalah teknik pembelajaran mesin yang dapat digunakan untuk tugas klasifikasi dan regresi. Model ini memiliki beberapa hyperparameter yang dapat dikonfigurasi untuk meningkatkan kinerjanya. Hyperparameter meliputi min_*child_weight*, *subsample*, dan *colsample_bytree*. Hyperparameter ini dapat diaturs menggunakan teknik seperti grid search untuk menemukan nilai optimal yang meningkatkan kinerja model XGBoost. Skenario eksperimen untuk nilai hyperparameter dapat dilihat pada **Tabel 1**.

Tabel 1 Skenario Eksperimen Hyperparameter

Hyperparameter	Nilai
eta/ learning rate	0.01, 0.1, 0.25, 0.5
gamma	0.1, 0.2, 0.3, 0.4, 0.5
max_depth	2, 3, 5, 10
min_child_weight	1, 6, 2
subsample	0.6, 0.7, 0.8, 0.9, 1
colsample_bytree	0.6, 0.7, 0.8, 0.9, 1
alpha	1e-5, 1e-2, 0.1, 1, 100

Nilai optimal untuk parameter XGBoost *eta/learning rate*, *gamma*, dan *max_depth* bervariasi tergantung pada data yang dianalisis. Dalam konteks pengolahan data teks pengaduan berdasarkan ekstraksi fitur TF-IDF, model yang diusulkan mencapai kinerja terbaik dengan tingkat pembelajaran sebesar 0,5, gamma sebesar 0, dan max_depth sebesar 3. Selain itu, ringkasan hasil mengenai nilai terbaik untuk setiap parameter disajikan pada **Gambar 3**.



Gambar 3 Hyperparameter XGBoost Terbaik

Algoritma XGBoost telah terbukti memiliki akurasi tinggi dalam klasifikasi teks pengaduan masyarakat. Dalam eksperimen ini, XGBoost mencapai akurasi 92,79% dalam tugas klasifikasi kelas *binary* menggunakan metode ekstraksi fitur TF-IDF. Studi ini menemukan bahwa XGBoost mengungguli algoritma lain seperti KNN, Naive Bayes, dan SVM dalam mengklasifikasikan teks pengaduan masyarakat. Secara keseluruhan, XGBoost telah terbukti efektif dalam mencapai akurasi tinggi dalam berbagai tugas klasifikasi teks pengaduan masyarakat dengan waktu komputasi yang diperlukan untuk menyesuaikan hyperparameter XGBoost adalah 13870.012468 detik.

4. KESIMPULAN

Penelitian ini bertujuan mengembangan model *extreme gradient boosting* berbasis *term frequency* (TFXGBoost) untuk pengolahan laporan pengaduan masyarakat. Penelitian ini akan melakukan klasifikasi data teks menggunakan XGBoost yang dioptimasi ekstraksi fitur TF-IDF dengan untuk memprediksi suatu teks apakah tergolong pengaduan atau bukan pengaduan berdasarkan data yang diteliti. Dalam eksperimen ini, XGBoost mencapai akurasi 92,79% dalam tugas klasifikasi multi-kelas menggunakan metode ekstraksi fitur TF-IDF dengan hyperparameter terbaik dengan tingkat eta / pembelajaran 0,5, gamma 0, dan max_depth 3. dan waktu komputasi yang diperlukan untuk menyesuaikan hyperparameter XGBoost adalah 13870.012468 detik.

UCAPAN TERIMA KASIH

Terima kasih kepada Pusat Penelitian dan Fakultas Ilmu Komputer, Universitas Mercu Buana yang telah mendukung pelaksanaan penelitian ini.

REFERENSI

[1] O. A. Lisjana and M. L. Khodra, "Classifying Complaint Reports Using RNN and Handling Imbalanced Dataset," pp. 303–307, 2022.

[2] K. Wabang, O. D. Nurhayati, and Farikhin, "Application of The Naïve Bayes Classifier Algorithm to Classify Community Complaints," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 6, no. 5, pp. 872–876, 2022.

[3] H. Sibyan, W. Suharso, E. Suharto, M. A. Manuhutu, and A. P. Windarto, "Optimization of Unsupervised Learning in Machine Learning," vol. 1783, no. 1, p. 12034, 2021.

[4] M. Purba *et al.*, "Effect of Random Splitting and Cross Validation for Indonesian Opinion Mining using Machine Learning Approach," *Int. J. Adv. Comput. Sci. Appl.*, vol. 13, no. 9, 2022.

[5] M. Purba, E. Ermatita, A. Abdiansah, V. Ayumi, H. Noprisson, and A. Ratnasari, "A Systematic Literature Review of Knowledge Sharing Practices in Academic Institutions," in *2021 International Conference on Informatics, Multimedia, Cyber and Information System (ICIMCIS*, 2021, pp. 337–342.

[6] D. Ramayanti *et al.*, "Tuberculosis Ontology Generation and Enrichment Based Text Mining," in *2020 International Conference on Information Technology Systems and Innovation (ICITSI)*, 2020, pp. 429–434.

[7] D. I. Sensuse, P. Prima, E. Cahyaningsih, and H. Noprisson, "Knowledge management practices in e-Government," in *2017 3rd International Conference on Science in Information Technology (ICSITech)*, 2017.

[8] H. Noprisson, "Analysis and Design e-Government Website for Special Allocation Fund," *Int. J. Comput. Sci. Eng.*, vol. 8, no. 02, 2019.

[9] M. Sadikin and A. Fauzan, "Evaluation of Machine Learning Approach for Sentiment Analysis using Yelp Dataset," *Eur. J. Electr. Eng. Comput. Sci.*, vol. 7, no. 6, pp. 58–64, 2023.

[10] M. Sadikin, M. I. Fanany, and T. Basaruddin, "A New Data Representation Based on Training Data Characteristics to Extract Drug Name Entity in Medical Text," *Comput. Intell. Neurosci.*, vol. 2016, 2016.

[11] T. Chen and C. Guestrin, "XGBoost," *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Min. - KDD '16*, pp. 785–794, 2016.

[12] B. Ghaddar and J. Naoum-sawaya, "High dimensional data classification and feature selection using support vector machines," *Eur. J. Oper. Res.*, vol. 0, pp. 1–12, 2017.

[13] A. F. Hidayatullah, C. I. Ratnasari, and S. Wisnugroho, "Analysis of Stemming Influence on Indonesian Tweet Classification," *Telkomnika (Telecommunication Comput. Electron. Control.*, vol. 14, no. 2, pp. 665–673, 2016.

[14] D. A. Firdlous, R. Andrian, and S. Widodo, "Sentiment Analysis Public Twitter on 2024 Election using the Long Short Term Memory Model," *J. Sist. Inf.*, vol. 12, no. 1, p. 52, 2023.

[15] H. Setyawan and L. M. Azizah, "Sentiment Analysis of Public Responses on Indonesia Government Using Naïve Bayes and Support Vector Machine," *Emerg. Inf. Sci. Technol.*, vol. 4, no. 1, pp. 1–7, 2023.

[16] A. D. Akmal, I. Permana, H. Fajri, and Y. Yuliarti, "Opini Masyarakat Twitter terhadap Kandidat Bakal Calon Presiden Republik Indonesia Tahun 2024," *J. Manaj. dan Ilmu Adm. Publik*, vol. 4, no. 4, pp. 292–300, 2022.

[17] D. Fitrianah and A. H. Wangsa, "Text classification to predict skin concerns over skincare using bidirectional mechanism in long short-term memory," *Comput. Sci. Inf. Technol.*, vol. 3, no. 3, pp. 137–

147, 2022.
[18] H. D. Wijaya, W. Gunawan, R. Avrizal, and S. M. Arif, "Designing chatbot for college information management," *IJISCS (International J. Inf. Syst. Comput. Sci.*, vol. 4, no. 1, pp. 8–13, 2020.
[19] H. D. Wijaya and W. Gunawan, "Implementation of Analytic Network Process Algorithm in E-Lowker System," *J. Syst. Eng. Inf. Technol.*, vol. 1, no. 1, pp. 18–26, 2022.
[20] Y. Devianto and S. Dwiasnati, "Application of Community Satisfaction Index in Service Units With Average Calculation Method," *Int. J. Comput. Sci. Inf. Secur.*, vol. 17, no. 8, 2019.
[21] M. Utami and D. Sunardi, "Pemodelan Arsitektur Mobile Commerce Usaha Mikro Menggunakan EAP Dan Togaf ADM Framework," *INTECOMS J. Inf. Technol. Comput. Sci.*, vol. 3, no. 2, pp. 290–297, 2020.
[22] E. D. Putra, E. Hidayat, and H. Noprisson, "Model Mobile Positioning System Berbasis Android," vol. III, no. September, pp. 113–121, 2016.
[23] D. P. Alamsyah, T. Arifin, Y. Ramdhani, F. A. Hidayat, and L. Susanti, "Classification of Customer Complaints: TF-IDF Approaches," pp. 1–5, 2022.
[24] A. Rahmat and N. P.-U. N. Mandiri, "Penerapan online service pekerja migran indonesia pada pt intan ayu lestari karawang," *IJNS - Indones. J. Netw. Secur.*, vol. 10, no. 4, 2022.
[25] I. M. K. Karo, S. A. A. K. Dewi, and P. M. Fadilah, "Hoax Detection on Indonesian Tweets using Naïve Bayes Classifier with TF-IDF," *J. Inf. Syst. Res.*, vol. 4, no. 3, pp. 914–919, 2023.