TIN: Terapan Informatika Nusantara
Vol 6, No 12, May 2026, page 2384-2392
ISSN 2722-7987 (Media Online)
Website https://ejurnal.seminar-id.com/index.php/tin
DOI 10.47065/tin.v6i12.9891

Analisis Komparatif MLP dan GraphSAGE dalam Deteksi Bot Twitter/X pada Benchmark TwiBot-22

Mochammad Fikri Chaerul Chalik Ramdhan*, Sigit Puspito Wigati Jarot

Program Studi Teknik Informatika, Sekolah Tinggi Teknologi Terpadu Nurul Fikri, Depok, Indonesia
Email: [1,*]chaerulchalikr@gmail.com, [2]sigit.jarot@nurulfikri.ac.id
Email Penulis Korespondensi: chaerulchalikr@gmail.com

Abstrak–Penyebaran akun bot pada Twitter/X tetap menjadi persoalan penting karena memengaruhi integritas informasi, kualitas diskursus publik, dan keandalan moderasi platform. Artikel ini mengevaluasi dua pendekatan deteksi bot pada *benchmark* TwiBot-22, yaitu Multilayer Perceptron (MLP) berbasis fitur profil dan GraphSAGE berbasis graf sosial, menggunakan Kerangka Kerja Evaluasi 12-Tahap yang menata validasi data, rekayasa fitur, pelatihan model, analisis ambang, ablasi fitur, dan evaluasi *multi-seed* secara terstruktur. Eksperimen dibatasi pada konteks *offline benchmark* dengan 1.000.000 akun berlabel, komposisi 13,99% bot dan 86,01% manusia, serta partisi tetap 70% pelatihan, 20% validasi, dan 10% pengujian. Hasil *single-seed* pada konfigurasi 15 fitur menunjukkan MLP mencapai F1(bot) 0,53 dan PR-AUC 0,48, sedangkan GraphSAGE mencapai F1(bot) 0,53 dan PR-AUC 0,46. Pada evaluasi konfirmatori tiga seed, konfigurasi user_only_8 menghasilkan F1(bot) 0,53 dan PR-AUC 0,49 dengan varians rendah, sementara all_15 menghasilkan F1(bot) 0,53 dan PR-AUC 0,47 dengan varians lebih tinggi. Temuan ini menunjukkan bahwa konfigurasi fitur profil yang lebih hemat tetap mempertahankan kualitas keputusan biner yang setara, memberi kualitas *ranking* probabilitas yang lebih baik, dan memiliki varians lebih rendah. Dengan demikian, kontribusi utama artikel ini adalah penegasan berbasis benchmark bahwa efisiensi fitur dapat menjadi pertimbangan yang lebih relevan daripada penambahan kompleksitas graf dan fitur pada TwiBot-22.

Kata Kunci: Deteksi Bot; Twitter/X; TwiBot-22; MLP; GraphSAGE; PR-AUC

Abstract–Bot accounts on Twitter/X remain a significant challenge because they affect information integrity, distort public discourse, and complicate platform moderation. This article evaluates two bot detection approaches on the TwiBot-22 benchmark: a profile-feature-based Multilayer Perceptron (MLP) and a graph-based GraphSAGE model, using a 12-Stage Evaluation Framework that covers data validation, feature engineering, model training, threshold analysis, feature ablation, and multi-seed evaluation. The study is limited to an offline benchmark setting with 1,000,000 labeled accounts, 13.99% bots and 86.01% humans, and a fixed split of 70% training, 20% validation, and 10% testing. In the single-seed 15-feature comparison, MLP achieved F1(bot) of 0.53 and PR-AUC of 0.48, while GraphSAGE reached F1(bot) of 0.53 and PR-AUC of 0.46. In the confirmatory three-seed evaluation, the user_only_8 configuration produced F1(bot) of 0.53 and PR-AUC of 0.49 with lower variance, whereas all_15 produced F1(bot) of 0.53 and PR-AUC of 0.47 with higher variance. These findings indicate that the more economical profile-only configuration preserves practically identical binary-decision quality, offers better probability ranking quality, and shows lower variance. The main contribution of this article is a feature-economy argument: on TwiBot-22, added graph and feature complexity does not automatically yield proportionate practical gains.

Keywords: Bot Detection; Twitter/X; TwiBot-22; MLP; GraphSAGE; PR-AUC

1. PENDAHULUAN

Twitter/X telah berkembang menjadi ruang komunikasi publik yang sangat berpengaruh, tetapi keterbukaan platform ini juga memberi ruang luas bagi akun otomatis untuk ikut membentuk arus informasi. Dalam literatur mutakhir, bot media sosial tidak lagi dipahami semata-mata sebagai akun yang berjalan otomatis, melainkan sebagai bagian dari perilaku inautentik yang dapat dipakai untuk memperbesar jangkauan narasi, mengganggu percakapan, dan memanipulasi persepsi publik melalui koordinasi yang sulit diamati pada pembacaan kasatmata (Chen et al., 2021; Cinelli et al., 2022; Mazza et al., 2022). Bot dapat tampil sebagai penyebar tautan, pengulang tagar, penguat kampanye tertentu, atau sekadar pembentuk ilusi bahwa suatu opini memperoleh dukungan luas. Dalam konteks ini, masalah deteksi bot bukan lagi isu teknis kecil, melainkan bagian dari persoalan integritas informasi digital.

Tantangan deteksi bot saat ini semakin kompleks, terutama dengan hadirnya model bahasa generatif (AI) yang mampu memproduksi teks mirip buatan manusia. Kondisi ini membuat analisis yang hanya mengandalkan teks permukaan tidak lagi memadai (Ferrara, 2023). Oleh karena itu, pendekatan deteksi bot modern telah bergeser ke arah analisis multidimensi yang memanfaatkan deep learning dan struktur graf (Feng, Tan, Li, et al., 2022; Huang et al., 2025; Li et al., 2026; Najari et al., 2022) . Pendekatan ini menggabungkan berbagai sinyal, mulai dari fitur profil pengguna hingga relasi jaringan antar-akun. Meskipun model-model canggih ini menawarkan representasi data yang lebih kaya, peningkatan kompleksitas arsitektur tersebut juga membawa tantangan baru dalam hal biaya komputasi dan evaluasi kinerja yang adil.

Di sisi lain, perkembangan metodologi tidak otomatis menyelesaikan masalah evaluasi. Pembatasan akses data Twitter/X sejak 2023 memperumit replikasi penelitian dan mengurangi peluang membangun dataset baru yang sebanding dalam skala besar. Küpfer (2024) dan Blakey (2024) menunjukkan bahwa hambatan akses data bukan sekadar persoalan teknis, tetapi juga ancaman terhadap transparansi ilmiah karena hasil penelitian menjadi lebih sulit diverifikasi secara independen. Dalam situasi seperti ini, *benchmark* publik yang telah terdokumentasi dengan baik menjadi sangat penting (Feng, Wan, Wang, Li, et al., 2021). Artikel ini karena itu menempatkan TwiBot-22 sebagai konteks evaluasi utama, bukan sebagai klaim representasi penuh terhadap seluruh dinamika platform mutakhir.

TIN: Terapan Informatika Nusantara
Vol 6, No 12, May 2026, page 2384-2392
ISSN 2722-7987 (Media Online)
Website https://ejurnal.seminar-id.com/index.php/tin
DOI 10.47065/tin.v6i12.9891

TwiBot-22 penting karena menyediakan 1.000.000 akun berlabel dengan 139.943 akun bot dan 860.057 akun manusia, disertai pembagian tetap 700.000 akun pelatihan, 200.000 validasi, dan 100.000 pengujian. Skala ini memberi landasan yang relatif kuat untuk membandingkan model pada kondisi yang identik. Namun, *benchmark* tersebut juga membawa dua konsekuensi metodologis yang harus dihadapi secara eksplisit. Pertama, distribusi kelas sangat tidak seimbang karena bot hanya 13,99% dari keseluruhan akun. Pada situasi seperti ini, metrik akurasi tunggal dapat menyesatkan karena model yang selalu memilih kelas mayoritas masih dapat terlihat baik secara dangkal. Literatur evaluasi klasifikasi pada data tidak seimbang menegaskan bahwa interpretasi harus lebih menekankan metrik yang sensitif terhadap kelas minoritas, khususnya F1 untuk kelas bot dan precision-recall area under curve (PR-AUC) (de la Cruz Huayanay et al., 2024; Williams, 2021). Kedua, data relasi graf yang benar-benar termanfaatkan dalam jalur eksperimen ini hanya tersedia pada sekitar 33,8% akun, sehingga manfaat teoretis model berbasis *message passing* tidak boleh diasumsikan muncul secara otomatis.

Beberapa penelitian terdahulu telah berupaya mengembangkan arsitektur untuk mengatasi persoalan tersebut, namun mayoritas berfokus pada peningkatan kompleksitas model tanpa membandingkannya dengan solusi sederhana secara ketat. Pertama, Hayawi et al. (2022) mengusulkan DeeProBot, sebuah model jaringan saraf hibrida berbasis fitur profil yang menunjukkan efektivitas tinggi, tetapi belum dievaluasi trade-off-nya ketika dihadapkan pada struktur graf yang terfragmentasi. Kedua, Feng, Wan, Wang, & Luo (2021) mengembangkan BotRGCN yang mengeksploitasi relasi graf pengguna; namun, pendekatan ini sangat bergantung pada ketersediaan data relasional yang utuh (*dense*). Ketiga, guna mengatasi keterbatasan graf yang tidak lengkap, Wei et al. (2024) memperkenalkan BotGSL berbasis *graph structure learning*, yang sayangnya berimplikasi pada peningkatan biaya komputasi yang tajam. Keempat, Wang et al. (2025) mengajukan HHG-Bot dengan graf hiperheterogen untuk menangkap struktur jaringan secara lebih mendalam, tetapi evaluasinya tidak secara langsung menimbang kelayakan efisiensi fitur dengan *baseline* arsitektur standar.

Berdasarkan analisis komparatif dari keempat penelitian tersebut, terlihat adanya celah atau kesenjangan penelitian (*research gap*). Mayoritas studi berlomba membangun arsitektur graf (GNN) yang semakin rumit demi menaikkan akurasi performa, tetapi mereka kerap mengabaikan bahwa pada data nyata seperti *benchmark* TwiBot-22, sebagian besar *node* (sekitar 66,2%) justru kekurangan struktur relasional yang kaya. Belum ada pengujian empiris dan komparatif yang menelaah apakah tambahan kerumitan model graf benar-benar menghasilkan keunggulan yang memadai dibandingkan *baseline* konvensional pada kondisi *sparse network* (jaringan jarang). Bagi praktisi dan peneliti dengan sumber daya terbatas, pertanyaan mendesak yang muncul bukan "model mana yang paling canggih", melainkan "apakah tambahan kompleksitas itu benar-benar memberi keuntungan praktis yang sepadan".

Di sinilah kontribusi artikel ini ditempatkan. Artikel ini membandingkan dua arsitektur, yaitu Multilayer Perceptron (MLP) sebagai *baseline* berbasis fitur profil dan GraphSAGE sebagai model berbasis graf, menggunakan dua konfigurasi fitur: *all_15* dan *user_only_8*. Perbandingan dilakukan melalui Kerangka Kerja Evaluasi 12-Tahap yang mendokumentasikan validasi data, rekayasa fitur, pelatihan model, analisis ambang (*threshold*), ablasi fitur, dan evaluasi *multi-seed* secara terstruktur. Artikel ini tidak bermaksud mengklaim superioritas absolut konfigurasi tertentu, melainkan bertujuan memberikan pembuktian terukur pada *benchmark* bahwa strategi yang lebih hemat fitur dapat tetap kompetitif, memiliki varians yang rendah, dan menghasilkan kualitas *ranking* probabilitas yang lebih unggul.

Berdasarkan latar belakang tersebut, artikel ini dirumuskan untuk menjawab tiga tujuan utama. Pertama, menguji apakah *baseline* MLP dapat menandingi GraphSAGE pada *benchmark* TwiBot-22 ketika keduanya dijalankan di bawah protokol evaluasi yang identik. Kedua, memastikan apakah konfigurasi *user_only_8* cukup solid dan andal untuk dijadikan standar efisiensi fitur dibandingkan konfigurasi agregat *all_15*. Ketiga, menganalisis fitur spesifik apa yang paling informatif melalui teknik ablasi. Jawaban dari penelitian ini akan memberikan pedoman metodologis bagi peneliti agar dapat memilih model pendeteksi bot yang efisien, hemat daya komputasi, dan dapat direproduksi (*reproducible*) tanpa mengorbankan performa klasifikasi yang krusial.

2. METODOLOGI PENELITIAN

2.1 Kerangka Penelitian

Penelitian ini menggunakan pendekatan kuantitatif eksperimental dengan rancangan *offline benchmark*. Seluruh eksperimen dibatasi pada dataset TwiBot-22 yang tersedia secara lokal, tanpa melakukan pengambilan data baru melalui API Twitter/X dan tanpa pengujian pada domain di luar *benchmark* tersebut. Pembatasan ini diterapkan secara sengaja untuk memastikan bahwa fokus penelitian berpusat pada perbandingan model yang adil, ketertelusuran artefak eksperimen, dan keterulangan prosedur (*reproducibility*). Kondisi akses data platform media sosial yang semakin tertutup saat ini membuat kualitas dan transparansi protokol evaluasi menjadi sama pentingnya dengan pemilihan arsitektur model itu sendiri (Blakey, 2024; Küpfer, 2024).

Secara operasional, eksperimen dijalankan melalui Kerangka Kerja Evaluasi 12-Tahap, yaitu sebuah *pipeline* sekuensial berbasis fail yang terdiri dari dua belas tahap. Tahap 01 sampai 05 berfokus pada validasi skema dataset, penyiapan label dan pemisahan data (*data splitting*), pembangunan graf relasi pengguna (*user-to-user graph*), ekstraksi delapan fitur profil, dan perakitan artefak data menggunakan pustaka komputasi graf (Fey & Lenssen, 2019). Tahap 07 sampai 08 menangani pengayaan fitur dari riwayat perilaku *tweet*, sedangkan Tahap 06 dan 09 didedikasikan untuk melatih model *baseline* tabular dan model graf yang diperkaya (*augmented*). Tahap 10 melaksanakan analisis ambang batas (*threshold*), Tahap 11 menjalankan ablasi fitur untuk menguji kontribusi setiap variabel, dan Tahap 12 berfungsi

TIN: Terapan Informatika Nusantara
Vol 6, No 12, May 2026, page 2384-2392
ISSN 2722-7987 (Media Online)
Website https://ejurnal.seminar-id.com/index.php/tin
DOI 10.47065/tin.v6i12.9891

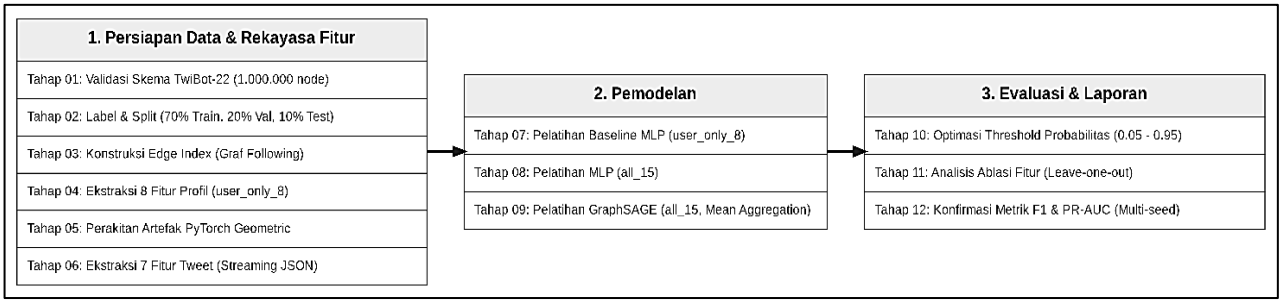
sebagai fase evaluasi konfirmatori menggunakan metode *multi-seed*. Seluruh konfigurasi dieksekusi melalui berkas YAML tunggal untuk meminimalisasi *human error*, di mana setiap eksekusi akan menghasilkan manifes yang mencatat status tahapan beserta artefak keluarannya. Desain sistematis ini sangat penting karena memungkinkan eksperimen dilanjutkan (*resume*) jika terjadi kegagalan proses komputasi, sekaligus mendukung audit yang konsisten pada seluruh alur eksperimen.

2.2 Dataset dan Pembagian Data

Dataset utama yang digunakan dalam penelitian ini adalah TwiBot-22, sebuah tolok ukur berskala besar yang secara spesifik dikembangkan untuk mendeteksi bot berbasis graf (Feng, Tan, Wan, et al., 2022). Dataset ini mencakup total 1.000.000 akun berlabel, yang terdiri atas 860.057 akun manusia dan 139.943 akun bot. Distribusi kelas pada dataset ini menghasilkan rasio ketimpangan (*class imbalance*) yang signifikan, yakni sekitar 13,99% entitas bot berbanding 86,01% entitas manusia. Pembagian partisi data secara ketat mengikuti protokol standar dari *benchmark* TwiBot-22 untuk memastikan perbandingan yang valid. Secara rincinya, sebanyak 700.000 akun dialokasikan untuk proses pelatihan (*training*), 200.000 akun untuk tahap validasi (*validation*), dan 100.000 akun disisihkan secara eksklusif untuk pengujian performa akhir (*testing*).

2.3 Tahapan Kerangka Kerja Evaluasi

Kerangka kerja evaluasi dirancang untuk memastikan seluruh proses, mulai dari prapemrosesan data awal hingga pelaporan metrik pengujian model, dapat ditelusuri secara komprehensif. Rincian alur arsitektural dari kerangka kerja dua belas tahap ini divisualisasikan pada Gambar 1, yang membagi alur komputasi ke dalam tiga blok utama.



Gambar 1. Alur kerangka kerja evaluasi deteksi bot

Gambar 1 merangkum urutan kerja evaluasi dari validasi data hingga evaluasi *multi-seed* yang dikonsolidasikan ke dalam tiga blok utama. Bagian pertama mengintegrasikan persiapan data dan rekayasa fitur (Tahap 01–06) agar sumber representasi fitur numerik maupun struktural dapat ditelusuri pembuatannya secara terpusat. Bagian tengah difokuskan sepenuhnya pada optimalisasi dan pelatihan model (Tahap 07–09), sementara bagian akhir memperlihatkan bahwa mekanisme analisis ambang dan ablasi fitur wajib dijalankan sebelum konfirmasi *multi-seed* akhir (Tahap 10–12) untuk menjaga integritas dan disiplin evaluasi. Dengan susunan yang ketat ini, interpretasi analitis pada bagian hasil dan pembahasan dapat dipetakan langsung kembali ke fase *pipeline* yang relevan tanpa menghasilkan klaim yang berlebihan di luar cakupan *benchmark*.

2.4 Konfigurasi Fitur dan Model

Konfigurasi fitur dalam penelitian ini dibagi menjadi dua skenario eksperimen. Konfigurasi pertama, *user_only_8*, memuat delapan fitur profil pengguna yang memiliki biaya komputasi rendah saat diekstraksi dari metadata akun. Delapan fitur tersebut meliputi *followers_count*, *following_count*, *tweet_count*, *listed_count*, *ff_ratio*, *account_age_days*, *verified*, dan *protected*. Konfigurasi kedua, *all_15*, menambahkan tujuh fitur perilaku yang diagregat dari korpus *tweet* pengguna, yaitu *tweet_count_seen*, *avg_text_len*, *url_ratio*, *mention_ratio*, *hashtag_ratio*, *retweet_ratio*, dan *reply_ratio*. Ekstraksi fitur berbasis *tweet* ini memerlukan pemrosesan secara *streaming* pada korpus berskala besar. Pemisahan kedua konfigurasi ini krusial untuk menguji hipotesis seleksi fitur (Matharaarachchi et al., 2021) , yakni apakah penambahan biaya komputasi dari ekstraksi fitur riwayat *tweet* sebanding dengan peningkatan performa model dibandingkan jika hanya menggunakan fitur profil.

Dua arsitektur algoritma *deep learning* dikomparasikan secara *head-to-head*. Model pertama adalah *Multilayer Perceptron* (MLP), yang diimplementasikan sebagai *baseline* non-relasional yang beroperasi sepenuhnya pada data tabular. Model kedua adalah GraphSAGE, sebuah algoritma *Graph Neural Network* (GNN) spasial yang mengumpulkan fitur dari ketetanggaan *node* (relasi *following*) melalui fungsi agregasi *mean*, sebagaimana diimplementasikan pada ekosistem tolok ukur ini (Feng, Tan, Wan, et al., 2022) . Kedua model dioptimalkan menggunakan algoritma Adam, nilai *weight decay* 5×10^{-4}, penerapan pembobotan kelas (*class weighting*) untuk mitigasi ketidakseimbangan data, dan mekanisme *early stopping* berbasis data validasi

TIN: Terapan Informatika Nusantara
Vol 6, No 12, May 2026, page 2384-2392
ISSN 2722-7987 (Media Online)
Website https://ejurnal.seminar-id.com/index.php/tin
DOI 10.47065/tin.v6i12.9891

2.5 Protokol Evaluasi

Protokol evaluasi dipecah menjadi fase eksplorasi dan fase konfirmasi. Fase eksplorasi mencakup pengujian *single-seed* (Tahap 09), analisis *threshold* (Tahap 10), dan eksperimen ablasi (Tahap 11). Tahap 10 secara khusus menelusuri fluktuasi kinerja pada rentang *threshold* probabilitas 0,05 hingga 0,95 untuk mencari titik klasifikasi optimal berbasis data validasi. Analisis ablasi (Tahap 11) menjalankan skenario *leave-one-out* untuk menakar nilai informasi dari tiap fitur maupun blok fitur secara independen. Fase konfirmasi (Tahap 12) mengeksekusi pengujian ulang model menggunakan tiga *random seed* (42, 7, 123) untuk membuktikan keandalan performa model terhadap inisialisasi bobot yang berbeda. Oleh karena dataset memiliki ketimpangan ekstrem, evaluasi ini merujuk pada prinsip analisis performa kelas minoritas, di mana metrik F1 untuk kelas bot dan nilai *Precision-Recall Area Under Curve* (PR-AUC) digunakan sebagai parameter keberhasilan utama, bukan metrik akurasi tunggal yang bias (de la Cruz Huayanay et al., 2024; Williams, 2021).

3. HASIL DAN PEMBAHASAN

3.1 Perbandingan MLP dan GraphSAGE

Lapisan bukti pertama berasal dari evaluasi *single-seed* Tahap 09 pada konfigurasi all_15. Pada titik ini, MLP mencapai F1(bot) 0,53, PR-AUC 0,48, presisi bot 0,40, recall bot 0,80, akurasi 0,59, dan ROC-AUC 0,71. GraphSAGE pada konfigurasi yang sama mencapai F1(bot) 0,53, PR-AUC 0,46, presisi bot 0,40, recall bot 0,79, akurasi 0,58, dan ROC-AUC 0,70. Selisih absolutnya memang tidak sangat besar, tetapi arahnya konsisten pada metrik yang paling relevan untuk kelas bot, yaitu F1(bot) dan PR-AUC. Rincian performa komparatif dari kedua model pada tahap *single-seed* ini dapat dilihat secara lengkap pada Tabel 1.

Tabel 1. Hasil pengujian *single-seed* Tahap 09 pada konfigurasi all_15

Model	Ambang	F1(bot)	PR-AUC	Presisi(bot)	Recall(bot)	Akurasi	ROC-AUC
MLP	0,50	0,5342	0,4782	0,4009	0,8003	0,5890	0,7066
GraphSAGE	0,55	0,5278	0,4590	0,3952	0,7940	0,5816	0,6982

Tabel 1 menegaskan konteks *single-seed* sebagai pembacaan awal, bukan simpulan final. Pada konfigurasi 15 fitur, MLP sedikit lebih baik daripada GraphSAGE pada metrik kelas bot. Hasil ini dipakai sebagai pijakan sebelum evaluasi *multi-seed* pada Tabel 2.

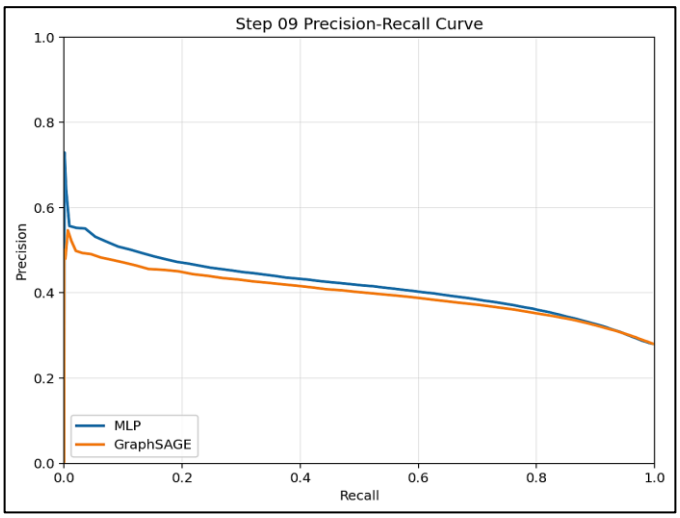
Interpretasi hasil ini perlu mempertimbangkan konteks struktur data. Secara teoritis, GraphSAGE dirancang untuk memperoleh keuntungan dari agregasi pesan antar-node. Namun pada *benchmark* yang dianalisis, hanya sekitar 33,8% akun memiliki relasi following yang termanfaatkan, sehingga sekitar 66,2% akun tidak memperoleh konteks tetangga yang substansial. Pada node-node tersebut, transformasi yang dilakukan model graf secara praktis mendekati pemetaan fitur individual, bukan eksploitasi struktur relasional yang kaya. Karena itu, kegagalan GraphSAGE melampaui MLP lebih tepat dibaca sebagai temuan *benchmark*-spesifik tentang kualitas dan cakupan graf, bukan penolakan umum terhadap pendekatan GNN. Ada dua implikasi dari hasil ini. Pertama, *baseline* sederhana perlu diperlakukan serius dalam studi deteksi bot. Ketika perbandingan dilakukan pada protokol yang identik, model yang lebih kompleks harus membuktikan manfaat yang jelas pada metrik inti, bukan sekadar membawa narasi arsitektur yang lebih canggih. Kedua, hasil ini mendukung pembacaan metodologis yang lebih konservatif: struktur graf yang parsial atau kurang informatif dapat mengurangi manfaat *message passing*, sehingga investasi pada model relasional belum tentu efisien bila kualitas relasi graf tidak memadai. Dengan kata lain, nilai tambah model graf tidak berdiri sendiri; ia bergantung pada mutu sumber sinyal relasional yang tersedia.

3.2 Evaluasi Multi-Seed dan Analisis Ambang

Tahap 10 memberikan konteks penting untuk membaca hasil single-seed. Pada MLP 15 fitur, ambang asli 0,50 menghasilkan F1(bot) 0,53, sedangkan ambang terbaik 0,52 meningkatkan F1(bot) menjadi 0,54. Delta sebesar +0,0021 menunjukkan bahwa penyesuaian titik operasi memang dapat memberi perbaikan, tetapi skalanya kecil. Pada GraphSAGE, ambang terbaik tetap berada pada 0,55 dan tidak mengubah F1(bot) secara berarti. Hal ini menandakan bahwa kualitas dasar model tidak berubah drastis hanya karena *threshold* disetel ulang. Bacaan ini sejalan dengan literatur evaluasi pada data tidak seimbang. F1(bot) merepresentasikan kualitas keputusan pada satu titik operasi, sedangkan PR-AUC merepresentasikan kualitas *ranking* probabilitas di seluruh rentang *threshold* yang mungkin (de la Cruz Huayanay et al., 2024; Williams, 2021). Dalam konteks ini, analisis ambang berguna untuk memahami bagaimana model bergerak di sepanjang kompromi presisi-recall, tetapi tidak cukup untuk mengubah kesimpulan substantif tentang kapasitas diskriminatif model. Karena PR-AUC bersifat lebih stabil terhadap pemilihan ambang, metrik tersebut tetap diperlukan sebagai pembacaan komplementer ketika dua konfigurasi atau dua arsitektur dibandingkan. Temuan Tahap 10 juga memperkuat disiplin evaluasi artikel ini. Ambang dipilih dari data validasi, bukan dari data pengujian, sehingga tidak terjadi penyesuaian titik operasi yang "mengintip" hasil akhir. Disiplin ini penting karena pada *benchmark* berskala besar, perbedaan kecil yang tampak menguntungkan sering kali dapat muncul hanya karena keputusan tuning yang terlalu dekat dengan data pengujian. Artikel ini sengaja menempatkan analisis ambang sebagai lapisan penjelas,

TIN: Terapan Informatika Nusantara
Vol 6, No 12, May 2026, page 2384-2392
ISSN 2722-7987 (Media Online)
Website https://ejurnal.seminar-id.com/index.php/tin
DOI 10.47065/tin.v6i12.9891

sementara sumber klaim utama tetap berada pada evaluasi konfirmatori *multi-seed* Tahap 12. Hasil konfirmatori utama menunjukkan bahwa all_15 menghasilkan F1(bot) 0,5328 ± 0,0026 dan PR-AUC 0,4740 ± 0,0040, sedangkan user_only_8 menghasilkan F1(bot) 0,5327 ± 0,0011 dan PR-AUC 0,4943 ± 0,0030. Selisih rerata F1(bot) antara kedua konfigurasi hanya 0,0001. Dalam praktik evaluasi *benchmark*, selisih sekecil ini tidak cukup untuk mendukung klaim bahwa salah satu konfigurasi unggul secara berarti pada kualitas keputusan biner. Sebaliknya, selisih PR-AUC sebesar 0,0203 menunjukkan bahwa user_only_8 memberi kualitas *ranking* probabilitas yang lebih baik. Temuan ini diperkuat oleh simpangan baku yang lebih kecil pada user_only_8, baik untuk F1(bot) maupun PR-AUC. Tabel 2 menyajikan perbandingan lengkap metrik agregat untuk kedua konfigurasi.



Gambar 2. Kurva precision-recall Tahap 09 pada konfigurasi all_15

Selain nilai ringkas pada Tabel 1, Gambar 2 menunjukkan bentuk kurva precision-recall untuk MLP dan GraphSAGE pada pengujian single-seed Tahap 09. Visual ini dipakai sebagai pembacaan kualitas ranking dan kompromi presisi-recall, bukan sebagai pengganti evaluasi konfirmatori multi-seed.

Tabel 2. Metrik pengujian agregat *multi-seed* Tahap 12

Konfigurasi	F1(bot)	PR-AUC	Presisi(bot)	Recall(bot)	Akurasi	Ambang Terbaik
all_15	0,5328 ± 0,0026	0,4740 ± 0,0040	0,4011 ± 0,0048	0,7934 ± 0,0073	0,5902 ± 0,0080	0,5333 ± 0,0094
user_only_8	0,5327 ± 0,0011	0,4943 ± 0,0030	0,4005 ± 0,0026	0,7950 ± 0,0055	0,5893 ± 0,0045	0,5400 ± 0,0141

Tabel 2 menunjukkan bahwa F1(bot) kedua konfigurasi praktis setara, sehingga perbedaannya tidak cukup untuk klaim unggul pada keputusan biner. Pada saat yang sama, user_only_8 mempertahankan PR-AUC yang lebih tinggi daripada all_15. Simpangan bakunya juga lebih kecil, jadi konfigurasi profil-saja ini tampak lebih stabil lintas seed. Karena itu, keuntungan utamanya ada pada kualitas *ranking* dan kestabilan, bukan pada lompatan performa absolut.
Di sinilah framing kontribusi artikel harus dijaga tetap jujur. Artikel ini tidak menyatakan bahwa user_only_8 mengalahkan all_15 dalam segala dimensi, dan tidak pula menyatakan bahwa MLP profil-saja adalah solusi terbaik secara universal untuk semua skenario deteksi bot. Yang dapat dinyatakan secara sah ialah bahwa pada TwiBot-22, konfigurasi profil-saja mampu mempertahankan F1(bot) yang praktis identik, memberikan PR-AUC yang lebih tinggi, dan menampilkan stabilitas lintas seed yang lebih baik. Oleh sebab itu, keunggulan user_only_8 berada pada ekonomi fitur dan kestabilan, bukan pada lompatan performa absolut yang dramatis. Pembacaan ini juga penting dari sisi biaya eksperimen. Konfigurasi all_15 memerlukan Tahap 08, yaitu pemrosesan korpus tweet besar secara *streaming*, sedangkan user_only_8 bertumpu pada fitur profil yang lebih murah diekstraksi. Jika keluaran *multi-seed* menunjukkan F1(bot) yang tetap setara sementara kualitas *ranking* pada user_only_8 justru lebih baik, maka penambahan fitur tweet tidak dapat langsung dianggap rasional hanya karena menambah kompleksitas representasi. Temuan ini sejalan dengan argumen seleksi fitur bahwa utilitas fitur harus dibaca sebagai keseimbangan antara nilai sinyal, stabilitas, dan biaya komputasi, bukan sekadar jumlah fitur yang tersedia (Matharaarachchi et al., 2021).

3.3 Ablasi Fitur dan Efisiensi Konfigurasi user_only_8

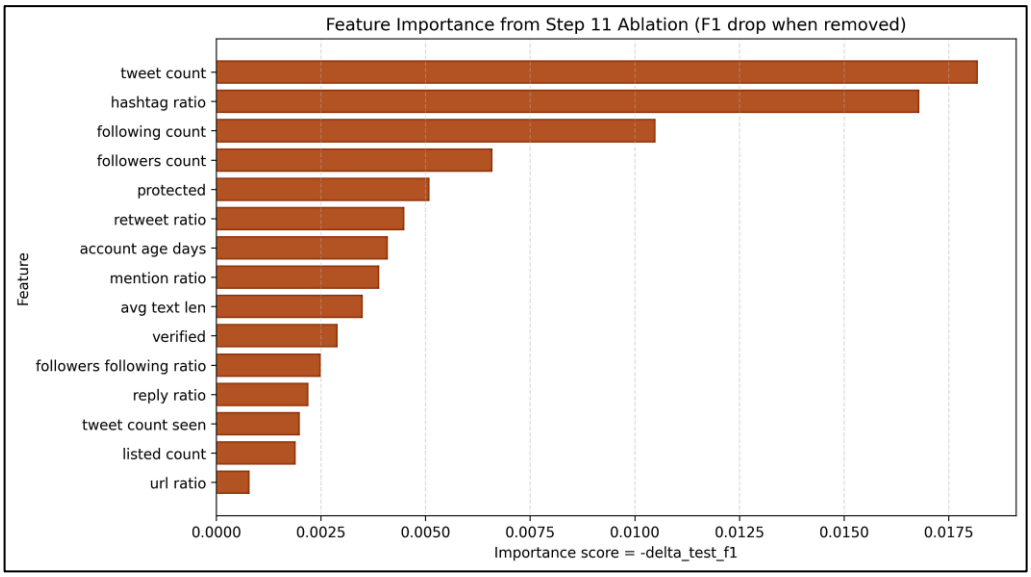
Pembahasan ablasi pada bagian ini dimulai dari pembacaan fitur individual, lalu bergerak ke pembandingan kelompok fitur. Urutannya penting karena hasil per fitur memberi konteks sebelum kualitas konfigurasi user_only_8 dan tweet_only dirangkum secara lebih luas pada Tabel 4. Daftar lima fitur dengan dampak penurunan performa terbesar saat dilakukan proses ablasi disajikan pada Tabel 3.

TIN: Terapan Informatika Nusantara
Vol 6, No 12, May 2026, page 2384-2392
ISSN 2722-7987 (Media Online)
Website https://ejurnal.seminar-id.com/index.php/tin
DOI 10.47065/tin.v6i12.9891

Tabel 3. Lima penurunan F1 terbesar pada ablasi *leave-one-out* Tahap 11

Peringkat	Eksperimen	F1(bot)	Delta F1	PR-AUC	Delta PR-AUC
1	Hapus tweet_count	0,5181	-0,0182	0,4424	-0,0358
2	Hapus hashtag_ratio	0,5195	-0,0168	0,4583	-0,0199
3	Hapus following_count	0,5258	-0,0105	0,4798	+0,0016
4	Hapus followers_count	0,5297	-0,0066	0,4776	-0,0006
5	Hapus protected	0,5312	-0,0051	0,4756	-0,0026

Tabel 3 menunjukkan bahwa penghapusan tweet_count dan hashtag_ratio memberikan penurunan F1 terbesar. Dengan demikian, Tabel 3 dipakai sebagai diagnosis fitur paling berdampak, sebelum perbandingan antar-kelompok fitur dirangkum pada Tabel 4. Untuk memperjelas distribusi kontribusi dari seluruh fitur yang dievaluasi, peringkat penurunan metrik F1 diilustrasikan secara visual pada Gambar 4.



Gambar 4. Peringkat kontribusi fitur berdasarkan penurunan F1 pada ablasi leave-one-out Tahap 11

Gambar 4 merangkum hasil ablasi leave-one-out sebagai peringkat penurunan F1 ketika setiap fitur dihapus. Visual ini membantu membaca Tabel 3 secara cepat, terutama untuk menempatkan tweet_count dan hashtag_ratio sebagai fitur yang paling sensitif dalam protokol ablasi ini tanpa memperluas klaim di luar benchmark TwiBot-22. Selanjutnya, perbandingan performa secara agregat antara kelompok fitur *all_15*, *user_only*, dan *tweet_only* dirangkum pada Tabel 4.

Tabel 4. Perbandingan kelompok fitur pada ablasi Tahap 11

Kelompok fitur	Jml fitur	F1(bot)	Delta F1	PR-AUC	Delta PR-AUC
all_15	15	0,5363	*baseline*	0,4782	*baseline*
user_only	8	0,5345	-0,0018	0,4962	+0,0180
tweet_only	7	0,4670	-0,0693	0,4473	-0,0309

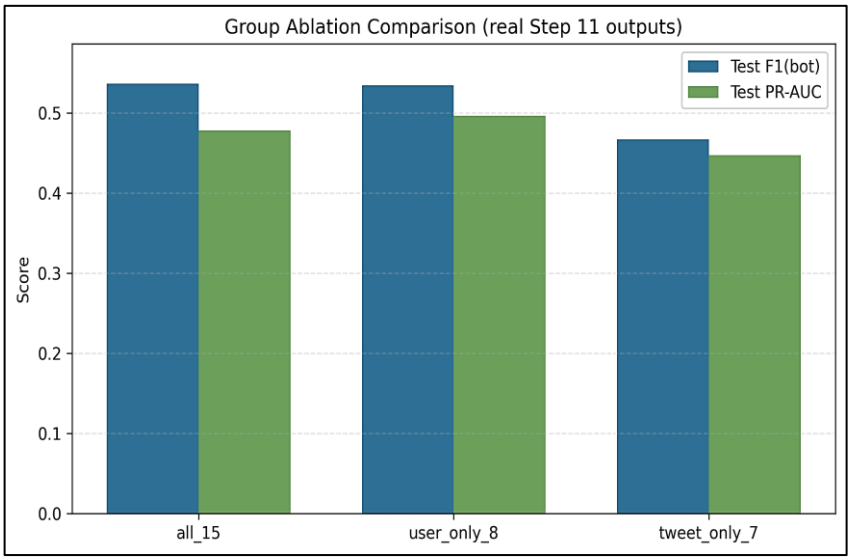
Ablasi Tahap 11 membantu menjelaskan mengapa all_15 tidak otomatis lebih baik pada evaluasi final. Pada skenario *leave-one-out*, penghapusan tweet_count menurunkan F1(bot) dari 0,54 ke 0,52, sedangkan penghapusan hashtag_ratio menurunkan F1(bot) ke 0,52. Dua penurunan terbesar ini menunjukkan bahwa ada fitur tambahan berbasis tweet yang memang membawa sinyal berguna.

Secara substantif, tweet_count merekam intensitas aktivitas akun, sementara hashtag_ratio memberi petunjuk tentang gaya penyebaran konten. Dari sudut pandang perilaku akun, masuk akal bila keduanya membantu membedakan aktivitas bot dari aktivitas manusia biasa. Karena itu, hasil ablasi individual layak dibaca sebagai diagnosis fitur, bukan sebagai daftar fitur terbaik yang berdiri sendiri. Namun, analisis individual tidak boleh dipisahkan dari analisis kelompok. Pada perbandingan grup, all_15 memperoleh F1(bot) 0,5363 dan PR-AUC 0,4782. Konfigurasi user_only masih sangat dekat pada F1(bot) dengan selisih -0,0018, tetapi PR-AUC-nya naik menjadi 0,4962. Sebaliknya, tweet_only turun jauh ke F1(bot) 0,4670 dan PR-AUC 0,4473. Pola ini memperlihatkan bahwa fitur tweet memang membantu saat melekat pada profil, tetapi blok fitur tweet secara keseluruhan belum cukup kuat untuk berdiri sendiri. Dengan kata lain, fitur perilaku tambahan lebih tepat dibaca sebagai penguat marjinal daripada fondasi utama pemisahan kelas.

Temuan ini menyatukan dua lapisan bukti yang tampak bertentangan di permukaan. Beberapa fitur tweet memang informatif secara individual, tetapi konfigurasi profil-saja tetap lebih ekonomis dan kompetitif secara

TIN: Terapan Informatika Nusantara
Vol 6, No 12, May 2026, page 2384-2392
ISSN 2722-7987 (Media Online)
Website https://ejurnal.seminar-id.com/index.php/tin
DOI 10.47065/tin.v6i12.9891

keseluruhan. Pada *benchmark* ini, nilai fitur tambahan belum cukup konsisten untuk menggantikan keunggulan sederhana dari user_only_8. Perbedaan performa metrik F1 dan PR-AUC antar-kelompok fitur tersebut dapat dilihat lebih jelas pada visualisasi Gambar 5.



Gambar 5. Perbandingan kelompok fitur pada skenario ablasi

Gambar 5 memperjelas bahwa konfigurasi user_only tetap sangat dekat dengan all_15 pada F1(bot), sementara tweet_only turun lebih jauh. Visual ini mendukung interpretasi bahwa fitur tweet berguna sebagai penguat tambahan, tetapi belum cukup kuat menjadi fondasi klasifikasi tanpa fitur profil.

3.4 Pembahasan terhadap Literatur dan Batas Klaim

Kaitan dengan literatur deteksi bot dan model graf perlu dibaca bersama kondisi data yang melingkupinya. TwiBot-20 dan TwiBot-22 memang dirancang untuk mendorong evaluasi yang lebih standar terhadap model deteksi bot berbasis graf, tetapi hasil pada *benchmark* ini tetap harus dipahami sebagai bukti yang spesifik pada data dan protokol yang dipakai.

BotRGCN dan model heterogeneity-aware menunjukkan bahwa struktur relasional dapat menambah kekuatan deteksi pada skenario tertentu. Model yang lebih baru seperti BotGSL, pendekatan semi-supervised relasional, HHG-Bot, dan NPGNN juga berupaya meningkatkan representasi relasional melalui pembelajaran struktur, heterogenitas, atau ketidakseimbangan pada level lingkungan node. Artikel ini tidak membantah relevansi arus penelitian tersebut.

Sebaliknya, artikel ini menyediakan pembanding yang lebih konservatif: jika *baseline* ringan sudah sangat kuat pada *benchmark* yang sama, maka model lanjutan harus dinilai dari nilai tambah empirisnya, bukan hanya dari kompleksitas desainnya. Posisi ini juga selaras dengan penelitian berbasis profil atau fitur perilaku yang menunjukkan bahwa data profil pengguna dapat menjadi basis yang kuat bagi model *deep learning*.

Dari sudut pandang ini, hasil user_only_8 dapat dibaca sebagai bukti bahwa fitur profil tetap merupakan sumber sinyal yang kuat dan relatif stabil pada TwiBot-22, bahkan ketika dibandingkan dengan konfigurasi yang telah diperkaya perilaku tweet. Kekuatan utama artikel ini bukan menemukan fenomena yang sepenuhnya baru, melainkan memperjelas batas praktis dari kompleksitas tambahan pada satu *benchmark* yang terdokumentasi baik.

Ada beberapa batas klaim yang harus ditegaskan agar pembahasan tidak melampaui evidensi. Pertama, seluruh hasil berlaku pada *benchmark* TwiBot-22 dan tidak secara otomatis mewakili performa pada data *real-time*, data lintas bahasa tertentu, atau populasi akun Twitter/X setelah perubahan kebijakan platform terbaru. Kedua, penelitian ini hanya membandingkan MLP dan GraphSAGE, sehingga tidak dapat dipakai untuk menyimpulkan bahwa semua model graf akan tampil lebih rendah dalam kondisi apa pun. Ketiga, artikel ini tidak menyajikan metrik per kelas manusia secara rinci karena fokus evaluasi diletakkan pada kelas bot sebagai kelas minoritas dan pada kualitas *ranking* probabilitas yang tersedia secara otoritatif. Keempat, analisis fitur Tahap 11 dibaca sebagai diagnostik, bukan konfirmasi statistik final. Dengan mempertahankan batas-batas ini, kontribusi utama artikel adalah memberikan pembacaan yang jujur dan berbasis *benchmark* bahwa pada kondisi data, protokol, dan artefak yang dianalisis, strategi hemat fitur lebih relevan daripada penambahan kompleksitas model dan fitur yang belum terbayar secara proporsional.

4. KESIMPULAN

Artikel ini menyimpulkan bahwa evaluasi deteksi bot pada *benchmark* TwiBot-22 perlu dibaca dengan disiplin metodologis dan disiplin klaim yang ketat. Pada konfigurasi 15 fitur, MLP menunjukkan hasil *single-seed* yang lebih baik daripada GraphSAGE, sehingga struktur graf yang tersedia pada jalur eksperimen ini belum memberikan

TIN: Terapan Informatika Nusantara
Vol 6, No 12, May 2026, page 2384-2392
ISSN 2722-7987 (Media Online)
Website https://ejurnal.seminar-id.com/index.php/tin
DOI 10.47065/tin.v6i12.9891

keuntungan empiris yang cukup untuk melampaui *baseline* tabular. Pada evaluasi konfirmatori *multi-seed*, konfigurasi user_only_8 dan all_15 menghasilkan F1(bot) yang praktis identik, tetapi user_only_8 memberikan PR-AUC yang lebih tinggi dan variasi hasil yang lebih kecil. Nilai utama konfigurasi profil-saja bukan terletak pada klaim superioritas absolut, melainkan pada efisiensi fitur: ia mempertahankan kualitas keputusan biner, meningkatkan kualitas *ranking* probabilitas, dan lebih stabil ketika eksperimen diulang. Analisis ablasi juga memperlihatkan bahwa beberapa fitur tweet, terutama tweet_count dan hashtag_ratio, tetap informatif, tetapi manfaatnya belum cukup konsisten untuk menjadikan konfigurasi kaya fitur sebagai pilihan default pada TwiBot-22. Namun demikian, perlu ditegaskan secara eksplisit bahwa temuan ini memiliki tiga keterbatasan penting. Pertama, seluruh hasil terbatas pada satu *benchmark*, yaitu TwiBot-22, sehingga tidak dapat digeneralisasi ke dataset lain, kondisi *real-time*, atau populasi bot yang berbeda karakteristiknya. Kedua, besaran F1(bot) yang dicapai, yaitu sekitar 0,53 pada kedua konfigurasi, merupakan angka yang moderat dan mencerminkan tantangan nyata deteksi bot pada data yang sangat tidak seimbang, bukan indikator performa tinggi yang dapat langsung diterapkan pada sistem produksi. Ketiga, perbandingan dalam artikel ini hanya mencakup MLP dan GraphSAGE, sehingga kesimpulan tentang keunggulan relatif kedua model tidak dapat diperluas ke arsitektur lain di luar kondisi yang diuji. Dengan mempertahankan batas-batas ini, kontribusi utama artikel adalah memberikan pembacaan yang jujur dan berbasis *benchmark* bahwa pada kondisi data, protokol, dan artefak yang dianalisis, strategi hemat fitur lebih relevan daripada penambahan kompleksitas model dan fitur yang belum terbayar secara proporsional. Kesimpulan ini diharapkan dapat menjadi dasar yang kuat bagi penelitian lanjutan yang ingin menguji model graf atau subset fitur baru dengan protokol evaluasi yang sama-sama ketat, dapat direproduksi, dan terbatas pada klaim yang benar-benar didukung oleh evidensi.

REFERENCES

Blakey, E. (2024). The day data transparency died: How Twitter/X cut off access for social research. *Contexts*, *23*(2), 30–35. https://doi.org/10.1177/15365042241252125

Chen, W., Pacheco, D., Yang, K.-C., & Menczer, F. (2021). Neutral bots probe political bias on social media. *Nature Communications*, *12*(1). https://doi.org/10.1038/s41467-021-25738-6

Cinelli, M., Cresci, S., Quattrociocchi, W., Tesconi, M., & Zola, P. (2022). Coordinated inauthentic behavior and information spreading on Twitter. *Decision Support Systems*, *160*, 113819. https://doi.org/10.1016/j.dss.2022.113819

De la Cruz Huayanay, A., Bazán, J. L., & Russo, C. M. (2024). Performance of evaluation metrics for classification in imbalanced data. *Computational Statistics*, *40*, 1447–1473. https://doi.org/10.1007/s00180-024-01539-5

Feng, S., Tan, Z., Li, R., & Luo, M. (2022). Heterogeneity-aware Twitter bot detection with relational graph transformers. *Proceedings of the AAAI Conference on Artificial Intelligence*, *36*(4), 3977–3985. https://doi.org/10.1609/aaai.v36i4.20314

Feng, S., Tan, Z., Wan, H., Wang, N., Chen, Z., Zhang, B., Zheng, Q., Zhang, W., Lei, Z., Yang, S., Feng, X., Zhang, Q., Wang, H., Liu, Y., Bai, Y., Wang, H., Cai, Z., Wang, Y., Zheng, L., … Luo, M. (2022). TwiBot-22: Towards graph-based Twitter bot detection. *Advances in Neural Information Processing Systems*, *35*, 35254–35269. https://arxiv.org/abs/2206.04564

Feng, S., Wan, H., Wang, N., Li, J., & Luo, M. (2021). TwiBot-20: A Comprehensive Twitter Bot Detection Benchmark. *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, 4485–4494. https://doi.org/10.1145/3459637.3482019

Feng, S., Wan, H., Wang, N., & Luo, M. (2021). BotRGCN: Twitter bot detection with relational graph convolutional networks. *Proceedings of the 2021 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, 236–239. https://doi.org/10.1145/3487351.3488336

Ferrara, E. (2023). Social bot detection in the age of ChatGPT: Challenges and opportunities. *First Monday*, *28*(6). https://doi.org/10.5210/fm.v28i6.13185

Fey, M., & Lenssen, J. E. (2019). *Fast Graph Representation Learning with PyTorch Geometric* (arXiv:1903.02428). arXiv. https://doi.org/10.48550/arXiv.1903.02428

Hayawi, K., Mathew, S., Venugopal, N., Masud, M. M., & Ho, P.-H. (2022). DeeProBot: A hybrid deep neural network model for social bot detection based on user profile data. *Social Network Analysis and Mining*, *12*(1). https://doi.org/10.1007/s13278-022-00869-w

Huang, D., Song, J., & Zhang, X. (2025). Semi-Supervised Social Bot Detection with Relational Graph Attention Transformers and Characteristics of the social environment. *Information Fusion*, *118*, 102956. https://doi.org/10.1016/j.inffus.2025.102956

Küpfer, A. (2024). Nonrandom tweet mortality and data access restrictions: Compromising the replication of sensitive Twitter studies. *Political Analysis*, *32*(4), 493–506. https://doi.org/10.1017/pan.2024.7

Li, Y., Lu, H., & Chen, W. (2026). Neighborhood perceivable graph neural network for relational heterogeneous Twitter bot detection. *PLOS ONE*, *21*(2), e0342686. https://doi.org/10.1371/journal.pone.0342686

Matharaarachchi, S., Domaratzki, M., & Muthukumarana, S. (2021). Assessing feature selection method performance with class imbalance data. *Machine Learning with Applications*, *6*, 100170. https://doi.org/10.1016/j.mlwa.2021.100170

TIN: Terapan Informatika Nusantara
Vol 6, No 12, May 2026, page 2384-2392
ISSN 2722-7987 (Media Online)
Website https://ejurnal.seminar-id.com/index.php/tin
DOI 10.47065/tin.v6i12.9891

Mazza, M., Avvenuti, M., Cresci, S., & Tesconi, M. (2022). Investigating the difference between trolls, social bots, and humans on Twitter. *Computer Communications*, *196*, 23–36. https://doi.org/10.1016/j.comcom.2022.09.022

Najari, S., Salehi, M., & Farahbakhsh, R. (2022). GANBOT: A GAN-based framework for social bot detection. *Social Network Analysis and Mining*, *12*(1). https://doi.org/10.1007/s13278-021-00800-9

Wang, T., Wang, Z., Li, H., Xia, C., & Zhao, C. (2025). HHG-Bot: A hyperheterogeneous graph-based Twitter bot detection model. *IEEE Transactions on Computational Social Systems*, *12*(5), 3416–3430. https://doi.org/10.1109/TCSS.2025.3543419

Wei, C., Liang, G., & Yan, K. (2024). BotGSL: Twitter bot detection with graph structure learning. *The Computer Journal*, *67*(7), 2486–2497. https://doi.org/10.1093/comjnl/bxae020

Williams, C. K. I. (2021). The effect of class imbalance on precision-recall curves. *Neural Computation*, *33*(4), 853–857. https://doi.org/10.1162/neco_a_01362