Semi-Supervised Phishing Email Detection Using Tri-Training with Transformer Models

Natzwa Novena Rantung, Ahya Radiatul Kamila*
Jurusan Sains data, Fakultas Teknologi dan Desain, Universitas Bunda Mulia, Jakarta, Indonesia

Received: 30 September 2026 Accepted: 10 November 2026 Published: 1 December 2026

*akamila@bundamulia.ac.id

Kata Kunci: Deteksi *phishing*, *Tri-Training*, *semi-supervised learning*, *transformer*, pergeseran domain

DSI: Jurnal Data Science Indonesia is licensed under a Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0).

Abstrak : Email *phishing* terus berubah, sehingga detektor yang dilatih pada korpus lama mudah kehilangan akurasi ketika menghadapi email modern, sementara pelabelan email baru membutuhkan waktu dan keahlian. Penelitian ini menguji *Tri-Training* versi Zhou dan Li dengan BERT, DistilBERT, dan RoBERTa untuk klasifikasi biner email *phishing* dan *legitimate* dengan data berlabel terbatas. Data berlabel berasal dari *Phishing Email Curated Datasets* di Zenodo (7.665 email), sedangkan data tanpa label dikumpulkan dari ScamWarners, Phishing Pot, dan arsip *mailing list* IETF. Data dibagi per kelompok email hampir sama untuk mencegah kebocoran, dan sebagian email modern disisihkan sebagai uji eksternal. Eksperimen dijalankan pada porsi label 1%, 2%, 5%, dan 10% dengan tiga *seed*, disertai ablasi *Weighted Cross-Entropy*, kontrol *pool in-domain*, pembanding TF-IDF, uji McNemar, *bootstrap*, dan koreksi Holm. Pada uji internal, *Tri-Training* tidak memberi perbedaan yang signifikan karena *baseline* sudah mencapai F1-macro 94,80% sampai 98,40%. Pada uji eksternal, F1-macro komite naik 5,51 sampai 10,78 poin pada keempat porsi, dan 10 dari 12 perbandingan tetap signifikan setelah koreksi Holm. DistilBERT 1,89 kali lebih cepat daripada BERT dengan F1-macro internal yang setara, sehingga layak menjadi penyaring pertama.

PENDAHULUAN

Email *phishing* kini jauh lebih rapi daripada pesan penipuan satu dekade lalu. Anti-Phishing Working Group mencatat 989.123 serangan *phishing* pada kuartal keempat 2024[1]. Pada tahun yang sama, Internet Crime Complaint Center milik FBI menerima 859.532 pengaduan dengan kerugian 16,6 miliar dolar AS, dan *phishing* menjadi jenis pengaduan yang paling sering[2]. Laporan Verizon menunjukkan bahwa unsur manusia terlibat pada 68% pelanggaran data[3]. Penyaringan berbasis daftar hitam dan aturan statis tidak lagi memadai, sehingga dibutuhkan detektor yang membaca isi pesan[4].

Deteksi berbasis teks berkembang dari TF-IDF dengan SVM atau *Logistic Regression*[5], metode *ensemble* untuk fitur tautan[6][7][8], sampai model *transformer* yang menangkap konteks dua arah[9][10]. DistilBERT, BERT, dan RoBERTa telah mencapai F1 atau akurasi di atas 0,93 pada email *phishing*[11][12][13], DistilBERT juga dipakai untuk deteksi waktu nyata[14], dan pendekatan serupa telah diterapkan pada *spam* serta email *phishing* berbahasa Indonesia[15][16]. Namun, hampir semua hasil tersebut diperoleh dengan data berlabel penuh dan diuji pada distribusi yang sama dengan data latih. Padahal email baru muncul lebih cepat daripada kemampuan analis untuk melabelinya[17], dan email modern sengaja dirancang untuk menghindari detektor[18].

Pembelajaran semi-terawasi memanfaatkan data tanpa label untuk mengurangi kebutuhan label[19][20][21], dan di Indonesia telah dipakai melalui *self-learning* dan *co-training* untuk teks [22][23][24]. *Tri-Training* memakai tiga pengklasifikasi yang saling memberi label dengan syarat galat dan batas jumlah *pseudo-label*[25]. Sejauh penelusuran penulis, belum ada penelitian yang menerapkannya dengan tiga

transformer berbeda arsitektur untuk email *phishing* sekaligus menguji hasilnya pada email modern dari sumber lain. Penelitian ini menjawab tiga pertanyaan, yaitu seberapa besar *Tri-Training* meningkatkan kinerja ketika label hanya 1% sampai 10%, komponen mana yang paling berperan, dan bagaimana keseimbangan kinerja serta efisiensi ketiga model.

TINJAUAN LITERATUR

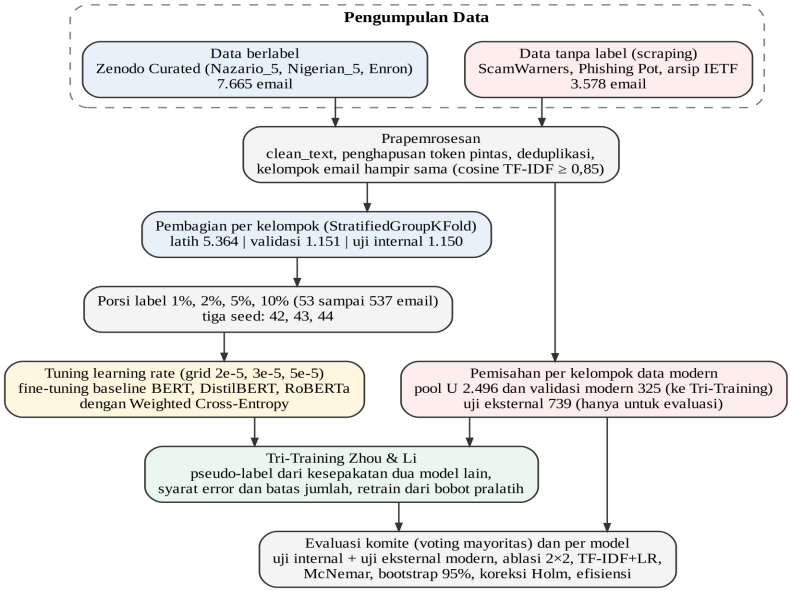
Kamila dkk. menunjukkan bahwa penambahan fitur meningkatkan kinerja algoritma *bagging* dan *boosting* pada deteksi tautan *phishing*[6], sedangkan Tampinongkol dkk. menyimpulkan bahwa pilihan antara *Random Forest* dan SVM perlu disesuaikan dengan kebutuhan dan sumber daya[7]. Penelitian lain dari kelompok yang sama memakai *Random Forest* dengan enam fitur URL dari data Kaggle, PhishTank, dan tautan resmi yang dikumpulkan sendiri, lalu memperoleh akurasi 0,775 [8]. Ketiga penelitian tersebut menegaskan bahwa kualitas fitur sama pentingnya dengan algoritma. Pada isi email, Otieno dkk. memperoleh F1 0,93 dengan BERT pada korpus Nazario dan Enron[12], Ibrahim dan Elhafiz memperoleh F1 0,94 untuk BERT dan 0,96 untuk RoBERTa[13], dan Meléndez dkk. melaporkan akurasi di atas 99% dengan RoBERTa[11]. Ketiganya memakai label penuh dan menguji pada distribusi yang sama.

Champa dkk. menunjukkan bahwa repositori email publik perlu dikurasi sebelum dipakai, lalu merilis kumpulan data terkurasi yang dipakai di sini[26], [27]. Boulila dkk. menganalisis 1.551 email yang lolos dari seluruh lapisan keamanan lima perusahaan dan menemukan bahwa serangan modern bervolume rendah dan memakai teknik penghindaran seperti kode QR[18]. *Tri-Training* memberi *pseudo-label* pada satu model jika dua model lainnya sepakat, dan pembaruan hanya dilakukan jika galat gabungan menurun serta jumlah *pseudo-label* memenuhi batas yang diturunkan dari analisis derau klasifikasi[25]. Keragaman anggota menjadi kunci, sehingga penelitian ini memakai BERT, DistilBERT, dan RoBERTa yang berbeda dalam data pralatih, *tokenizer*, dan ukuran.

METODE PENELITIAN

Rancangan Sistem dan Data

Alur penelitian ditunjukkan pada Gambar 1. Data berlabel diambil dari *Phishing Email Curated Datasets* di Zenodo dengan lisensi CC BY 4.0 [28], yaitu file Nazario_5 dan Nigerian_5 ditambah email *ham* Enron [29]. Setelah duplikat dibuang, *phishing* Nigerian dibatasi sama dengan Nazario (1.533 email), dan email *legitimate* dibatasi 1,5 kali email *phishing*, diperoleh 7.665 email (4.599 *legitimate* dan 3.066 *phishing*) seperti pada Tabel 1. Email yang hampir sama, yaitu yang *cosine similarity* TF-IDF-nya paling sedikit 0,85, dikelompokkan menjadi 6.931 kelompok, lalu dibagi 70/15/15 dengan StratifiedGroupKFold menjadi 5.364 email latih, 1.151 email validasi, dan 1.150 email uji internal.



Gambar 1. Diagram alur penelitian deteksi email *phishing* dengan *Tri-Training*

Tabel 1. Komposisi Data Berlabel per Asal File dan per Split

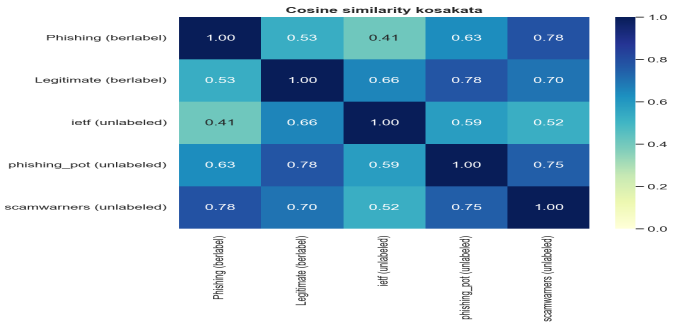
Kelompok	Legitimate	Phishing	Total
Enron (ham)	1.604	0	1.604
Nazario_5	3	1.533	1.536
Muncul di Nazario_5 dan Nigerian_5	1.494	0	1.494
Nigerian_5	1.498	1.533	3.031
Total	**4.599**	**3.066**	**7.665**
Data latih (70%)	3.218	2.146	5.364
Data validasi (15%)	691	460	1.151
Data uji internal (15%)	690	460	1.150

Data tanpa label dikumpulkan dari forum ScamWarners [30], repositori *honeypot* Phishing Pot [31], dan arsip *mailing list* IETF [32], seperti dirinci pada Tabel 2. Pengambilan mematuhi robots.txt sesuai RFC 9309 [33], sehingga tautan dengan parameter sid, versi cetak, dan daftar anggota forum tidak pernah dibuka. Setiap permintaan diberi jeda acak 3 sampai 6 detik dengan User-Agent yang menyatakan tujuan riset akademik. Dari ScamWarners hanya post pertama setiap *thread* yang diambil, dan pengambilan dihentikan tanpa diakali ketika situs menampilkan proteksi Cloudflare setelah 197 email terkumpul. Kekurangannya diisi dari arsip ZIP resmi Phishing Pot, sedangkan email IETF diambil dari 22 file mbox bulanan periode 2015 sampai 2022. Alamat email dan nomor telepon dianonimkan sesuai UU Pelindungan Data Pribadi [34]. Setelah penyaringan dan penyeimbangan diperoleh 3.578 email, dan setelah 18 duplikat lagi dibuang pada prapemrosesan tersisa 3.560 email. Sebanyak 1.064 email modern disisihkan per kelompok, yaitu 325 email sebagai validasi modern dan 739 email sebagai uji eksternal (347 Phishing Pot, 42 ScamWarners, dan 350 IETF) dengan label berdasarkan sumber. Sisa 2.496 email menjadi *pool scraped*. Sebagai kontrol dibentuk *pool in-domain* berukuran sama dari data latih Zenodo dengan label disembunyikan, sehingga efek prosedur *Tri-Training* dapat dipisahkan dari efek isi *pool*.

Tabel 2. Rincian Pengumpulan Data Tanpa Label per Sumber

Sumber	Jenis email	Cara pengambilan	Kepatuhan dan lisensi	Terkumpul	Setelah penyaringan
ScamWarners (www.scamwarners.com)	Email penipuan yang dilaporkan korban	Web scraping post pertama dari 175.917 URL thread pada 19 sitemap	robots.txt dipatuhi, jeda 3 sampai 6 detik, berhenti saat diblokir Cloudflare	197	184
Phishing Pot (github.com/rf-peixoto/phishing_pot)	Email phishing asli dari honeypot	Arsip ZIP resmi commit 80a4867012, 12.864 file .eml diacak	Lisensi CC BY-NC 4.0, atribusi dicantumkan	2.803	1.605
Arsip IETF (ietf@ietf.org)	Email profesional mailing list	Unduhan mbox bulanan 2015 sampai 2022, 22 dari 88 bulan, maks. 150 email per bulan	Arsip publik resmi, robots.txt dipatuhi, jeda 3 sampai 6 detik	3.000	1.789
Total				**6.000**	**3.578**

Gambar 2 memperlihatkan kemiripan kosakata TF-IDF antarkelompok. Email ScamWarners paling mirip dengan *phishing* berlabel (0,78), sedangkan email Phishing Pot justru lebih mirip dengan email *legitimate* berlabel (0,78) daripada dengan *phishing* berlabel (0,63). Pola ini sesuai dengan pengamatan bahwa *phishing* modern meniru notifikasi bisnis yang wajar [18], dan menjadi alasan utama uji eksternal jauh lebih sulit daripada uji internal.



Gambar 2. *Cosine similarity* kosakata antara data berlabel dan data tanpa label per sumber

Seluruh teks dibersihkan dengan fungsi yang sama. HTML dihapus, teks diubah ke huruf kecil, lalu URL, alamat email, nomor telepon, tahun, dan angka panjang diganti token [URL], [EMAIL], [PHONE], [YEAR], dan [NUM]. Sebanyak 21 token pintas yang hanya mencerminkan asal korpus, misalnya enron, ect, kaminski, jose, dan nazario, juga dihapus. Sebelum pembersihan, penanda Enron muncul pada 24,9% email *legitimate* dan hanya 0,1% email *phishing*, dan tahun 1995 sampai 2008 muncul pada 52,9% email *legitimate*. Setelah pembersihan, penanda Enron tersisa 1,5% dan pola tahun 0%. Kata yang paling membedakan kelas *phishing* kemudian adalah beneficiary, kin, deceased, dan fund, yaitu kosakata khas penipuan *advance-fee*, sehingga model belajar dari isi pesan dan bukan dari asal korpus. Teks dipotong 3.000 karakter sebelum tokenisasi dengan panjang maksimum 256 token. Median panjangnya 273 token, sehingga sekitar 52% email terpotong, tetapi subjek, sapaan, dan permintaan utama pelaku yang berada di awal email tetap terbaca.

Model dan *Fine-Tuning*

BERT adalah *encoder* dua arah 12 lapisan berbasis *self-attention* [10], [9]. RoBERTa memakai arsitektur yang sama dengan pralatih lebih lama pada korpus lebih besar [35], sedangkan DistilBERT adalah hasil distilasi BERT dengan enam lapisan, sekitar 40% lebih kecil dan 60% lebih cepat [36], [37]. Karena kelas *phishing* hanya 40%, fungsi rugi yang dipakai adalah *Weighted Cross-Entropy* (WCE) pada Persamaan 1 dengan bobot kelas pada Persamaan 2.

$$\mathcal{L}_{WCE} = -\frac{\sum_{i=1}^{B} w_{y_i} \log \hat{p}_{i,y_i}}{\sum_{i=1}^{B} w_{y_i}} \quad \text{.....(1)}$$

$$w_c = \frac{N}{C \cdot N_c} \quad \text{.....(2)}$$

B adalah ukuran *batch*, w_{yi} adalah bobot kelas label sampel ke-*i*, *N* adalah jumlah sampel latih, *C* = 2, dan N_c adalah jumlah sampel kelas *c*, sehingga bobotnya 0,83 untuk *legitimate* dan 1,25 untuk *phishing*. Konfigurasi pada Tabel 3 memakai AdamW [38] dan perpanjangan *epoch* untuk porsi yang sangat kecil karena *fine-tuning* pada data kecil dikenal tidak stabil [39]. *Learning rate* dipilih dari 2×10^{-5}, 3×10^{-5}, dan 5×10^{-5} [10] pada porsi 5% *seed* 42, dengan selisih antarnilai kurang dari satu poin. Porsi label 1% (53 email), 2% (107), 5% (268), dan 10% (537) diambil bertingkat dan terstratifikasi, sedangkan pelatihan dengan seluruh 5.364 email latih menjadi batas atas.

Tabel 3. Konfigurasi Fine-Tuning

Parameter	Nilai
Model pralatih	bert-base-uncased, distilbert-base-uncased, roberta-base
Learning rate (hasil tuning)	BERT 5×10^{-5}; DistilBERT dan RoBERTa 3×10^{-5}
Batch efektif dan panjang masukan	16 (2 micro-batch berukuran 8); 256 token
Epoch dan early stopping	Maks. 4, diperpanjang hingga minimal 100 langkah (maks. 15); patience 2 pada F1-macro validasi
Optimizer dan penjadwal	AdamW, weight decay 0,01; warmup linear 10%, lalu turun linear
Fungsi rugi dan presisi	Weighted Cross-Entropy (Persamaan 1 dan 2); mixed precision FP16
Seed dan perangkat	42, 43, 44; NVIDIA GeForce RTX 4050 Laptop GPU 6 GB, RAM 16 GB

***Tri-Training* Zhou dan Li**

Pada setiap iterasi, model h_i mendapat kandidat *pseudo-label* dari email *pool* yang diprediksi sama oleh dua model lainnya, tanpa ambang kepercayaan [25]. Keragaman diperoleh dari tiga arsitektur berbeda, sehingga

sampel *bootstrap* tidak dipakai. Galat gabungan dua model diestimasi pada validasi gabungan (1.151 email Zenodo dan 325 email modern) dengan penghalusan Laplace pada Persamaan 3, dan h_i hanya diperbarui jika syarat pada Persamaan 4 terpenuhi.

$$\hat{e}_t = \frac{m_t+1}{n_t+2} \quad \text{.....(3)}$$

$$\hat{e}_t < e'_{t-1} \quad \text{dan} \quad |L_t|\,\hat{e}_t < |L_{t-1}|\,e'_{t-1} \quad \text{.....(4)}$$

n_t adalah jumlah sampel validasi yang disepakati kedua model, m_t adalah yang disepakati tetapi salah, dan e'_{t-1} serta $|L_{t-1}|$ adalah galat dan jumlah *pseudo-label* pembaruan sebelumnya dengan $e'_0 = 0{,}5$. Jika kandidat terlalu banyak, diambil acak sebanyak s pada Persamaan 5 dengan proporsi kelas mengikuti data berlabel.

$$s = \left\lceil \frac{e'_{t-1}\,|L_{t-1}|}{\hat{e}_t} - 1 \right\rceil \quad \text{.....(5)}$$

Model yang diperbarui dilatih ulang dari bobot pralatih pada gabungan data berlabel dan *pseudo-label* dengan *seed* tetap, mengikuti temuan bahwa inisialisasi ulang mencegah akumulasi kesalahan [21]. Prediksi komite adalah voting mayoritas ketiga model. Iterasi terbaik dipilih dari F1-macro komite validasi, dan proses berhenti jika tidak ada model yang diperbarui, F1-macro tidak naik selama dua iterasi, atau sudah lima iterasi. Uji eksternal tidak pernah dipakai pada tahap ini.

Evaluasi

Evaluasi dilakukan pada uji internal dan uji eksternal untuk setiap model dan komite dengan *phishing* sebagai kelas positif. Metrik yang dipakai adalah *Accuracy*, *Precision*, *Recall*, F1 pada Persamaan 6, F1-macro pada Persamaan 7, FPR pada Persamaan 8, serta AUC.

$$F1 = \frac{2\times\text{Precision}\times\text{Recall}}{\text{Precision}+\text{Recall}} \quad \text{.....(6)}$$

$$F1_{macro} = \frac{1}{C}\sum_{c=1}^{C} F\,1_c \quad \text{.....(7)}$$

$$\text{FPR} = \frac{FP}{FP+TN} \quad \text{.....(8)}$$

Precision adalah TP/(TP + FP) dan *Recall* adalah TP/(TP + FN), dengan TP email *phishing* yang terdeteksi benar, FP email *legitimate* yang salah ditandai *phishing*, FN email *phishing* yang lolos, TN email *legitimate* yang dikenali benar, dan $F1_c$ adalah F1 kelas *c*. F1-macro dipakai sebagai metrik utama karena tidak bias terhadap kelas mayoritas, sedangkan FPR penting karena setiap email *legitimate* yang diblokir langsung mengganggu pengguna. Perbedaan *baseline* dan *Tri-Training* diuji dengan McNemar eksak [40] pada Persamaan 9, dengan *b* dan *c* adalah jumlah email yang berubah dari benar ke salah dan sebaliknya.

$$p = \min\left(1,\ 2\sum_{k=0}^{\min(b,c)}\binom{b+c}{k}\,0{,}5^{\,b+c}\right) \quad \text{.....(9)}$$

Selang kepercayaan 95% selisih F1-macro dihitung dengan 1.000 *bootstrap*, dan nilai *p* dari 60 perbandingan pada setiap jenis uji dikoreksi dengan metode Holm [41]. Selain itu dilakukan ablasi 2×2 WCE dan *Tri-Training* pada porsi 5%, pembanding TF-IDF dengan *Logistic Regression*, dan pengukuran latensi.

HASIL PENELITIAN

Uji Internal dan Uji Eksternal

Tabel 4 menyajikan kinerja setiap model dan komite pada porsi 5%. Pada uji internal, komite *baseline* dengan 268 email berlabel sudah mencapai F1-macro 97,67%, dan *Tri-Training* dengan *pool scraped* menghasilkan 97,64%, sehingga praktis tidak berubah. Dari 12 perbandingan komite pada uji internal tidak ada yang signifikan setelah koreksi Holm. Pola ini merupakan efek plafon, karena ruang perbaikan yang tersisa kurang dari dua poin. BERT justru turun 2,05 poin dengan simpangan baku besar, terutama pada *seed* 43 ketika FPR-nya naik menjadi 10,58%, tetapi efek ini diredam komite sehingga FPR komite hanya naik dari 1,26% menjadi 1,74%.

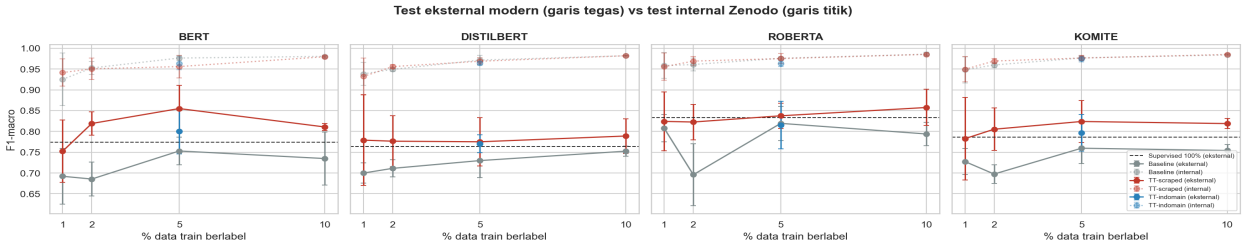
Tabel 4. Kinerja Uji Internal dan Uji Eksternal Modern pada Porsi 5% (%, Rata-rata ± Simpangan Baku Tiga Seed)

Skenario	Model	F1-macro internal	AUC internal	F1-macro eksternal	Recall phishing eksternal	FPR eksternal	AUC eksternal
Baseline	BERT	97,64 ± 0,60	99,43 ± 0,37	75,29 ± 3,30	55,10 ± 5,66	0,48 ± 0,16	93,57 ± 3,00
	DistilBERT	97,14 ± 1,17	99,36 ± 0,41	72,98 ± 4,07	51,50 ± 6,69	0,76 ± 0,16	94,07 ± 3,43
	RoBERTa	97,59 ± 0,29	99,68 ± 0,15	81,93 ± 4,11	66,75 ± 7,46	0,57 ± 0,29	97,02 ± 2,72
	Komite	97,67 ± 0,55	99,58 ± 0,26	75,95 ± 3,73	56,30 ± 6,14	0,57 ± 0,29	96,03 ± 2,52
TT-scraped	BERT	95,59 ± 2,71	99,43 ± 0,16	85,49 ± 5,62	74,04 ± 11,04	1,43 ± 1,03	95,95 ± 2,33
	DistilBERT	96,87 ± 0,84	99,44 ± 0,21	77,50 ± 5,82	59,21 ± 10,26	0,76 ± 0,16	95,19 ± 2,91
	RoBERTa	97,50 ± 1,21	99,70 ± 0,13	83,75 ± 3,06	70,01 ± 5,69	0,57 ± 0,29	97,18 ± 2,35
	Komite	97,64 ± 0,69	99,63 ± 0,14	82,38 ± 5,09	67,69 ± 9,38	0,67 ± 0,33	97,59 ± 1,84
TT-in-domain	BERT	96,37 ± 1,55	99,33 ± 0,39	80,06 ± 5,47	63,92 ± 10,03	1,05 ± 0,72	92,60 ± 3,74
	DistilBERT	96,46 ± 0,47	99,42 ± 0,09	77,06 ± 2,15	58,44 ± 3,85	0,95 ± 0,16	94,44 ± 2,33
	RoBERTa	96,28 ± 0,70	99,58 ± 0,15	81,52 ± 5,71	66,50 ± 10,65	1,05 ± 0,44	98,14 ± 1,16
	Komite	97,40 ± 0,73	99,55 ± 0,28	79,65 ± 4,40	62,81 ± 7,99	0,67 ± 0,33	95,66 ± 2,31
Supervised 100%	BERT	99,30 ± 0,10	99,84 ± 0,06	77,40 ± 2,61	58,53 ± 4,53	0,29 ± 0,00	98,35 ± 0,77
	DistilBERT	99,21 ± 0,05	99,77 ± 0,04	76,39 ± 3,03	56,81 ± 5,20	0,29 ± 0,00	97,87 ± 1,08
	RoBERTa	99,12 ± 0,10	99,91 ± 0,04	83,28 ± 1,85	68,89 ± 3,36	0,29 ± 0,00	99,48 ± 0,07
	Komite	99,37 ± 0,16	99,86 ± 0,04	78,65 ± 1,85	60,67 ± 3,21	0,29 ± 0,00	99,22 ± 0,17

Hasil berbeda tajam pada uji eksternal. Tabel 5 dan Gambar 3 memperlihatkan bahwa komite kehilangan 15 sampai 27 poin ketika berpindah ke email modern. *Tri-Training* menaikkan F1-macro komite eksternal sebesar 5,51, 10,78, 6,43, dan 6,47 poin pada porsi 1%, 2%, 5%, dan 10%. Pada porsi 5%, komite *Tri-Training* (82,38%) melampaui model *supervised* dengan seluruh data latih (78,65%). Kenaikan terbesar per model terjadi pada BERT (75,29% menjadi 85,49%) karena *recall phishing*-nya naik dari 55,10% menjadi 74,04%, sedangkan FPR semua model tetap di bawah 1,5%. Dari 12 perbandingan komite eksternal, 10 tetap signifikan setelah koreksi Holm, misalnya 103 email berubah dari salah menjadi benar dan hanya 1 sebaliknya pada porsi 2% *seed* 42. Satu perbandingan pada porsi 1% *seed* 43 turun 1,82 poin. Rincian uji McNemar per *seed* disajikan pada Tabel 6. Pada kontrol *in-domain*, dua dari tiga perbandingan signifikan.

Tabel 5. F1-macro Komite pada Uji Internal dan Uji Eksternal Modern (%)

Porsi	Internal baseline	Internal Tri-Training	Signifikan internal (Holm)	Eksternal baseline	Eksternal Tri-Training	Selisih eksternal (poin)	Signifikan eksternal (Holm)
1%	94,80 ± 3,11	94,94 ± 3,06	0 dari 3	72,74 ± 3,14	78,25 ± 9,90	+5,51	2 dari 3
2%	95,97 ± 0,65	96,92 ± 0,60	0 dari 3	69,74 ± 2,23	80,52 ± 5,08	+10,78	3 dari 3
5%	97,67 ± 0,55	97,64 ± 0,69	0 dari 3	75,95 ± 3,73	82,38 ± 5,09	+6,43	2 dari 3
5% (in-domain)	97,67 ± 0,55	97,40 ± 0,73	0 dari 3	75,95 ± 3,73	79,65 ± 4,40	+3,70	2 dari 3
10%	98,40 ± 0,14	98,49 ± 0,14	0 dari 3	75,41 ± 1,45	81,88 ± 1,23	+6,47	3 dari 3
100% (supervised)	99,37 ± 0,16	–	–	78,65 ± 1,85	–	–	–



Gambar 3. F1-macro uji eksternal modern (garis tegas) dan uji internal (garis titik) terhadap porsi data berlabel

Tabel 6. Uji McNemar Komite pada Uji Eksternal Modern (F1-macro dalam %)

Porsi	Seed	Baseline	Tri-Training	Selisih	CI 95% selisih	b (benar ke salah)	c (salah ke benar)	p Holm	Signifikan
1%	42	72,13	78,70	+6,57	[+4,74; +8,51]	2	45	< 0,001	Ya
1%	43	69,96	68,14	−1,82	[−3,59; −0,24]	20	9	0,799	Tidak
1%	44	76,14	87,92	+11,77	[+9,40; +14,32]	3	85	< 0,001	Ya
2%	42	67,92	83,45	+15,53	[+12,71; +18,07]	1	103	< 0,001	Ya
2%	43	69,08	74,65	+5,57	[+3,60; +7,43]	3	38	< 0,001	Ya
2%	44	72,22	83,45	+11,23	[+8,99; +13,67]	5	80	< 0,001	Ya
5%	42	75,23	77,33	+2,09	[+0,63; +3,50]	5	19	0,093	Tidak
5%	43	72,64	82,32	+9,68	[+7,40; +11,89]	0	65	< 0,001	Ya
5%	44	79,98	87,50	+7,52	[+5,60; +9,60]	4	57	< 0,001	Ya
5% in-domain	42	75,23	76,12	+0,89	[−0,50; +2,29]	9	15	1,000	Tidak
5% in-domain	43	72,64	78,26	+5,62	[+3,93; +7,26]	0	37	< 0,001	Ya
5% in-domain	44	79,98	84,58	+4,60	[+3,01; +6,45]	5	37	< 0,001	Ya
10%	42	73,98	81,03	+7,05	[+5,21; +8,91]	2	49	< 0,001	Ya
10%	43	75,36	81,31	+5,95	[+4,10; +7,84]	3	43	< 0,001	Ya
10%	44	76,88	83,29	+6,41	[+4,65; +8,30]	0	44	< 0,001	Ya

Tabel 7 merinci uji eksternal komite per sumber. Kesulitan utama ada pada email Phishing Pot, dengan *recall* komite *baseline* hanya 43,52% sampai 54,56% yang naik menjadi 59,46% sampai 66,38% setelah *Tri-Training*. *Recall* pada ScamWarners lebih tinggi karena gayanya dekat dengan penipuan pada data berlabel, sedangkan FPR pada email IETF tetap rendah, yaitu 0,29% sampai 1,33% atau setara dua sampai lima email dari 350 email.

Tabel 7. Kinerja Komite per Sumber Uji Eksternal Modern (%, Rata-rata ± Simpangan Baku Tiga Seed)

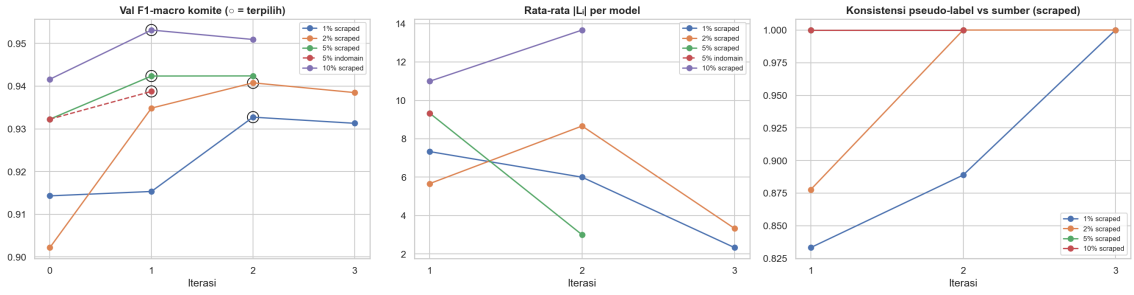
Porsi	Skenario	Recall Phishing Pot (n = 347)	Recall ScamWarners (n = 42)	FPR IETF (n = 350)
1%	Baseline	48,90 ± 6,09	71,43 ± 10,38	1,14 ± 0,99
1%	TT-scraped	59,46 ± 18,88	76,19 ± 0,00	1,33 ± 0,59
2%	Baseline	43,52 ± 3,60	65,87 ± 1,37	0,38 ± 0,33
2%	TT-scraped	62,63 ± 9,62	80,16 ± 9,01	0,86 ± 0,00
5%	Baseline	54,56 ± 5,44	70,63 ± 11,98	0,57 ± 0,29
5%	TT-scraped	66,38 ± 9,38	78,57 ± 10,38	0,67 ± 0,33
5%	TT-in-domain	60,90 ± 7,99	78,57 ± 8,25	0,67 ± 0,33
10%	Baseline	53,41 ± 2,24	69,05 ± 6,30	0,29 ± 0,00
10%	TT-scraped	65,03 ± 2,08	80,16 ± 1,37	0,67 ± 0,33
100%	Supervised 100%	59,08 ± 3,60	73,81 ± 0,00	0,29 ± 0,00

Proses *Tri-Training* dan Ablasi

Ringkasan proses *Tri-Training* disajikan pada Tabel 8 dan Gambar 4. Semua *run* berhenti sendiri setelah satu sampai tiga iterasi. Jumlah *pseudo-label* hanya 3 sampai 16 email per model per iterasi, karena dengan galat gabungan sekitar 0,05 Persamaan 5 hanya mengizinkan sekitar 10 email dari lebih dari 1.400 kandidat. Pada *pool in-domain*, akurasi *pseudo-label* rata-rata 98,1%, sedangkan pada *pool scraped* kesesuaiannya dengan label sumber 98,7% untuk IETF, 97,5% untuk ScamWarners, dan 88,4% untuk Phishing Pot. Ablasi pada Tabel 9 menunjukkan bahwa pada uji eksternal efek *Tri-Training* pada komite mencapai 7,79 ± 3,77 poin, jauh lebih besar daripada efek WCE (0,59 ± 2,56 poin), dengan interaksi −2,72 ± 1,09 poin. Pada uji internal kedua efek hampir nol, dan WCE hanya menaikkan *recall phishing* komite *baseline* dari 95,51% menjadi 96,30%.

Tabel 8. Ringkasan Proses Tri-Training Zhou dan Li (Tiga Seed)

Porsi	Pool	Iterasi berjalan	Iterasi terpilih (seed 42, 43, 44)	Pseudo-label per model	F1-macro komite validasi awal (%)	F1-macro komite validasi terpilih (%)	Waktu per run (menit)
1%	scraped	1 s.d. 3	2, 1, 3	7,2	89,57 ± 3,69	91,50 ± 4,76	6,9
2%	scraped	2 s.d. 3	2, 1, 2	8,6	90,47 ± 1,04	93,65 ± 1,50	7,1
5%	scraped	2 s.d. 3	1, 2, 3	7,1	93,49 ± 1,29	94,62 ± 1,20	6,3
5%	in-domain	1 s.d. 2	1, 1, 1	10,0	93,49 ± 1,29	93,88 ± 1,18	4,0
10%	scraped	2 s.d. 3	1, 2, 1	8,6	94,19 ± 0,18	95,55 ± 0,25	7,0



Gambar 4. F1-macro komite validasi per iterasi, jumlah *pseudo-label* per model, dan kesesuaian *pseudo-label* dengan sumber (*seed* 42)

Tabel 9. Ablasi 2×2 Weighted Cross-Entropy dan Tri-Training pada Komite Porsi 5% (%)

Model	Kode	WCE	Tri-Training	F1-macro internal	Recall phishing internal	FPR internal	AUC internal	F1-macro eksternal
Komite	a	Tidak	Tidak	97,54 ± 0,67	95,51 ± 1,81	0,92 ± 0,17	99,58 ± 0,24	74,00 ± 3,19
	b	Ya	Tidak	97,67 ± 0,55	96,30 ± 1,21	1,26 ± 0,22	99,58 ± 0,26	75,95 ± 3,73
	c	Tidak	Ya	97,59 ± 0,88	97,39 ± 0,95	2,13 ± 1,31	99,54 ± 0,31	83,14 ± 3,19
	d	Ya	Ya	97,64 ± 0,69	96,96 ± 1,64	1,74 ± 0,87	99,63 ± 0,14	82,38 ± 5,09

Pembanding TF-IDF dan Efisiensi

Tabel 10 menunjukkan bahwa TF-IDF dengan *Logistic Regression* tertinggal 0,8 sampai 4,7 poin dari komite *baseline* pada uji internal, tetapi pada uji eksternal hanya mencapai 64,15% sampai 72,93%. Dengan seluruh data latih, TF-IDF mencapai 99,27% pada uji internal tetapi hanya 74,57% pada uji eksternal. Tabel 11 memperlihatkan bahwa DistilBERT memiliki 61% jumlah parameter BERT dengan latensi 5,29 ms berbanding 10,00 ms, atau 1,89 kali lebih cepat, dan F1-macro internalnya setelah *Tri-Training* (96,87%) lebih tinggi daripada BERT (95,59%).

Tabel 10. Perbandingan F1-macro TF-IDF dan Logistic Regression dengan Komite Transformer (%)

Porsi	TF-IDF + LR internal	TF-IDF + LR eksternal	Komite baseline internal	Komite baseline eksternal	Komite Tri-Training internal	Komite Tri-Training eksternal
1%	90,11	64,15	94,80	72,74	94,94	78,25
2%	94,58	70,17	95,97	69,74	96,92	80,52
5%	96,61	72,06	97,67	75,95	97,64	82,38
10%	97,60	72,93	98,40	75,41	98,49	81,88
100% (supervised)	99,27	74,57	99,37	78,65	–	–

Tabel 11. Efisiensi Inferensi dan Kinerja Tri-Training Porsi 5% per Model

Model	Parameter (juta)	Ukuran FP32 (MB)	Latensi (ms/email)	Throughput (email/detik)	Kecepatan relatif BERT	F1-macro internal (%)	F1-macro eksternal (%)
BERT	109,48	417,65	10,00	309,22	1,00×	95,59	85,49
DistilBERT	66,96	255,41	5,29	585,41	1,89×	96,87	77,50
RoBERTa	124,65	475,49	8,59	311,96	1,16×	97,50	83,75

PEMBAHASAN

Pertanyaan pertama terjawab dengan dua hasil yang tampak berlawanan. Pada uji internal, *Tri-Training* tidak bermanfaat karena efek plafon, sedangkan pada uji eksternal ruang perbaikannya besar dan di sanalah *Tri-Training* bekerja. Temuan ini sejalan dengan Shu dkk. bahwa data tanpa label paling bermanfaat ketika label terbatas [20] dan dengan Boulila dkk. bahwa email modern berbeda dari korpus lama [18]. Model *supervised* penuh yang kalah dari komite *Tri-Training* porsi 5% pada uji eksternal menunjukkan bahwa menambah label dari distribusi lama tidak menyelesaikan pergeseran domain.

Untuk pertanyaan kedua, efek WCE kecil, sedangkan efek *Tri-Training* dominan pada uji eksternal. Pada porsi 5%, *pool scraped* menaikkan F1-macro eksternal 6,43 poin dan *pool in-domain* 3,70 poin, sehingga sekitar 2,7 poin dapat dianggap sumbangan email modern. Sisa kenaikan berasal dari pelatihan ulang dan pemilihan iterasi pada validasi yang mencakup 325 email modern. Fluktuasi F1-macro per model selama iterasi tetap muncul walaupun *pseudo-label* 100% benar, sehingga sumbernya adalah variasi *fine-tuning* pada data kecil [39], dan komite meredam variasi tersebut. Perbandingan dengan penelitian terdahulu pada Tabel 12 menunjukkan bahwa F1 di atas 0,93 pada uji internal belum menggambarkan kinerja di lapangan.

Tabel 12. Perbandingan dengan Penelitian Terdahulu

Penelitian	Data	Skema pelatihan dan pengujian	Model	Hasil
Otieno dkk. [12]	11.270 email Nazario dan Enron (seimbang)	Supervised penuh, split acak 80/20	BERT	F1 per kelas 0,93; akurasi 0,93
Ibrahim dan Elhafiz [13]	19.000 email Nazario dan Enron	Supervised penuh, split acak 80/20	BERT; RoBERTa	F1 0,94; 0,96 (jenis rata-rata tidak disebutkan)
Meléndez dkk. [11]	Gabungan beberapa korpus phishing dan Enron	Supervised penuh	DistilBERT, BERT, RoBERTa	Akurasi tertinggi di atas 0,99 (RoBERTa)
Penelitian ini	7.665 email Zenodo, split per kelompok; uji eksternal 739 email modern	Supervised penuh (5.364 label)	Komite tiga model	F1-macro internal 0,9937; eksternal 0,7865
Penelitian ini	7.665 email Zenodo, split per kelompok; uji eksternal	Tri-Training, 5% label (268 email)	Komite tiga model	F1-macro internal 0,9764; eksternal 0,8238

	739 email modern			

Untuk pertanyaan ketiga, RoBERTa paling tahan terhadap pergeseran domain, BERT paling banyak memperoleh manfaat *Tri-Training* tetapi paling tidak stabil, dan DistilBERT paling efisien. Susunan dua tahap dapat menjadi kompromi, yaitu DistilBERT menyaring seluruh email lalu RoBERTa atau komite memeriksa email yang probabilitasnya dekat ambang [14]. Keterbatasan penelitian ini adalah label uji eksternal yang ditentukan dari sumber, jumlah email ScamWarners yang kecil, pemakaian validasi modern untuk pemilihan iterasi, sisa token message- pada 7,48% email *legitimate*, serta tiga *seed*, bahasa Inggris, dan panjang 256 token.

KESIMPULAN

Tri-Training Zhou dan Li dengan BERT, DistilBERT, dan RoBERTa tidak mengubah kinerja secara signifikan pada uji internal karena *baseline* sudah mencapai F1-macro 94,80% sampai 98,40%. Pada uji eksternal modern, F1-macro komite naik 5,51 sampai 10,78 poin pada keempat porsi label dengan 10 dari 12 perbandingan signifikan setelah koreksi Holm, dan pada porsi 5% melampaui model *supervised* penuh. Kenaikan tersebut terutama berasal dari *Tri-Training* (7,79 poin), bukan dari WCE (0,59 poin), dan sekitar 2,7 poin berasal dari *pool* email modern. Aturan Zhou dan Li menghasilkan 3 sampai 16 *pseudo-label* per model per iterasi dan berhenti sendiri setelah satu sampai tiga iterasi. DistilBERT 1,89 kali lebih cepat daripada BERT, sedangkan RoBERTa paling tahan terhadap email modern. Penelitian selanjutnya disarankan melakukan anotasi manual pada email modern, menambah *seed*, memisahkan validasi dari sumber uji eksternal, dan menguji email berbahasa Indonesia.

REFERENCES

[1] Anti-Phishing Working Group, "Phishing Activity Trends Report, 4th Quarter 2024," APWG, Cambridge, MA, USA, Mar. 2025. [Online]. Available: https://docs.apwg.org/reports/apwg_trends_report_q4_2024.pdf

[2] Federal Bureau of Investigation, Internet Crime Complaint Center, "2024 Internet Crime Report," FBI IC3, Washington, DC, USA, Apr. 2025. [Online]. Available: https://www.ic3.gov/AnnualReport/Reports/2024_IC3Report.pdf

[3] C. D. Hylender, P. Langlois, A. Pinto, and S. Widup, "2024 Data Breach Investigations Report," Verizon Business, 2024. [Online]. Available: https://www.verizon.com/business/resources/reports/2024-dbir-data-breach-investigations-report.pdf

[4] A. Basit, M. Zafar, X. Liu, A. R. Javed, Z. Jalil, and K. Kifayat, "A comprehensive survey of AI-enabled phishing attacks detection techniques," *Telecommunication Systems*, vol. 76, no. 1, pp. 139–154, 2021, doi: 10.1007/s11235-020-00733-2.

[5] S. Salloum, T. Gaber, S. Vadera, and K. Shaalan, "A systematic literature review on phishing email detection using natural language processing techniques," *IEEE Access*, vol. 10, pp. 65703–65727, 2022, doi: 10.1109/ACCESS.2022.3183083.

[6] A. R. Kamila, F. Adikara, S. Sutrisno, and C. Herdian, "Analisa pengaruh penambahan fitur dengan perbandingan algoritma berbasis bagging dan boosting pada deteksi phishing link," *JEPIN (Jurnal Edukasi dan Penelitian Informatika)*, vol. 10, no. 3, pp. 460–467, Dec. 2024. [Online]. Available: https://jurnal.untan.ac.id/index.php/jepin/article/view/83366

[7] F. F. Tampinongkol, A. R. Kamila, A. C. Wardhana, A. W. C. Kusuma, and D. Revaldo, "Implementation of random forest classification and support vector machine algorithms for phishing link detection," *Journal of Informatics, Information System, Software Engineering and Applications (INISTA)*, vol. 7, no. 1, pp. 127–137, 2024, doi: 10.20895/inista.v7i1.1588.

[8] F. F. Tampinongkol, R. Ilham, A. R. Kamila, Y. Purnomo, C. Herdian, and S. Virginia, "Deteksi ciri link phishing menggunakan algoritma random forest untuk meningkatkan keamanan cyber," *Techno Xplore: Jurnal Ilmu Komputer dan Teknologi Informasi*, vol. 9, no. 2, pp. 60–67, Oct. 2024. [Online]. Available: https://journal.ubpkarawang.ac.id/index.php/TeknikInformatikaSistemInfor/article/view/7664

[9] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, vol. 30, 2017, pp. 5998–6008. [Online]. Available: https://arxiv.org/abs/1706.03762

[10] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. NAACL-HLT*, Minneapolis, MN, USA, 2019, pp. 4171–4186, doi: 10.18653/v1/N19-1423.

[11] R. Meléndez, M. Ptaszynski, and F. Masui, "Comparative investigation of traditional machine-learning models and transformer models for phishing email detection," *Electronics*, vol. 13, no. 24, Art. no. 4877, 2024, doi: 10.3390/electronics13244877.

[12] D. O. Otieno, A. Siami Namin, and K. S. Jones, "The application of the BERT transformer model for phishing email classification," in *Proc. IEEE 47th Annual Computers, Software, and Applications Conference (COMPSAC)*, Torino, Italy, 2023, pp. 1303–1310, doi: 10.1109/COMPSAC57700.2023.00198.

[13] M. Ibrahim and R. Elhafiz, "Phishing email detection using BERT and RoBERTa," *Computation*, vol. 14, no. 2, Art. no. 46, 2026, doi: 10.3390/computation14020046.

[14] E. M. Damatie, A. Eleyan, and T. Bejaoui, "Real-time email phishing detection using a custom DistilBERT model," in *Proc. 2024 International Symposium on Networks, Computers and Communications (ISNCC)*, Washington, DC, USA, 2024, pp. 1–6, doi: 10.1109/ISNCC62547.2024.10759011.

[15] M. B. M. Amin, G. Hakim, M. T. Maulana, M. F. Alwan, H. S. Anggraheni, M. J. Naufal, and N. Yudistira, "Deteksi spam berbahasa Indonesia berbasis teks menggunakan model BERT," *Jurnal Teknologi Informasi dan Ilmu Komputer (JTIIK)*, vol. 11, no. 6, pp. 1291–1302, 2024, doi: 10.25126/jtiik.2024118121.

[16] Y. R. Purnamadewi and A. Zahra, "Enhancing detection of zero-day phishing email attacks in the Indonesian language using deep learning algorithms," *Bulletin of Electrical Engineering and Informatics*, vol. 14, no. 1, pp. 505–512, 2025, doi: 10.11591/eei.v14i1.8759.

[17] O. N. Mbadiwe, O. C. Nwokonkwo, A. I. Otuonye, C. O. Ikerionwu, and C. Etus, "Challenges of data collection and preprocessing for phishing email detection," *International Journal of Novel Research in Computer Science and Software Engineering*, vol. 11, no. 2, pp. 45–59, 2024, doi: 10.5281/zenodo.12651047.

[18] E. Boulila, M. Dacier, S. P. V. Peroumal, N. Veys, and S. Aonzo, "A closer look at modern evasive phishing emails," in *Proc. 55th Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN)*, Naples, Italy, Jun. 2025. [Online]. Available: https://www.eurecom.fr/publication/8200

[19] M.-R. Amini, V. Feofanov, L. Pauletto, L. Hadjadj, É. Devijver, and Y. Maximov, "Self-training: A survey," *Neurocomputing*, vol. 616, Art. no. 128904, 2025, doi: 10.1016/j.neucom.2024.128904.

[20] R. Shu, T. Xia, H. Tu, L. Williams, and T. Menzies, "Reducing the cost of training security classifier (via optimized semi-supervised learning)," arXiv:2205.00665, 2022, doi: 10.48550/arXiv.2205.00665.

[21] P. Cascante-Bonilla, F. Tan, Y. Qi, and V. Ordonez, "Curriculum labeling: Revisiting pseudo-labeling for semi-supervised learning," in *Proc. AAAI Conference on Artificial Intelligence*, vol. 35, no. 8, 2021, pp. 6912–6920, doi: 10.1609/aaai.v35i8.16852.

[22] S. Saifullah, R. Dreżewski, F. A. Dwiyanto, A. S. Aribowo, Y. Fauziah, and N. H. Cahyana, "Automated text annotation using a semi-supervised approach with meta vectorizer and machine learning algorithms for hate speech detection," *Applied Sciences*, vol. 14, no. 3, Art. no. 1078, 2024, doi: 10.3390/app14031078.

[23] S. Handayani, R. Isnanto, and B. Warsito, "Co-training pseudo-labeling for text classification with support vector machine and long short-term memory," *IAES International Journal of Artificial Intelligence (IJ-AI)*, vol. 14, no. 3, pp. 2158–2168, 2025, doi: 10.11591/ijai.v14.i3.pp2158-2168.

[24] A. S. Aribowo, S. Khomsah, and S. Saifullah, "Semi-supervised sentiment classification using self-learning and enhanced co-training," *Jurnal Infotel*, vol. 17, no. 3, pp. 472–489, 2025, doi: 10.20895/infotel.v17i3.1344.

[25] Z.-H. Zhou and M. Li, "Tri-training: Exploiting unlabeled data using three classifiers," *IEEE Transactions*

on Knowledge and Data Engineering, vol. 17, no. 11, pp. 1529–1541, Nov. 2005, doi: 10.1109/TKDE.2005.186.

[26] A. I. Champa, M. F. Rabbi, and M. F. Zibran, "Why phishing emails escape detection: A closer look at the failure points," in *Proc. 12th International Symposium on Digital Forensics and Security (ISDFS)*, San Antonio, TX, USA, 2024, pp. 1–6, doi: 10.1109/ISDFS60797.2024.10527344.

[27] A. I. Champa, M. F. Rabbi, and M. F. Zibran, "Curated datasets and feature analysis for phishing email detection with machine learning," in *Proc. 3rd IEEE International Conference on Computing and Machine Intelligence (ICMI)*, Mt Pleasant, MI, USA, 2024, pp. 1–7, doi: 10.1109/ICMI60790.2024.10585821.

[28] A. I. Champa, M. F. Rabbi, and M. F. Zibran, "Phishing Email Curated Datasets," Zenodo, 2023, doi: 10.5281/zenodo.8339691.

[29] B. Klimt and Y. Yang, "The Enron corpus: A new dataset for email classification research," in *Machine Learning: ECML 2004*, Lecture Notes in Computer Science, vol. 3201. Berlin, Germany: Springer, 2004, pp. 217–226, doi: 10.1007/978-3-540-30115-8_22.

[30] ScamWarners, "ScamWarners: Scam warning and anti-fraud forum," 2026. [Online]. Available: https://www.scamwarners.com (diakses 1 Nov. 2026).

[31] rf-peixoto, "Phishing Pot: A collection of phishing samples for researchers and detection developers," GitHub repository, lisensi CC BY-NC 4.0, 2026. [Online]. Available: https://github.com/rf-peixoto/phishing_pot (diakses 2 Nov. 2026).

[32] Internet Engineering Task Force, "IETF mail archive: ietf@ietf.org monthly text archives 2015–2022," 2026. [Online]. Available: https://www.ietf.org/ietf-ftp/ietf-mail-archive/ietf/ (diakses 2 Nov. 2026).

[33] M. Koster, G. Illyes, H. Zeller, and L. Sassman, "Robots Exclusion Protocol," Internet Engineering Task Force, RFC 9309, Sep. 2022, doi: 10.17487/RFC9309.

[34] Republik Indonesia, "Undang-Undang Nomor 27 Tahun 2022 tentang Pelindungan Data Pribadi," Lembaran Negara Republik Indonesia Tahun 2022 Nomor 196, Jakarta, 2022. [Online]. Available: https://peraturan.bpk.go.id/Details/229798/uu-no-27-tahun-2022

[35] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A robustly optimized BERT pretraining approach," arXiv:1907.11692, 2019, doi: 10.48550/arXiv.1907.11692.

[36] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter," arXiv:1910.01108, 2019, doi: 10.48550/arXiv.1910.01108.

[37] J. Gou, B. Yu, S. J. Maybank, and D. Tao, "Knowledge distillation: A survey," *International Journal of Computer Vision*, vol. 129, no. 6, pp. 1789–1819, 2021, doi: 10.1007/s11263-021-01453-z.

[38] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *Proc. International Conference on Learning Representations (ICLR)*, New Orleans, LA, USA, 2019. [Online]. Available: https://arxiv.org/abs/1711.05101

[39] M. Mosbach, M. Andriushchenko, and D. Klakow, "On the stability of fine-tuning BERT: Misconceptions, explanations, and strong baselines," in *Proc. International Conference on Learning Representations (ICLR)*, 2021. [Online]. Available: https://arxiv.org/abs/2006.04884

[40] T. G. Dietterich, "Approximate statistical tests for comparing supervised classification learning algorithms," *Neural Computation*, vol. 10, no. 7, pp. 1895–1923, 1998, doi: 10.1162/089976698300017197.

[41] S. Holm, "A simple sequentially rejective multiple test procedure," *Scandinavian Journal of Statistics*, vol. 6, no. 2, pp. 65–70, 1979. [Online]. Available: https://www.jstor.org/stable/4615733