KLASIFIKASI PDF MALWARE PADA LAYANAN AGREGATOR NASIONAL (GARUDA) KEMDIKBUD DIKTI DENGAN METODE NAIVE BAYES CLASSIFIER

OLEH :

Nata Arista

09011181823128

JURUSAN SISTEM KOMPUTER

FAKULTAS ILMU KOMPUTER

UNIVERSITAS SRIWIJAYA

2023

Klasifikasi *PDF Malware* Pada Layanan Agregator Nasional (GARUDA) Kemdikbud Dikti dengan Metode *Naive Bayes Classifier*

TUGAS AKHIR

Diajukan Untuk Melengkapi Salah Satu Syarat

Memperoleh Gelar Sarjana Komputer

OLEH :

Nata Arista

09011181823128

JURUSAN SISTEM KOMPUTER

FAKULTAS ILMU KOMPUTER

UNIVERSITAS SRIWIJAYA

2023

LEMBAR PENGESAHAN

Klasifikasi *PDF Malware* Pada Layanan Agregator Nasional (GARUDA) Kemdikbud Dikti dengan Metode *Naive Bayes Classifier*

TUGAS AKHIR

Diajukan untuk Melengkapi Salah Satu Syarat

Memperoleh Gelar Sarjana Komputer

Oleh :

Nata Arista

09011181823128

Indralaya, 7 Juli 2023

Mengetahui,

Pembimbing I	Pembimbing II
Deris Stiawan, M.T., Ph.D	**Nurul Afifah, M.Kom**
NIP. 197806172006041002	**NIP. -**

Ketua Jurusan Sistem Komputer

Dr. Ir. Sukemi, M.T

NIP. 196612032006041001

HALAMAN PERSETUJUAN

Telah diuji dan lulus pada :

Hari : Jum'at

Tanggal : 07 Juli 2023

Tim Penguji :

1. **Ketua** : **Ahmad Heryanto, M.T.**
2. **Sekretaris** : **Aditya Putra Perdana P, M.T**
3. **Penguji** : **Sarmayanta Sembiring, M.T**
4. **Pendamping I** : **Deris Stiawan, M.T., Ph.D**
5. **Pendamping II** : **Nurul Afifah, M.Kom**

Mengetahui,

Ketua Jurusan Sistem Komputer

Dr.Ir.Sukemi, M.T.

NIP.196612032006041001

HALAMAN PERNYATAAN

Yang bertanda tangan di bawah ini :

Nama : Nata Arista

NIM : 09011181823128

Judul : Klasifikasi PDF *Malware* pada Layanan Agregator Nasional (GARUDA) Kemdikbud Dikti dengan Metode *Naive Bayes Classifier*

Hasil Pengecekan Software iThenticate/Turnitin : 3%

Menyatakan bahwa laporan tugas akhir saya merupakan hasil karya sendiri dan bukan hasil penjiplakan atau plagiat. Apabila ditemukan unsur penjiplakan atau plagiat dalam laporan tugas akhir ini, maka saya bersedia menerima sanksi akademik dari Universitas Sriwijaya.

Demikian, pernyataan ini saya buat dalam keadaan sadar dan tanpa paksaan dari siapapun.

Indralaya, Juli 2023

Nata Arista
09011181823128

KATA PENGANTAR

Assalamu'alaikum Warahmatullahi Wabarakatuh

Puji dan syukur penulis haturkan atas kehadirat Allah SWT, yang telah memberikan rahmat dan karunia-Nya berupa akal pikiran, ilmu pengetahuan kesehatan dan kekuatan sehingga penulis dapat menyelesaikan penyusunan tugas akhir ini dengan judul **Klasifikasi PDF *Malware* Pada Layanan Agregator Nasional (GARUDA) Kemdikbud Dikti dengan Metode *Naive Bayes Classifier*.**

Pada penyusunan tugas akhir ini, tidak lepas dari motivasi, semangat, bimbingan dan dukungan dari berbagai pihak. Oleh karena itu, pada kesempatan ini penulis mengucapkan rasa syukur dan terima kasih kepada :

1. Allah Subhanahu Wata'ala yang telah memberikan rahmat dan karunia-Nya kepada penulis dalam penyusunan tugas akhir ini.
2. Orangtua tercinta (Sarijo dan Ngatmiati) yang selalu memberikan dukungan baik moral maupun finansial, semangat serta do'a yang tiada hentinya.
3. Keluarga besar penulis yang tersayang. Terima kasih atas semua kebaikan dan dukungan yang diberikan.
4. Adik satu-satunya (Hanan Witjaya) yang selalu memberi semangat dan do'a.
5. Bapak Jaidan Jauhari, S.Pd. M.T selaku Dekan Fakultas Ilmu Komputer Universitas Sriwijaya.
6. Bapak Dr. Ir. H. Sukemi, M.T. selaku Ketua Jurusan Sistem Komputer Universitas Sriwijaya.
7. Bapak Dr. Erwin, M.Si. selaku Dosen Pembimbing Akademik penulis di Jurusan Sistem Komputer
8. Bapak Deris Stiawan, M.T., Ph.D., IPU., ASEAN-Eng selaku Pembimbing I Tugas Akhir penulis di Jurusan Sistem Komputer yang telah meluangkan waktu untuk membimbing dan memberikan motivasi selama kuliah dan pengerjaan Tugas Akhir.

9. Mbak Nurul Afifah, M.Kom. selaku pembimbing II yang telah membimbing penulis dalam pengerjaan Tugas Akhir dari awal pembuatan dataset sampai pembenaran dalam penulisan laporan Tugas Akhir.
10. Kak Tri Wanda Septian, M.Sc. yang telah meluangkan waktu untuk membimbing penulis dalam pembuatan Tugas Akhir dari awal pembuatan dataset sampai selesai laporan Tugas Akhir, serta dukungan dan motivasi yang diberikan selama masa kuliah.
11. Mbak Renny Virgasari selaku Admin Jurusan Sistem Komputer yang baik dan ramah dalam membantu administrasi Tugas Akhir.
12. Orang tersayang yang selalu support, selalu ada, dan selalu menemani selama proses Tugas Akhir (M. Daul Angfal), terima kasih yang sebesar-besarnya ayang, love you.
13. Teman- teman satu kelompok riset yang selalu memberi solusi dan semangat Tri Shena Orivia Pasin, Alfiah Nur Fatmawati, Rani Octaviani, Alifah Fidela, Indah Cahya Resti, dan Novi Yuningsih. Sukses untuk kita semua guys!
14. Teman – teman Kerja Praktik yang jadi teman terdekat selama satu bulan, terima kasih yaa Haqiqi Oktaviani, Jarna Ajda, Ahmad Ramdoni Kusduandi, M. Al Insyirah Satria Harahap, dan Yusdiansya Putra.
15. Sahabat SMP yang selalu ada dalam setiap kondisi, Ria Nita Anggraini, Rika Astuti, Eriska Witantri Budiarti, Ermawati, Nurdiana Putri. Sehat selalu ya kalian.
16. Liana Indriani dan Mia Kurnia selaku teman kost, teman satu daerah yang selalu menjadi tempat cerita tentang apa pun yang penulis rasakan. Teman yang selalu ada dan membantu saat susah, serta merawat saat sakit. Terima kasih banyak, sukses bareng yaaa !
17. Kakak- kakak tingkat yang menjadi panutan, teman-teman seperjuangan Jurusan Sistem Komputer Angkatan 2018 terkhusus kelas A, serta semua orang baik yang sempat hadir dalam kehidupan penulis yang tidak dapat penulis cantumkan satu persatu.
18. Civitas Akademika Fakultas Ilmu Komputer Universitas Sriwijaya.
19. Almamater Universitas Sriwijaya.

Penulis menyadari bahwa masih ada banyak kekurangan dalam penulisan laporan tugas akhir ini. Mengingat kurangnya pengetahuan dan pengalaman penulis dalam hal ini. Oleh karena itu kritik dan saran yang mendukung sangat penting bagi penulis.

Akhir kata dengan segala keterbatasan, penulis berharap semoga laporan ini menghasilkan sesuatu yang bermanfaat bagi kita semua khususnya bagi mahasiswa Fakultas Ilmu Komputer Universitas Sriwijaya secara langsung ataupun tidak langsung sebagai sumbangan pikiran dalam peningkatanmutu pembelajaran.

Wassalamu'alaikum Warahmatullahi Wabarakatuh.

Indralaya, Juli 2023
Penulis

Nata Arista
NIM :09011181823128

PDF MALWARE CLASSIFICATION ON NATIONAL AGGREGATOR SERVICE (GARUDA) KEMDIKBUD DIKTI USING NAIVE BAYES CLASSIFICATION METHOD

NATA ARISTA (09011181823128)

Computer Engineering Department, Computer Science Faculty, Sriwijaya University

Email : nataarista245111@gmail.com

ABSTRACT

Garba Rujukan Digital (GARUDA) is one of the e-library in Indonesian academic that uses PDF as a file extension. The dataset used in this study came from the GARUDA repository with 10000 data consisting of 9800 benign, 196 malicious html, and six malicious pdf. The dataset was analyzed using VirusTotal, PDF-parser and PDFid. The classification process is carried out in three stages using the Gaussian Naive Bayes method, we called : (i) classification on the imbalance dataset which result an accuracy value of 0.2280, (ii) classification with the addition of SMOTE to overcome the imbalance dataset which results in accuracy value of 0.7361, (iii) classification with a combination of SMOTE and Near Miss to overcome the imbalance dataset which results in accuracy value of 0.9643, precision 0.9640, recall 0.9675 and f1-score 0.9639.

Keyword *: Classification, PDF Malware, VirusTotal, PDFid, Gaussian Naive Bayes, SMOTE, Near Miss.*

KLASIFIKASI *PDF MALWARE* PADA LAYANAN AGREGATOR NASIONAL (GARUDA) KEMDIKBUD DIKTI DENGAN *METODE NAIVE BAYES CLASSIFIER*

NATA ARISTA (09011181823128)

Jurusan Sistem Komputer, Fakultas Ilmu Komputer, Universitas Sriwijaya

Email : nataarista245111@gmail.com

ABSTRAK

Garba Rujukan Digital (GARUDA) merupakan salah satu *e-library* akademisi Indonesia yang memakai PDF sebagai ekstensi *file*. *Dataset* yang digunakan dalam penelitian ini berasal dari portal GARUDA dengan data sebanyak 10000 yang terdiri dari 9800 *benign*, 196 *malicious html*, dan enam *malicious pdf*. *Dataset* dianalisis menggunakan VirusTotal, PDF-parser dan PDFid. Proses klasifikasi dilakukan sebanyak tiga tahap menggunakan metode *Gaussian Naive Bayes*, yaitu: (i) klasifikasi pada *dataset imbalance* yang menghasilkan nilai akurasi sebesar 0,2280, (ii) klasifikasi dengan penambahan SMOTE untuk mengatasi *dataset imbalance* yang meghasilkan nilai akurasi sebesar 0,7361, (iii) klasifikasi dengan gabungan SMOTE dan *Near Miss* untuk mengatasi *dataset imbalance* yang menghasilkan nilai akurasi sebesar 0,9643, presisi 0,9640, *recall* 0,9675 dan *f1-score* 0,9639.

Kata Kunci : Klasifikasi, *PDF Malware*, VirusTotal, PDFid, *Gaussian Naive Bayes*, SMOTE, *Near Miss*.

DAFTAR ISI

Halaman

DAFTAR GAMBAR

Halaman

DAFTAR TABEL

Halaman

DAFTAR LAMPIRAN

BAB I

PENDAHULUAN

1.1 Latar Belakang

File Portabel Document Format atau sering disebut dengan PDF sudah sangat umum digunakan seperti pada e-*journal*, *essay*, karya ilmiah, makalah dan lain sebagainya. Hal ini dikarenakan kemudahannya dalam penggunaan dan pengiriman. *File* dengan format PDF juga dapat dibuka pada hampir seluruh sistem operasi. *File* dengan format ini dapat berisi tulisan, angka, gambar, skrip dan masih banyak lagi. Banyak perpustakaan digital yang menyajikan dokumen ilmiah dengan format PDF yang dapat diakses dengan bebas. Garba Rujukan Digital (GARUDA) [1] merupakan salah satu portal yang memuat rujukan ilmiah, karya ilmiah, hasil penelitian, *e-journal*, skripsi atau tesis dari peneliti akademisi Indonesia. *Library* yang tersedia dalam portal ini berasal dari berbagai perguruan tinggi, ilmuwan dan lembaga penelitian di Indonesia.

Diperkirakan lebih dari 114 juta dokumen PDF di internet, dimana lebih dari 27 juta (24%) dapat diakses dengan mudah dan tanpa berlangganan [2]. Hal ini menjadikan banyak orang lebih memilih untuk mencari referensi di internet karena dapat diakses kapan dan dimana saja serta tidak terbatas pada satu sistem operasi. *File* PDF memiliki beberapa fitur fungsional seperti *binary file, compression, management font, single-pass file generation, random access, security, incremental update, extensibility* [3]. Dari beberapa fitur tersebut, ada beberapa fitur yang membuat para *hacker* tertarik untuk menyisipkan berbagai jenis konten *malware* ke dalam *file* PDF.

Malware merupakan program komputer berbahaya yang sengaja diciptakan *hacker* dengan tujuan jahat seperti untuk menyusup dan merusak *software* atau sistem operasi target. Adapun serangan yang paling populer dilakukan pada *file* PDF adalah penyematan kode berbahaya *Java Script* kedalam *file* PDF. Untuk menghindari deteksi, *hacker* akan memakai berbagai cara untuk menyamarkan

malware, yaitu dengan cara menyuntikkan *malware* ke PDF *benign* tanpa melewati ambang batas *benign*, sehingga *file* tidak terdeteksi jika mengandung *malware*. Serangan ini disebut dengan mimikri [4]. Seiring berkembangnya teknologi, volume *malware* semakin meningkat hingga mencapai angka ribuan setiap harinya. Sebagian sampel merupakan varian *malware* yang ada, kemudian diproduksi dengan mengubah atau mengaburkan sehingga terhindar dari deteksi anti virus [5].

Penelitian ini akan berfokus pada proses klasifikasi *file* PDF *malware*. Proses klasifikasi ini bertujuan untuk memisahkan antara PDF *benign* dan PDF *malware*. Penelitian ini menggunakan dari 10.000 *file* PDF dengan sampel PDF *malware* sebanyak 200 *file*. Banyak metode yang telah dievaluasi menggunakan *dataset* yang seimbang (*balance*), sehingga hal ini menjadi menarik karena *dataset* yang digunakan dalam penelitian ini merupakan data tidak seimbang (*imbalance*) dimana jumlah *file* PDF *benign* lebih banyak daripada PDF *malware*.

Dataset yang tidak seimbang seringkali menurunkan kinerja dari algoritma pembelajaran. Hasil yang diperoleh akan cenderung menguntungkan kelas mayoritas. Bahkan dalam kasus klasifikasi *multiclass*, ketidakseimbangan data ini mengakibatkan representasi data minoritas yang rendah dan cenderung diabaikan [6]. *Synthetic Minority Oversampling Technique* (SMOTE) dan *Near Miss* merupakan metode *oversampling* dan *undersampling* yang banyak digunakan dalam menangani masalah *dataset imbalance*. Beberapa penelitian menunjukkan pengaruh dari SMOTE [7][8][9] dan *Near Miss* [6] yang dapat mengatasi *dataset imbalance* serta meningkatkan performasi dari metode *classifier*.

Naive Bayes Classifier merupakan salah satu dari sepuluh metode teratas dalam proses klasifikasi pada *machine learning*. *Naive Bayes Classifier* menggambarkan probabilitas dari sebuah peristiwa berdasarkan pengetahuan sebelumnya terkait dengan peristiwa tersebut. Beberapa penelitian [10][11][12][13] menghasilkan nilai akurasi yang tinggi ketika menggunakan algoritma *Naive Bayes Classifier* yaitu lebih dari 85%. Bahkan pada penelitian

cancer classification [14], *Gaussian Naive Bayes* menunjukkan performa yang sangat baik yaitu dengan memperoleh nilai akurasi sebesar 98%.

Berdasarkan pembahasan diatas, penulis bermaksud untuk melakukan penelitian terhadap PDF *malware* GARUDA dengan menggunakan metode klasifikasi *Naive Bayes Classifier*. Penelitian ini dibagi menjadi tiga model klasifikasi yaitu klasifikasi dengan *dataset imbalance*, klasifikasi dengan SMOTE, dan klasifikasi dengan *Over-UnderSampling* menggunakan SMOTE dan *Near Miss*. Diharapkan dari penelitian ini dapat menghasilkan nilai akurasi, presisi, *recall* dan *f1-score* yang baik sehingga dapat menjadi referensi untuk ilmu pengetahuan terkait.

1.2 Perumusan Masalah

Berikut adalah rumusan masalah dalam penulisan Tugas Akhir ini:

1. Bagaimana penerapan serta pengaruh SMOTE dan *Near Miss* sebagai metode *resampling* untuk menangani masalah *dataset imbalance* pada *dataset* PDF *malware* GARUDA?
2. Bagaimana penerapan *Naive Bayes Classifier* dalam proses klasifikasi pada *dataset* PDF *malware* GARUDA?
3. Bagaimana hasil validasi dari model klasifikasi menggunakan algoritma *Naive Bayes Classifier* pada *dataset* PDF *malware* GARUDA?

1.3 Batasan Masalah

Agar penelitian mengarah pada pemaparan yang diharapkan, maka diperlukan batasan masalah dalam penelitian ini. Adapun batasan masalah tersebut adalah sebagai berikut.

1. *Dataset* yang digunakan dalam penelitian ini berbatas pada *dataset* yang berasal dari Layanan Agregator Garba Rujukan Digital (GARUDA) Kemdikbud Dikti.
2. Menggunakan SMOTE dan *Near Miss* sebagai metode *resampling* untuk menangani *dataset imbalance* pada *dataset* PDF *malware* GARUDA.
3. Melakukan klasifikasi PDF *benign*, *mal-html* dan *mal-pdf* dengan menggunakan program *Python* dan algoritma *Naive Bayes Classifier*.

4. Tidak membahas bagaimana *malware* dapat masuk ke dalam *file* PDF dan cara pencegahannya.
5. Nilai performasi yang diukur adalah akurasi, presisi, *recall* dan *f1-score*.

1.4 Tujuan

Berikut adalah tujuan dari penulisan Tugas Akhir ini :

1. Menerapakan SMOTE dan *Near Miss* untuk menangani masalah *dataset imbalance* pada *dataset* PDF *malware* GARUDA.
2. Melakukan klasifikasi pada *dataset* PDF *malware* GARUDA menggunakan algoritma *Naive Bayes Classifier*.
3. Melakukan analisa dan menarik kesimpulan pada hasil validasi yang didapatkan menggunakan *Naive Bayes Classifier* untuk memperoleh model terbaik.

1.5 Manfaat

Berikut adalah manfaat dari penulisan Tugas Akhir ini :

1. Menemukan solusi untuk masalah *dataset imbalance* pada *dataset* PDF *malware* GARUDA yaitu dengan menggunakan SMOTE dan *Near Miss*.
2. Dapat menerapkan *Naive Bayes Classifier* sebagai metode klasifikasi pada *dataset* PDF *malware* GARUDA.
3. Mampu menganalisa dan menarik kesimpulan pada hasil validasi yang didapatkan menggunakan *Naive Bayes Classifier* sehingga diperoleh model terbaik.

1.6 Metodologi Penelitian

Metodologi yang digunakan dalam penelitian ini adalah sebagai berikut.

1. Metode Studi Pustaka (*Literature*)

   Metode ini dilakukan dengan cara mencari dan mengumpulkan referensi yang berupa *literature* yang terdapat pada buku dan internet mengenai PDF *Malware*, *Naive Bayes Classifier*, *dataset imbalance*, *resampling*, dan hal-hal yang dibutuhkan dalam penelitian.

2. Metode Konsultasi

   Metode ini melakukan konsultasi kepada pihak-pihak yang memiliki pengetahuan serta wawasan yang baik dalam mengatasi permasalahan yang ditemui dalam penulisan tugas akhir.

3. Metode Pengumpulan Data

   Metode ini dilakukan dengan mengumpulkan *dataset*, yang mana dalam penelitian ini digunakan *dataset* dari portal Layanan Agregator Garba Rujukan Digital (GARUDA) Kemdikbud Dikti. *Dataset* yang diperoleh merupakan *raw* data dengan jumlah lebih dari 20.000 *file pdf*.

4. Metode Pengolahan Data

   Metode ini dilakukan dengan menganalisis *file* PDF terbelih dahulu sehingga menjadi *dataset* yang siap diolah. *Dataset* yang diperoleh merupakan *dataset imbalance* sehingga diperlukan tahap *resampling*. Pada penelitian ini tahap *resampling* dilakukan dengan menggabungkan antara teknik *oversampling* dan *undersampling*. *Oversampling* dilakukan dengan menggunakan algoritma SMOTE dan *undersampling* dilakukan dengan menggunakan algoritma *Near Miss*. Algoritma *Naive Bayes Classifier* diterapkan untuk melakukan klasifikasi *PDF benign*, *mal-html*, dan *mal-pdf*. Pembagian data dilakukan secara acak menggunakan algoritma *Satrified Kfold cross validation*.

5. Metode Analisa

   Metode ini dilakukan dengan menganalisa hasil dari pengolahan data yang kemudian divalidasi untuk mendapatkan hal-hal penting untuk dijadikan kesimpulan.

6. Metode Kesimpulan dan Saran

   Metode ini adalah metode terakhir yang dilakukan setelah mendapat hal-hal penting yang kemudian akan menjadi kesimpulan pada penelitian tugas akhir ini serta saran yang dapat dijadikan referensi bagi yang tertarik untuk meneliti lebih lanjut.

1.7 Sistematika Penulisan

Dalam penyusunan laporan tugas akhir ini, penulis membuat sistematika penulisan agar mempermudah mengetahui isi dari setiap bab yang dibuat pada laporan tugas akhir ini. Adapun sestematika penulisan laporan tugas akhir sebagai berikut :

BAB I. PENDAHULUAN

Bab ini akan menjelaskan tentang Latar Belakang Masalah, Tujuan dan Manfaat, Perumusan Masalah, Batasan Masalah, Metodologi Penelitian, serta Sistematika Penulisan.

BAB II.TINJAUAN PUSTAKA

Bab ini berisi dasar teori dari penelitian terkait dengan PDF *malware*, proses analisis *dataset*, proses *resampling* menggunakan SMOTE dan *Near Miss*, pembagian data menggunakan *Satrified Kfold*, klasifikasi dengan mengunakan metode *Naive Bayes Classifier*, dan hal-hal yang berkaitan langsung dengan penelitian.

BAB III.METODELOGI

Bab ini akan menjelaskan tentang langkah-langkah (metodologi), diagram alur (*flow chart*) dalam setiap tahap perancangan sistem pada tugas akhir.

BAB IV. ANALISA DAN PEMBAHASAN

Bab ini akan menjelaskan tentang hasil dari pengolahan data yang telah dilakukan, dari hasil tersebut akan dilakukan analisa agar mendapatkan data yang akurat. Analisa dilakukan dengan menghitung secara manual nilai yang ada pada *Confusion Matrix*. Kemudian membandingkan hasil antara klasifikasi pada *dataset PDF malware* GARUDA yang *imbalance* dan pada *dataset* PDF *malware* GARUDA yang diterapkan algoritma *resampling* menggunakan SMOTE dan *Near Miss* untuk menangani masalah *dataset imbalance*.

BAB V. KESIMPULAN DAN TINDAK LANJUT

Bab ini akan menjelaskan tentang kesimpulan yang didapat dari data penelitian yang telah dilakukan. Dan saran yang diharapkan dapat membuat penelitian ini dikembangkan lebih baik.

DAFTAR PUSTAKA

LAMPIRAN

DAFTAR PUSTAKA

[1] "Garuda - Garba Rujukan Digital," [Online]. Available: https://garuda.kemdikbud.go.id/.

[2] N. Nissim *et al.*, "Sec-lib: Protecting scholarly digital libraries from infected papers using active machine learning framework," *IEEE Access*, vol. 7, pp. 110050–110073, 2019, doi: 10.1109/ACCESS.2019.2933197.

[3] C. Ulucenk, V. Varadharajan, V. Balakrishnan, and U. Tupakula, "Techniques for analysing PDF malware," *Proc. - Asia-Pacific Softw. Eng. Conf. APSEC*, pp. 41–48, 2011, doi: 10.1109/APSEC.2011.41.

[4] A. Corum, D. Jenkins, and J. Zheng, "Robust PDF Malware Detection with Image Visualization and Processing Techniques," *Proc. - 2019 2nd Int. Conf. Data Intell. Secur. ICDIS 2019*, pp. 108–114, 2019, doi: 10.1109/ICDIS.2019.00024.

[5] I. J. Cruickshank and K. M. Carley, "Analysis of malware communities using multi-modal features," *IEEE Access*, vol. 8, pp. 77435–77448, 2020, doi: 10.1109/ACCESS.2020.2989689.

[6] A. R. B. Alamsyah, S. Rahma, N. S. Belinda, and A. Setiawan, "A R B Alamsyah et al SMOTE and Nearmiss Methods for Disease Classification with Unbalanced Data Case Study: IFLS 5," pp. 305–314, 2021, [Online]. Available: https://proceedings.stis.ac.id/icdsos/article/download/240/29/2098.

[7] V. Rattan, V. Malik, R. Mittal, and J. Singh, "Analyzing the Application of SMOTE on Machine Learning Classifiers," pp. 692–695, 2021.

[8] N. Qazi, "Effect Of Feature Selection , Synthetic Minority Over-sampling (SMOTE) And Under-sampling On Class imbalance Classification," 2012, doi: 10.1109/UKSim.2012.116.

[9] A. C. Flores and K. D. Gorro, "on Sentiment Analysis Data Set," pp. 1–4,

2018.

[10] T. M. Ma, K. Yamamori, and A. Thida, “A Comparative Approach to Naïve Bayes Classifier and Support Vector Machine for Email Spam Classification,” *2020 IEEE 9th Glob. Conf. Consum. Electron. GCCE 2020*, pp. 324–326, 2020, doi: 10.1109/GCCE50665.2020.9291921.

[11] G. Aksoy and M. Karabatak, “Performance comparison of new fast weighted Naïve bayes classifier with other bayes classifiers,” *7th Int. Symp. Digit. Forensics Secur. ISDFS 2019*, pp. 1–5, 2019, doi: 10.1109/ISDFS.2019.8757558.

[12] N. R. Fatahillah, P. Suryati, and C. Haryawan, “Implementation of Naive Bayes classifier algorithm on social media (Twitter) to the teaching of Indonesian hate speech,” *Proc. - 2017 Int. Conf. Sustain. Inf. Eng. Technol. SIET 2017*, vol. 2018-Janua, pp. 128–131, 2018, doi: 10.1109/SIET.2017.8304122.

[13] A. O. Adi, “20 Haber Grubu ’ nun Naïve Bayes Yöntemi ile S ı n ı fland ı r ı lmas ı Classification of 20 News Group with Naïve Bayes Classifier,” no. Siu, pp. 2150–2153, 2014.

[14] H. Kamel, D. Abdulah, and J. M. Al-Tuwaijari, “Cancer Classification Using Gaussian Naive Bayes Algorithm,” *Proc. 5th Int. Eng. Conf. IEC 2019*, pp. 165–170, 2019, doi: 10.1109/IEC47844.2019.8950650.

[15] R. J. Maulana and G. P. Kusuma, “Malware classification based on system call sequences using deep learning,” *Adv. Sci. Technol. Eng. Syst.*, vol. 5, no. 4, pp. 207–216, 2020, doi: 10.25046/aj050426.

[16] A. Adam, M. I. Shapiai, Z. Ibrahim, and M. Khalid, “Artificial Neural Network - Naïve Bayes Fusion for Solving Classification Problem of Imbalanced Dataset Universiti Teknologi Malaysia,” pp. 0–4, 2011.

[17] M. Su, J. Rhee, B. Kim, and B. Zhang, “AESNB : Active Example Selection with Naïve Bayes Classifier for Learning from Imbalanced Biomedical Data,” pp. 15–21, 2009, doi: 10.1109/BIBE.2009.63.

[18] S. G. Sayed and M. Shawkey, “Data Mining Based Strategy for Detecting Malicious PDF Files,” *Proc. - 17th IEEE Int. Conf. Trust. Secur. Priv. Comput. Commun. 12th IEEE Int. Conf. Big Data Sci. Eng. Trust. 2018*, pp. 661–667, 2018, doi: 10.1109/TrustCom/BigDataSE.2018.00097.

[19] S. R. Gopaldinne, H. Kaur, P. Kaur, G. Kaur, and Madhuri, “Overview of PDF Malware Classifiers,” *Proc. 2021 2nd Int. Conf. Intell. Eng. Manag. ICIEM 2021*, pp. 337–341, 2021, doi: 10.1109/ICIEM51511.2021.9445341.

[20] H. Bae, Y. Lee, Y. Kim, U. Hwang, S. Yoon, and S. Member, “Learn2Evade : Learning-Based Generative Model for Evading PDF Malware Classifiers,” vol. 2, no. 4, pp. 299–313, 2021.

[21] D. Ocuments, “Jansen, F. or P.,” *Benezit Dict. Artist.*, pp. 18–21, 2018, doi: 10.1093/benz/9780199773787.article.b00093920.

[22] M. Li, Y. Liu, M. Yu, G. Li, Y. Wang, and C. Liu, “FEPDF: A robust feature extractor for malicious PDF detection,” *Proc. - 16th IEEE Int. Conf. Trust. Secur. Priv. Comput. Commun. 11th IEEE Int. Conf. Big Data Sci. Eng. 14th IEEE Int. Conf. Embed. Softw. Syst.*, pp. 218–224, 2017, doi: 10.1109/Trustcom/BigDataSE/ICESS.2017.240.

[23] A. Charim, S. Basuki, D. R. Akbi, and U. M. Malang, “Detect Malware in Portable Document Format Files (PDF) Using Support Vector Machine and Random Decision Forest,” vol. 3, no. 2, pp. 99–102, 2019, doi: 10.15575/join.v3i2.196.

[24] N. Fleury, T. Dubrunquez, and I. Alouani, “PDF-Malware: An Overview on Threats, Detection and Evasion Attacks,” 2021, [Online]. Available: http://arxiv.org/abs/2107.12873.

[25] T. Mokoena and T. Zuva, “Malware analysis and detection in enterprise systems,” *Proc. - 15th IEEE Int. Symp. Parallel Distrib. Process. with Appl. 16th IEEE Int. Conf. Ubiquitous Comput. Commun. ISPA/IUCC 2017*, pp. 1304–1310, 2018, doi: 10.1109/ISPA/IUCC.2017.00199.

[26] “VirusTotal - Home,” [Online]. Available: https://www.virustotal.com/gui/home/upload.

[27] H. Bae, Y. Lee, Y. Kim, U. Hwang, S. Yoon, and Y. Paek, “Learn2Evade: Learning-Based Generative Model for Evading PDF Malware Classifiers,” *IEEE Trans. Artif. Intell.*, vol. 2, no. 4, pp. 299–313, 2021, doi: 10.1109/tai.2021.3103139.

[28] T. S. Sujana, N. M. S. Rao, and R. S. Reddy, “An Efficient Feature Selection using Parallel Cuckoo Search and Naïve Bayes classifier,” no. July, pp. 168–173, 2017.

[29] D. A. Cieslak, N. V. Chawla, and A. Striegel, “Combating imbalance in network intrusion datasets,” *2006 IEEE Int. Conf. Granul. Comput.*, pp. 732–737, 2006, doi: 10.1109/grc.2006.1635905.

[30] J. Lee and J. Lee, “A Classification System for Visualized Malware based on Multiple Autoencoder Models,” *IEEE Access*, vol. 9, pp. 144786–144795, 2021, doi: 10.1109/ACCESS.2021.3122083.