AICOMS
Applied Information Technology and Computer Science
e-ISSN: 2964-7703
Vol. 3 No. 1, Juni-2024, pp. 21-27

Klasifikasi Opini Masyarakat Terhadap Naturalisasi Pemain Sepak Bola Menggunakan KNN dan SMOTE

Michelle Graciela [1], Rikky [2] dan Hafiz Irsyad [3]

1 Program Studi Informatika Fakultas Ilmu Komputer dan Rekayasa Universitas Multi Data Palembang; michelle.graciela@mhs.mdp.ac.id; rikky25@mhs.mdp.ac.id; hafidirsyad@mdp.ac.id
* Michelle Graciela: michelle.graciela@mhs.mdp.ac.id

Info Artikel:
Dikirim: 22 Juni 2024
Direvisi: 24 Juni 2024
Diterima: 28 Juni 2024

Intisari: Penelitian ini menganalisis sentimen masyarakat terhadap naturalisasi pemain sepak bola menggunakan metode K-Nearest Neighbor (KNN) dan Synthetic Minority Oversampling Technique (SMOTE). KNN digunakan untuk klasifikasi sentimen, sedangkan SMOTE menangani ketidakseimbangan kelas dalam dataset. Metodologi mencakup pengumpulan data, pelabelan, pembersihan, preprocessing, klasifikasi, dan evaluasi model dengan Google Colab dan Python. Hasil menunjukkan bahwa tanpa SMOTE, performa model lebih baik dengan presisi, recall, F1 score, dan akurasi tinggi. Sebaliknya, penggunaan SMOTE menurunkan performa, terutama dalam presisi dan F1 score. Model "Manhattan Neighbor 7" dan "Manhattan Neighbor 3" tanpa SMOTE menunjukkan hasil hampir sempurna, sementara SMOTE menurunkan beberapa metrik evaluasi secara signifikan. Selain itu, analisis opini masyarakat di YouTube menunjukkan kecenderungan sentimen negatif terhadap podcast tentang naturalisasi pemain sepak bola yang dipandu oleh Bung Towel dan Anjas Asmara, yang mencerminkan persepsi publik yang kritis dan skeptis terhadap topik tersebut. Penelitian ini memberikan wawasan penting tentang sentimen masyarakat dan efektivitas metode klasifikasi dalam konteks isu olahraga nasional.

Kata Kunci: K-Nearest Neighbor; Naturalisasi; Synthetic Minority Oversampling Technique

1. Pendahuluan

Sepakbola merupakan permainan yang memerlukan teknik, kekompakan dan kerja sama tim [1]. Pergerakan pemain dalam pemainan yang sangat presisi, baik dengan bola maupun tanpa bola sangat tepat dan dengan berlari sangat cepat dan dengan berlari mencari celah-celah di area yang dapat ditembus untuk memasukkan bola ke dalam gawang. Keadaan ini berlangsung dalam waktu yang cukup lama, sehingga begitu menguras energi dan berujung pada kelelahan. Dengan demikian, seorang pemain sepakbola harus melakukan latihan fisik dengan baik untuk menunjang kemampuan kondisi tubuhnya selama pertandingan berlangsung [2], karena persiapan ini sangat penting untuk menjaga performa optimal sepanjang permainan. Biasanya latihan tersebut mencakup berbagai aspek fisik dan kebugaran.

Mengklasifikasikan opini masyarakat mengenai naturalisasi pemain sepakbola dapat memberikan wawasan yang berharga bagi pengambil keputusan di bidang olahraga, khususnya dalam menentukan kebijakan terkait. Naturalisasi adalah proses yang dilakukan oleh warga asing agar menjadi warga negara indonesia secara sah dan utuh. Mengenai perpindahan status kewarganegaraan, setiap negara memiliki kebijakan-kebijakan yang berbeda sesuai dengan kepentingan dan tujuan suatu negara tersebut [3].

Metode yang efektif untuk klasifikasi data adalah algoritma K-Nearest Neighbors (KNN) dan SMOTE. Metode KNN memiliki beberapa keunggulan, yaitu pelatihan yang sederhana, cepat, mudah dimengerti, dan efektif apabila ukuran data pelatihan besar [4]. Namun, KNN ini juga terdapat kelemahan, yaitu nilai k bias [5]. Menurut penelitian yang dilakukan oleh [6] pada tahun 2023, tentang Implementasi Algoritma KNN untuk Klasifikasi Penerimaan Program

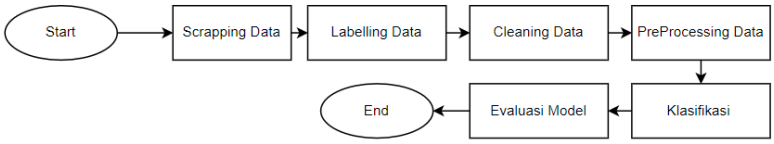
Beasiswa Program Indonesia Pintar, hasil dari penelitian ini adalah evaluasi algoritma KNN dengan menggunakan Confusion Matrix menunjukkan hasil berupa nilai rata - rata akurasi sebesar 77.06%.

Metode Synthetic Minority Over-sampling Technique (SMOTE) merupakan salah satu algoritma yang menangani imbalanced data [7]. SMOTE juga termasuk metode yang populer diterapkan dalam rangka menangani ketidak seimbangan kelas. Teknik ini mensintesis sampel baru dari kelas minoritas untuk menyeimbangkan dataset dengan cara sampling ulang sampel kelas minoritas. SMOTE juga adalah salah satu turunan dari oversampling.

Dengan demikian, penelitian ini bertujuan untuk menganalisis sentimen opini masyarakat mengenai naturalisasi pemain sepakbola menggunakan metode KNN dan SMOTE. Melalui penelitian ini, diharapkan dapat memperoleh pemahaman yang lebih mendalam mengenai distribusi sentimen di kalangan masyarakat. Hasil analisis ini tidak hanya akan memberikan gambaran mengenai persepsi umum publik terhadap naturalisasi pemain sepakbola, tetapi juga menawarkan wawasan yang berguna bagi pembuat kebijakan dan pengambil keputusan dalam mengelola kebijakan naturalisasi di bidang olahraga.

2. Metode

Penelitian ini menggunakan Google Colab dan untuk bahasa pemrogramannya menggunakan Python. Metode yang digunakan yaitu K-Nearest Neighbor dan SMOTE untuk klasifikasi sentimen. Dimana dalam penelitian ini terdapat beberapa tahap yaitu scrapping data, labeling data, cleaning data, preprocessing data, klasifikasi, dan evaluasi model. Berikut merupakan flowchart dari penelitian yang dilakukan.



Gambar 1. Flowchart Tahapan Penelitian

2.1 Scrapping Data

Scraping data adalah proses ekstraksi informasi dari suatu sumber, seperti situs web, dengan menggunakan program komputer atau skrip. Tujuannya adalah untuk mengambil data yang spesifik dari halaman web dan menyimpannya dalam format yang dapat diakses dan diolah oleh komputer.

Penelitian ini menggunakan data yang diambil dari platform YouTube yang tersedia secara publik. Video yang menjadi fokus penelitian memiliki judul "PEDAS! Bung Towel & Anjas Asmara Kritik STY: Naturalisasi Jangan Asal Comot! Bagian 02". Dalam video ini, diskusi yang hangat terjadi antara Bung Towel dan Anjas Asmara membahas topik sensitif mengenai kebijakan naturalisasi pemain sepakbola. Meskipun topiknya serius, jumlah komentar yang tercatat mencapai 3.264, menunjukkan minat dan partisipasi yang signifikan dari pengguna YouTube terhadap isu ini.

2.2 Labeling Data

Labeling data adalah proses menambahkan label atau kategori ke setiap data dalam dataset. Tujuan dari labeling data adalah untuk memberikan identifikasi atau penandaan yang jelas terhadap setiap data, sehingga data tersebut dapat digunakan untuk melatih model pembelajaran mesin yang memiliki tujuan tertentu, seperti klasifikasi atau prediksi. Misalnya, jika Anda memiliki dataset komentar YouTube tentang topik tertentu, Anda mungkin ingin memberi label "positif", "negatif", atau "netral" kepada setiap komentar berdasarkan sentimen yang terkandung di dalamnya. Proses ini akan melibatkan membaca dan menganalisis setiap komentar, kemudian menetapkan label yang sesuai berdasarkan analisis tersebut. Berikut merupakan contoh pelabelan sentimen:

Tabel 1. Tabel Pemberian Label Sentimen pada Komentar

Description	Sentiment
Towel itu mafia dy ga suka timnas maju karna dy ga dpt cuan	negatif
Mntap bung kami tnggu aksi apikmu	positif
Si towel ini mewakili siapa	netral

2.3 Cleaning Data

Cleaning data merupakan proses penting dalam pre-processing data yang bertujuan untuk membersihkan, menormalisasi, dan mempersiapkan data agar dapat digunakan untuk analisis. Pembersihan data ini bertujuan untuk meningkatkan kualitas data, dan meningkatkan kinerja model.

2.4 Preprocessing Data

Preprocessing data adalah teknik awal data mining untuk mengubah data mentah atau bisa dikenal dengan raw-data yang dikumpulkan dari berbagai sumber menjadi informasi yang bersih dan dapat digunakan pada tahap selanjutnya .

2.4.1 CaseFold

Casefold adalah proses mengubah semua huruf dalam suatu dokumen atau kalimat menjadi huruf kecil [8], tetapi lebih kuat dan lebih komprehensif. Ini adalah teknik yang digunakan untuk mengubah teks menjadi format yang seragam dalam hal perbandingan kasus (case-insensitive).

2.4.2 Tokenized

Tokenisasi adalah proses membagi teks kalimat atau paragraf menjadi bagian-bagian tertentu[9]. Token bisa berupa kata-kata, frasa, simbol, atau entitas lain yang relevan dalam konteks analisis teks. Tujuan dari tokenisasi adalah untuk mengubah teks mentah menjadi representasi yang lebih terstruktur dan dapat diproses lebih lanjut lagi.

2.4.3 Stopword

Penghapusan stopword adalah proses mengidentifikasi dan menghapus kata-kata umum dan kata yang sering muncul dalam teks namun cenderung tidak memberikan banyak informasi penting atau relevan terhadap konten teks tersebut. Stopword juga mengurangi indeks dari teks dengan menghapus beberapa kata kerja, kata sifat, dan kata keterangan lainnya[9]. Contohnya seperti di, ke, dari, atau, yang dan lain lain

2.4.4 Stemming

Stemming adalah metode yang dapat menghilangkan kata dengan porter guna meniadakan suffix dan order prefix dari suatu kata menjadi tidak bermunculan [10]. Sebagai contohnya seperti kata “berenang” diubah menjadi renang.

2.5 Evaluasi Model

Evaluasi model dilakukan untuk menguji hasil persiapan dengan memperkirakan nilai pada sebuah sistem. Batas yang digunakan untuk mengukur apresiasi realitas adalah akurasi. Akurasi sendiri adalah tingkat atau hasil laporan yang telah disiapkan secara efektif oleh system[11]. Berikut ini persamaan yang digunakan dalam *confusion matrix.*

$$Accuracy = \frac{(TP+TN)}{(TP+FP+TN+FN)} \quad (1)$$

$$Precision = \frac{TP}{(TP+FP)} \quad (2)$$

$$Recall = \frac{TP}{(TP+FN)} \quad (3)$$

Keterangan:

TP = True Positif

TN =True Negatif

FP = False Positif

FN = False Negatif

2.6 Pembobotan TF-IDF

TF-IDF (Term Frequency-inverse Document Frequency) adalah suatu proses pembobotan setiap kata agar bisa mengoptimalkan kemampuan analisis sentimen pada proses text mining[12]. Pembobotan ini menggunakan TfidfVectorizer dari library Sklearn.

Terma frekuensi (TF) adalah jumlah kata yang muncul dalam suatu dokumen, yang dihitung dengan rumus sebagai berikut:

$$TF(D, DF) = \frac{D}{DF} \quad (4)$$

Inverse document Frequency (IDF) digunakan untuk menghitung banyak dokumen yang mengandung term; untuk mencegah IDF bernilai 0 digunakan penambahan angka 1.IDF menggunakan rumus sebagai berikut:

$$IDF\left(\frac{D}{DF}\right) = \log_{10}(1 + \frac{D}{DF}) \quad (5)$$

Nilai TDF-IDF didapatkan dengan mengalikan nilai TF dan IDF. TF-IDF menggunakan rumus sebagai berikut:

$$TF - IDF = TF(D, DF) * IDF(\frac{D}{DF}) \quad (6)$$

Keterangan:

D = jumlah kalimat

DF = jumlah kalimat yang mengandung kata F

4. Hasil dan Pembahasan Penelitian

Untuk tahap klasifikasi, penelitian dilakukan dengan teknik *cross-validations* untuk menguji model *KNN* dengan parameter k yang berbeda, yaitu k = 3, k = 5, k = 7. Berikut merupakan tahap pengujian parameter k = 3, k = 5 dan k = 7. Berikut adalah perbandingan tabel yang berisi *accuracy, precision, recall, F1-Sco*re sebelum dan setelah di *SMOTE.*

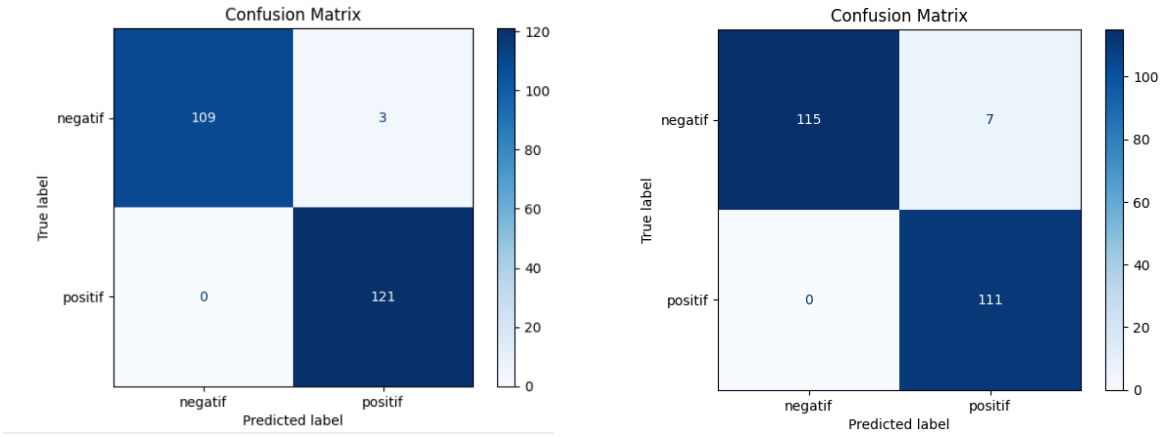
Tabel 2. Accuracy, Precision, Recall, dan F1-score pada Manhattan Neighbor sebelum SMOTE

	K = 3		**K = 5**		**K = 7**	
	Positif	Negatif	Positif	Negatif	Positif	Negatif
Accuracy	0.98		0.85		0.87	
Precision	0.97	1.00	0.78	1.00	0.77	1.00
Recall	1.00	0.97	1.00	0.71	1.00	0.73
F1-Score	0.98	0.98	0.85	0.83	0.87	0.84

Tabel 3. Accuracy, Precision, Recall, dan F1-score pada Manhattan Neighbor setelah SMOTE

	K = 3		**K = 5**		**K = 7**	
	Positif	Negatif	Positif	Negatif	Positif	Negatif
Accuracy	0.96		0.80		0.87	
Precision	1.00	0.94	1.0	0.71	0.79	1.00
Recall	0.94	1.00	0.62	1.00	1.00	0.77
F1-Score	0.97	0.96	0.76	0.83	0.88	0.87

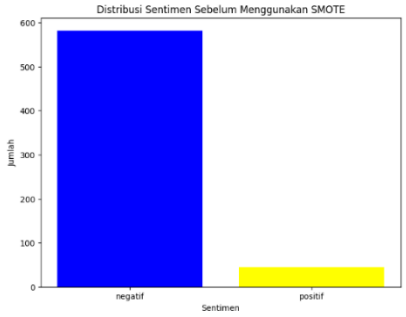
Jika merujuk pada tabel yang diberikan sebelumnya, dapat diperhatikan bahwa tingkat akurasi tertinggi tercapai ketika menggunakan parameter k=3 pada model. Confusion matrix yang berkaitan dengan tingkat akurasi ini menunjukkan hasil yang dihasilkan mencapai 98%, dimana akurasi sebesar 98% berada pada tabel yang belum dilakukan SMOTE. Berikut merupakan perbandingan antara confusion matrix sebelum dilakukan SMOTE dan setelah dilakukan SMOTE.



(a) (b)

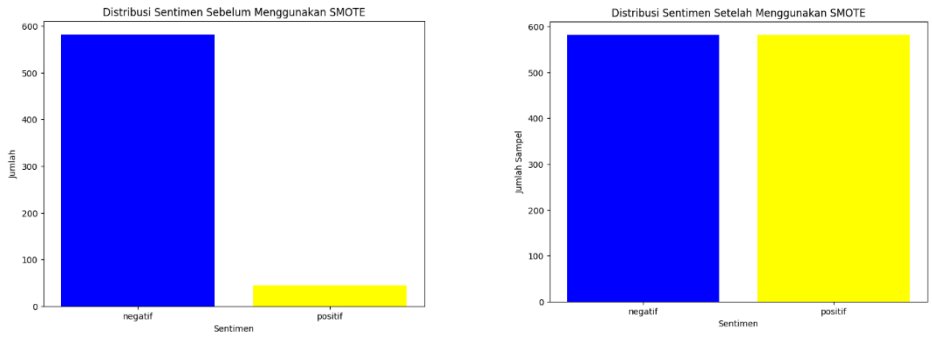
Gambar 2. Confusion Matrix
(**a**) Confusion Matrix KNN Sebelum SMOTE; (**b**) Confusion Matrix KNN Setelah SMOTE

Berikut merupakan hasil visualisasi jumlah sentimen positif dan negatif dimana jumlah data yang digunakan adalah sebanyak 627 setelah proses preprocessing. Dari data tersebut, terdapat 45 sentimen positif dan 582 sentimen negatif. Berikut merupakan visualisasi jumlah sentimen positif dan negatif sebelum di SMOTE:



Gambar 3. Visualisasi Jumlah Sentimen Positif dan Negatif

Setelah selesai menguji model K-Nearest Neighbors (KNN), tahap selanjutnya melibatkan penerapan teknik Synthetic Minority Over-sampling Technique (SMOTE) dalam visualisasi data sentimen awal yang telah dilakukan sebelumnya.



(a) (b)

Gambar 4. Visualisasi Sentimen
(**a**) Visualisasi Sentimen Positif dan Negatif Sebelum SMOTE; (**b**) Visualisasi Sentimen Positif dan Negatif Setelah SMOTE

Gambar 5 menunjukkan visualisasi sentimen menggunakan *library wordcloud.* Pada gambar dibawah merepresentasi visual dari frekuensi kata dalam sebuah teks, di mana kata-kata yang paling sering muncul akan ditampilkan dengan ukuran yang lebih besar.

Gambar 5. Visualisasi Sentimen dalam library wordcloud

5. Kesimpulan

Berdasarkan data yang disajikan, dapat disimpulkan bahwa tanpa menggunakan SMOTE, performa model cenderung lebih baik dibandingkan dengan menggunakan SMOTE. Hal ini terlihat dari beberapa metrik evaluasi seperti presisi, recall, F1 score, dan akurasi.

Untuk kelas "Manhattan Neighbor 7", model tanpa SMOTE memiliki presisi, recall, F1 score, dan akurasi yang cukup tinggi, dengan nilai yang stabil di atas 0.87. Sementara itu, ketika SMOTE diterapkan, terjadi penurunan performa terutama dalam hal presisi dan F1 score. Hal serupa terjadi pada kelas "Manhattan Neighbor 3", di mana model tanpa SMOTE memiliki nilai presisi, recall, F1 score, dan akurasi yang hampir sempurna. Namun, ketika SMOTE diterapkan, meskipun performa masih tinggi, terjadi penurunan kecil dalam beberapa metrik evaluasi.

Oleh karena itu, dapat disimpulkan bahwa dalam kasus ini, penggunaan SMOTE tidak menghasilkan peningkatan signifikan dalam performa model, bahkan dapat menyebabkan penurunan performa terutama pada kelas minoritas dna jika kita mengamati komentar dan opini masyarakat di platform YouTube, terlihat bahwa masyarakat cenderung menunjukkan sentimen negatif terhadap podcast tentang naturalisasi pemain sepakbola yang dipandu oleh Bung Towel dan Anjas Asmara

Daftar Pustaka

[1] H. . Pratama, S. Sulendro, and G. . Prasetyo, "Pengaruh Latihan Tingkat Keterampilan Teknik Dasar Menggiring Bola Dalam Permainan Sepakbola Peserta Putra Ekstrakurikuler SMPN 1 Gandusari," *J. Phys. Act.*, vol. 3, no. 1, pp. 1–9, 2022, doi: 10.58343/jpa.v3i1.28.

[2] A. Hidayat, I. Imanudin, and S. Ugelta, "Analisa Kebutuhan Latihan Fisik Pemain Sepakbola Dalam Kompetisi AFF U-19 (Studi Analisis Terhadap Pemain Gelandang Timnas Indonesia U-19)," *J. Terap. Ilmu Keolahragaan*, vol. 4, no. 1, pp. 1–4, 2019.

[3] G. K. Annas and N. M. Hazzar, "ANALISIS PERSAMAAN HAK KEWARGANEGARAAN BAGI PEMAIN NATURALISASI SEPAKBOLA DI INDONESIA," *J. Wicarana*, vol. 2, no. 2, pp. 127–143, 2024, [Online]. Available: https://www.ejournal-kumhamdiy.com/wicarana/article/view/37/29

[4] H. Dhery, A. Assyam, and F. N. Hasan, "Analisis Sentimen Twitter Terhadap Perpindahan Ibu Kota Negara Ke IKN Nusantara Menggunakan Orange Data Mining," *KLIK Kaji. Ilm. Inform. dan Komput.*, vol. 4, no. 1, pp. 341–349, 2023, doi: 10.30865/klik.v4i1.957.

[5] D. Cahyanti, A. Rahmayani, and S. A. Husniar, "Analisis performa metode Knn pada Dataset pasien pengidap Kanker Payudara," *Indones. J. Data Sci.*, vol. 1, no. 2, pp. 39–43, 2020, doi: 10.33096/ijodas.v1i2.13.

[6] C. Yanasari and T. Arifin, "Implementasi Algoritma K-Nearest Neighbor Untuk Klasifikasi Penerimaan Beasiswa Program Indonesia Pintar," *J. Sist. Inf. dan Ilmu Komput.*, vol. 1, no. 4, pp. 178–194, 2023, [Online]. Available: https://doi.org/10.59581/jusiik-widyakarya.v1i4.1862

[7] R. A. Nurdian, Mujib Ridwan, and Ahmad Yusuf, "Komparasi Metode SMOTE dan ADASYN dalam Meningkatkan Performa Klasifikasi Herregistrasi Mahasiswa Baru," *J. Tek. Inform. dan Sist. Inf.*, vol. 8, no. 1, pp. 24–32, 2022, doi: 10.28932/jutisi.v8i1.4004.

[8] E. E. P. Billy Gunawan, Helen Sasty Pratiwi, "Pengembangan Analisis Sentimen dalam Rekayasa Software

Engineering menggunakan tinjauan literatur sistematis," *J. MENTARI Manajemen, Pendidik. dan Teknol. Inf.*, vol. 2, no. 1, pp. 95–103, 2023, doi: 10.33050/mentari.v2i1.377.

[9] Syahril Dwi Prasetyo, Shofa Shofiah Hilabi, and Fitri Nurapriani, "Analisis Sentimen Relokasi Ibukota Nusantara Menggunakan Algoritma Naïve Bayes dan KNN," *J. KomtekInfo*, vol. 10, pp. 1–7, 2023, doi: 10.35134/komtekinfo.v10i1.330.

[10] ichsan nur irmasnyah Nurul chafid, luqman mujianto, "Penerapan Filter Kata Menggunakan Metode Stemming Pada Aplikasi Chatting Berbasis Web," vol. 1, no. 1, pp. 1–9, 2020.

[11] M. I. W. Slamet Harry Ramadhani, "Analisis Sentimen Terhadap Vaksinasi Astra Zeneca pada Twitter Menggunakan Metode Naïve Bayes dan K-NN," *J. Teknol. Inf. dan Komun.*, p. 530, 2022, [Online]. Available: https://journal.lembagakita.org/index.php/jtik/article/view/530

[12] O. I. Gifari, M. Adha, F. Freddy, and F. F. S. Durrand, "Analisis Sentimen Review Film Menggunakan TF-IDF dan Support Vector Machine," *J. Inf. Technol.*, vol. 2, no. 1, pp. 36–40, 2022, doi: 10.46229/jifotech.v2i1.330.