Analisis Sentimen Komentar Berdasarkan Geo Tagged Menggunakan Algoritma Naïve Bayes

1st Nurrafi Bagus Pratama
Fakultas Teknik Elektro
Universitas Telkom
Bandung, Indonesia
nurrafibp@student.telkomuniversity.ac.id

2nd Casi Setianingsih
Fakultas Teknik Elektro
Universitas Telkom
Bandung, Indonesia
setiacasie@telkomuniversity.co.id

3rd M. Faris Ruriawan
Fakultas Teknik Elektro
Universitas Telkom
Bandung, Indonesia
muhammadfaris@telkomuniversity.ac.id

***Abstrak* — Perkembangan teknologi dan media sosial kini pesat, khususnya Instagram. Banyak tokoh seperti presiden, menteri, artis, dan masyarakat umum menggunakan platform ini. Di Instagram, pengguna dapat berbagi gambar, video, pesan, dan menandai lokasi. Setiap postingan memiliki kolom komentar dengan beragam tanggapan, baik positif maupun negatif. Sentimen ini penting dalam menilai daya tarik objek wisata untuk masyarakat. Machine learning saat ini mampu otomatis mengklasifikasikan komentar sebagai positif atau negatif. Algoritma yang dipakai adalah Naïve Bayes, yang menggunakan probabilitas sederhana. Caranya, masukkan komentar, lalu gunakan Naive Bayes untuk kategorisasi dan menampilkan hasil sentimen. Penelitian ini bertujuan untuk menilai persentase komentar terhadap objek wisata dalam kategori positif dan negatif. Model sistem ini menggunakan rasio data latih dan tes terbaik, yaitu 80% dan 20%. Pengujian sistem dengan model tersebut menghasilkan presisi 87.72%, recall 89.27%, f1-score 87.60%, dan akurasi 87.72%. Hasil klasifikasi ini diharapkan menjadi panduan bagi masyarakat dalam kunjungan wisata.**

***Kata Kunci* — Instagram, Komentar Sentimen, Naïve Bayes, TF-IDF**

I. PENDAHULUAN

Media sosial adalah medium yang sangat efektif untuk berbagi cerita dalam bentuk gambar dan tulisan. Dalam perkembangannya, puluhan media sosial telah muncul di seluruh negeri. Lebih dari setengah populasi Indonesia menggunakan media sosial, menjadikannya negara dengan tingkat penggunaan tertinggi[1]. Saat ini, media sosial menjadi penting dalam menyebarkan berita dan juga memungkinkan setiap individu berbagi lokasi tempat wisata. Instagram, populer di kalangan masyarakat, terutama remaja, menawarkan fitur dan tampilan menarik. Pengguna tidak hanya berbagi gambar, tetapi juga menambah deskripsi. Tagging lokasi memungkinkan orang lain mengetahui lokasi. Komentar positif/negatif juga diberikan.

Penelitian lain telah mengklasifikasikan daya tarik turis dengan algoritma X-means di Flickr[2]. Tugas Akhir ini fokus pada menganalisis sentimen komentar Instagram dengan algoritma Naive Bayes. Tujuannya menghitung persentase komentar positif/negatif terhadap objek wisata. Hasilnya menentukan daya tarik tempat tersebut bagi turis.

II. KAJIAN TEORI

A. *Geo Tagged*

Geo Tagged merupakan fitur yang berbasis GPS di media sosial, umumnya dipakai untuk info lokasi posting. Setiap perangkat mobile memiliki teknologi GPS, termasuk Smartphone yang populer karena kemudahannya. Fitur ini digunakan menganalisis komentar geo tagged, mengungkap perilaku dan preferensi pengguna media sosial dalam menilai daya tarik tempat wisata[3]. Pada penelitian lain, *Geo Tagged* digunakan untuk mengukur keakuratan estimasi lintasan perjalanan waktu seseorang. *Geo Tagged* terdiri dari koordinat geografi yang menunjukkan lokasi postingan diunggah[4].

B. Analisa Sentimen

Sentimen saat ini memiliki 3 nilai positif, negatif, dan netral[5]. Ini disebabkan oleh interaksi melalui media sosial. Analisa sentimen di media sosial dapat mengungkap perilaku dan preferensi masyarakat suatu daerah[6]. Komentar sentimen merupakan analisa perasaan orang di media sosial. Hasilnya membantu klasifikasi daya tarik wisata berdasarkan sentimen positif, negatif, dan netral.

C. *Preprocessing*

Preprocessing adalah tahap awal dalam proses analisa teks untuk membuat sebuah data yang nanti akan diproses oleh sistem. Pembuatan data ini berfungsi untuk membuat algoritma mesin mengenali teks menjadi data numerik. Teks yang sudah dikumpulkan harus dipisahkan hal ini dapat

memiliki beberapa tingkatan. Tingkatan pada preprocessing yaitu bab, sub-bab, paragraf, kalimat, kalimat, dan pecahan kata[7].

D. TF-IDF

TF-IDF merupakan metode yang memberikan bobot terhadap kata (*term*) berdasarkan hubungannya dengan dokumen. Metode ini menggabungkan dua konsep pembobotan, yakni frekuensi kata (*term frequency*) dan frekuensi dokumen terbalik (*inverse document frequency*). TF-IDF efisien dalam memberikan bobot kata dengan menghitung frekuensi kemunculan suatu kata dalam dokumen tertentu dan frekuensi inverse dokumen yang mengandung kata tersebut[8]..

E. TF (*Term Frequency*)

Perhitungan *Term Frequency* adalah jumlah kata yang ingin dinilai dari suatu dokumen. Penjelasan pada rumus TF dibawah untuk 49 adalah jumlah kata yang ingin dihitung dalam suatu dokumen, lalu untuk 1739 merupakan jumlah kata yang terdapat pada satu dokumen sesuai dengan rumus dari[9]

$TF = 49 \div 1739 = 0.028$

F. IDF

Perhitungan *Inverse Document Frequency* adalah jumlah kata yang ingin dinilai dari seluruh dokumen yang dimiliki. Penjelasan pada rumus IDF diatas untuk 8 adalah jumlah dokumen total yang dimiliki, lalu untuk 49 merupakan jumlah kata yang ingin dihitung dalam suatu dokumen

$IDF = \log(8 \div 49) = 0.787$

G. SMOTE

SMOTE berfungsi ntuk mengatasi masalah data yang tidak seimbang, teknik SMOTE (*Synthetic Minority Oversampling Technique*) mengambil sampel dari kelompok minoritas kemudian duplikasinya untuk menghasilkan data yang seimbang di antara kelompok minoritas[10]. Menurut [11] proses SMOTE secara rinci dapat diurutkan sebagai berikut:

a. Setiap sample $x_i \in S_{min}$, Lakukan perhitungan k terdekat dengan sampel dari kelas minoritas menggunakan metode jarak Euclidean.
b. Pilih nilai x_j acak dari k terdekat dari x_i.
c. Sampel baru dihasilkan dari x_i dan x_j berdasarkan rumus:

$$x_{new} = x_i + |x_i \times x_j| \times \delta$$

H. *Naïve Bayes*

Algoritma *Naïve bayes* adalah algoritma pengklasifikasi menggunakan probabilitas sederhana yang dihitung dari satu set probabilitas dengan menghitung kombinasi dan frekuensi nilai dalam suatu kumpulan data[12]. Algoritma Naïve bayes ini menggunakan Teorema bayes untuk mengasumsikan bahwa variabel memiliki pertimbangan *independent* antara variabel.[12]. Teorema bayes merupakan rumus matematika yang digunakan untuk menghitung probabilitas bersyarat, yang ditemukan pada abad ke 18 oleh matematikawan di inggris yaitu Thomas Bayes. Persamaan dari teorema Bayes dapat dijelaskan sebagai berikut:

$$P(A|X) = \frac{p(X|A).p(A)}{p(X)}$$

Keterangan:

X : Data dengan kategori kelas yang belum diketahui

A : Hipotesis bahwa data X merupakan kelas yang spesifik

P(A|X) : Kemungkinan hipotesis (A) dengan memperhatikan kondisi (X)

P(A) : Kemungkinan hipotesis (A)

P(X|A) : Kemungkinan X dengan memperhatikan kondisi hipotesis A

P(X) : Kemungkinan X

Dalam metode Naïve Bayes, langkah klasifikasi memerlukan panduan untuk menentukan kelas yang tepat bagi data yang sedang diidentifikasi. Oleh karena itu, rumus di atas diadaptasi menjadi seperti berikut:

$$P(C|H1 \dots Hn) = \frac{p(C)p(H1 \dots Hn|C)}{p(H1 \dots Hn)}$$

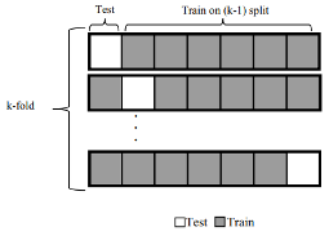
Keterangan:

C : *Class*

H1…Hn : menunjukan nilai klasifikasi

I. *K-Fold Cros Validation*

Langkah penting dalam meramalkan akurasi sampel baru ialah tahap klasifikasi. *K-Fold Cross Validation* mewakili pendekatan klasifikasi yang membantu dalam mengestimasi akurasi dari model pembelajaran mesin yang sudah di-training. Berbeda dari metode klasifikasi lain yang sekadar membagi data menjadi satu set latih dan satu set uji, *K-Fold Cross Validation* membagi data latih dan uji secara berulang, menjadikan seluruh dataset terlibat sebagai data latih dan uji secara bergantian. [13].



GAMBAR 1
Penggambaran *K-Fold Cross Validation.*

Ketika digunakan, pengguna memilih jumlah k yang menunjukkan seberapa banyak lipatan data akan dibuat. Kemudian, salah satu dari lipatan data akan dijadikan sebagai data uji, sementara lipatan-lipatan data lainnya digunakan sebagai data latih. Langkah ini diulang sebanyak nilai k yang telah ditentukan, untuk mengujikan setiap lipatan data

sebagai data uji. Proses ini berlangsung berulang-ulang hingga setiap lipatan data telah diuji sebagai data uji[14]. Akurasi dari *K-Fold Cross Validation* dihitung dengan mengambil nilai rata-rata dari hasil pengujian pada setiap *fold*.

J. *Confusion Matrix*

Confusion matrix sering dimanfaatkan dalam pembelajaran mesin sebagai instrumen untuk mengukur prestasi model klasifikasi yang dipakai. Matriks konfusi diilustrasikan dalam bentuk tabel yang merepresentasikan nilai dari empat parameter yang ada di dalamnya, yakni *true positive* (TP), *true negative* (TN), *false positive* (FP), dan *false negative* (FN). **[15]**.

TABEL 1
Tabel representasi nilai parameter *confusion matrix*

		Aktual	
		True	False
Prediksi	True	TP Hasil Tepat	FP Hasil Salah
	False	FN Hasil Salah	TN Hasil Tepat

True positive (TP) adalah data positif yang berhasil diidentifikasi dengan benar sebagai data positif oleh model, sementara *false positive* (FP) mengacu pada data negatif yang keliru diidentifikasi oleh model sebagai data positif. Sebaliknya, *false negative* (FN) mencakup data positif yang tidak berhasil diidentifikasi oleh model dan secara keliru dianggap sebagai data negatif, sedangkan *true negative* (TN) merujuk pada data negatif yang berhasil diidentifikasi dengan benar sebagai data negatif oleh model. [16].

Kemudian, hasil dari confusion matrix dihitung untuk mengukur nilai accuracy, precision, recall, dan F1 Score. Menggunakan rumus dibawah ini:

Accuracy

$$accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

Precicsion

$$Precision = \frac{TP}{TP + FP}$$

Recall

$$Recall = \frac{TP}{TP + FN}$$

F1 Score

$$F1\ Score = 2 * \frac{Precision\ *\ Recall}{Precision\ + Recall}$$

III. METODE

A. Desain Sistem

Pada analisis sentimen komentar ini, akan diterapkan metode algoritma *Naïve Bayes*. Serangkaian langkah kerja yang akan dijalankan oleh sistem termasuk langkah-langkah berikut: input data, tahap *preprocessing*, pembobotan kata, klasifikasi, dan hasil output.

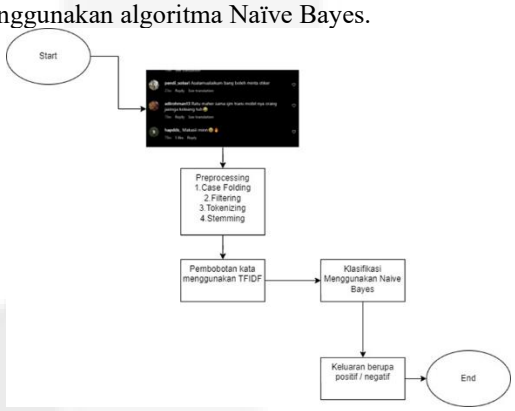
Langkah awal adalah menyediakan data yang akan digunakan, yang diambil dari komentar pada postingan Instagram. Data ini telah disaring berdasarkan tag lokasi, khususnya tempat wisata di kawasan Bandung dan sekitarnya.

Tahap berikutnya adalah *preprocessing*. Di dalam tahap ini, dilakukan penghapusan kata-kata yang tidak diperlukan dan berpotensi mengganggu proses klasifikasi dari sistem.

Setelah tahap *preprocessing*, dilakukan langkah pembobotan kata, juga dikenal sebagai ekstraksi fitur. Tujuan di balik langkah ini adalah mengubah kata-kata menjadi representasi vektor dengan menggunakan metode TF-IDF. Karena hakikatnya, kata-kata atau kalimat tidak memiliki struktur yang dapat dievaluasi secara langsung.

Kemudian, langkah klasifikasi dijalankan. Data yang telah melalui *preprocessing* akan dikelompokkan ke dalam dua kategori keluaran, yakni nilai positif dan negatif.

Inilah garis besar dari proses analisis sentimen yang akan diterapkan pada komentar tersebut menggunakan algoritma Naïve Bayes.

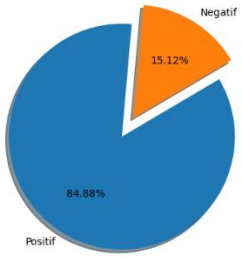


GAMBAR 3
Desain Sistem

B. Kebutuhan Data

Dataset pada sistem ini diambil dari komentar postingan Instagram yang mempunyai *tagging location* berasal dari tempat wisata Kota Bandung dan sekitarnya. Data yang berhasil terkumpul berjumlah 4014 komentar yang terbagi menjadi 2 label, yaitu label positif berjumlah 3407 dan label negatif berjumlah 607. Label pada setiap komentar diperoleh dari hasil evaluasi yang dilakukan oleh pakar bahasa di Balai Bahasa Universitas Pendidikan Indonesia. Setelah data ini telah dievaluasi, langkah selanjutnya adalah memasukkannya ke dalam proses berikutnya sebelum masuk ke dalam proses klasifikasi. Dari hasil *dataset* tersebut menghasilkan grafik lingkaran

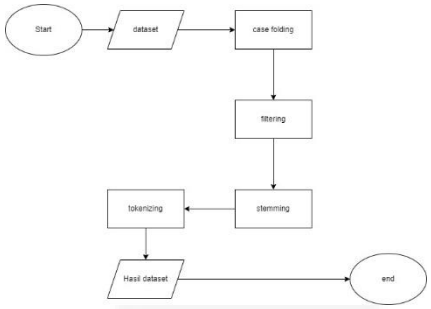
yang menggambarkan berapa persen komentar positif dan negatif pada **Gambar** dibawah ini.



GAMBAR 4
Grafik Lingkaran perbandingan dataset

C. *Preprocessing*

Tahap preprocessing berfungsi agar kalimat dan kata yang digunakan untuk proses klasifikasi tidak terganggu pada sistem tugas akhir ini. Dengan cara menghilangkan kata-kata yang tidak bersangkutan, emoji, huruf yang tidak lowercase. Proses preprocessing memiliki alur seperti Gambar dibawah ini.



GAMBAR 4
Tahap *preprocessing*

D. *Feature Extraction*

Tahap Ekstraksi Fitur merupakan langkah di mana kata-kata diberikan representasi vektor. Karena, kata-kata tidak memiliki nilai vektor kecuali jika proses ekstraksi fitur dilakukan. Dalam tahap ini, digunakan metode TF-IDF (*Term Frequency-Inverse Document Frequency*). Frekuensi Term akan menghitung seberapa sering kata tertentu muncul dalam dokumen atau kumpulan data. Setelah dilakukan Frekuensi Term, inverse Frekuensi Dokumen akan memberikan bobot pada kata-kata yang muncul dalam dokumen atau dataset khusus. Bobot yang diberikan oleh invers Frekuensi Dokumen mencerminkan sejauh mana kata tersebut umum atau jarang muncul dalam berbagai dokumen, sehingga akan membedakan antara kata-kata yang penting dan tidak penting. Setelah itu, sistem akan menghasilkan nilai untuk setiap kata yang terdapat dalam dokumen atau dataset khusus.

E. Klasifikasi *Naïve Bayes*

Lalu pada tahap klasifikasi Naïve Bayes ini, nilai pada tahap TFIDF akan diproses oleh proses klasifikasi Naïve Bayes. Selanjutnya diasumsikan akan diuji dengan data uji kata “keren”. Langkah-langkah perhitungan seperti bawah ini:

1. Prior

F(a)=Na/N

Diberikan informasi bahwa N mewakili total jumlah dokumen dan Na mengindikasikan total jumlah dokumen dengan label a. Ini dapat diterapkan dalam sebuah contoh kasus, menghasilkan nilai sebagai berikut:

F(p) = 1/8
F(n) = 1/8
Penjelasan:
p = positif
n = negatif

2. *Condtiional Probabalitties*

F(w|a) = (count(w,a)+1)/(count(a)+|V|)

Dari persamaan tersebut dapat diidentifikasi bahwa count(w,a) mencerminkan jumlah kata w dalam data yang memiliki label a. Sedangkan |v| merepresentasikan total jumlah kata dalam semua dokumen. Apabila diaplikasikan pada suatu contoh kasus, akan menghasilkan nilai sebagai berikut:

F(keren|p) = (1+1) / (49+1) = 2/50
F(keren|n) = (0+1) / (49+1) = 1/50

3. *Class Selection*

Pada class selection, nilai prior akan dikali dengan nilai conditional probabilities. Sehingga hasil perhitungan menjadi:

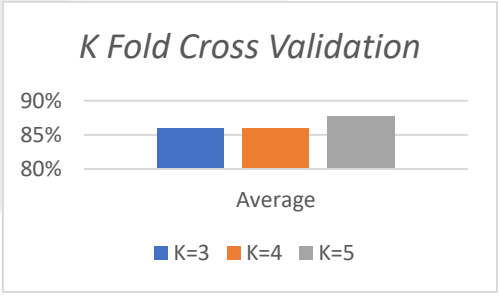
F(p|tebak) = F(p) × F(keren|p)
F(p|tebak) = 1/8 × 2/50 = 0.005
F(n|tebak) = F(n) × F(keren|n)
F(n|tebak) = 1/8 × 1/50 = 0.0025

Dari perhitungan di atas dengan klasifikasi Naïve Bayes, dapat disimpulkan bahwa data uji mendapat nilai lebih tinggi untuk kelas positif dibandingkan kelas negatif, maka label dari kata keren yaitu positif.

IV. HASIL DAN PEMBAHASAN

A. *K-Fold Cross Validation*

Pengujian K-Fold Cross Validation akan dilakukan dengan nilai K dari 3 sampai 5



GAMBAR 5
Hasil *K Fold Cross Validation*

Pada iterasi pertama mendapatkan nilai akurasi 87%, pada iterasi kedua mendapatkan nilai akurasi 80.51%, pada iterasi ketiga mendapatkan nilai akurasi 88.03%, pada iterasi keempat mendapatkan nilai akurasi 89.45%, dan pada iterasi kelima mendapatkan nilai akurasi 88.65%. Dari kelima

iterasi tersebut mendapatkan nilai rata-rata akurasi sebesar 87.72%, pada iterasi K=5 ini memiliki nilai akurasi terbesar dari seluruh iterasi yang dicoba. Maka untuk melakukan pengujian selanjutnya saya akan menggunakan K=5.

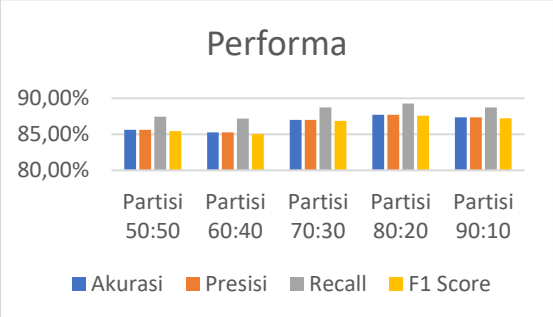
B. Pengujian Partisi Data

Pada pengujian partisi data ini dilakukan dengan menjalankan rasio pembagian dataset *training* dan *test* dari 50:50 hingga 90:10. Menghasilkan seperti Berikut:

TABEL 2
Confusion Matrix partisi data 50:50 hingga 90:10

Data Latih	Data Uji			Label Aktual	
				Positif	Negatif
50%	50%	Label Prediksi	Positif	1269	432
			Negatif	57	1644
60%	40%		Positif	1006	355
			Negatif	46	1315
70%	30%		Positif	781	240
			Negatif	25	996
80%	20%		Positif	529	151
			Negatif	16	665
90%	10%		Positif	265	75
			Negatif	11	330

C. Pengujian Performa



GAMBAR 6
Pengujian performa

Pada grafik diatas menggambarkan hasil pengujian performa dari sistem yang dibuat. Untuk partisi 60:40 memiliki nilai paling kecil dalam pengujian performa. Lalu pada partisi data 80:20 memiliki nilai paling baik dalam pengujian performa kali ini. Parameter yang digunakan pada pengujian performa kali ini adalah akurasi, presisi, *recall*, dan *f1 score*. Menghasilkan 87.72% akurasi, 87.72% presisi, 89.27% *recall*, dan *f1 score* 87.60%.

V. KESIMPULAN

Berdasarkan hasil implementasi sistem yang berhasil dikembangkan, serta evaluasi dan analisis yang dilakukan, dapat disimpulkan bahwa K-Fold dengan nilai K=5 mendapatkan akurasi terbaik pada sistem ini dengan nilai 87.72%, Precision 87.72%, Recall 89.27%, dan F1 Score 87.60%

REFERENSI

[1] Supratman, L. P. "Penggunaan Media Sosial oleh Digital Native". Jurnal Ilmu Komunikasi, vol. 15. 2018

[2] Lin J-Y, Wen S-M, Hirota M, Araki T, Ishikawa H. "Less-Known Tourist Attraction Discovery Based on Geo-Tagged Photographs," Machine Learning and Knowledge Extraction. 2020

[3] Kim, Dongeun, Youngok Kang, Yerim Park, Nayeon Kim, and Juyoon Lee. 2020. "Understanding Tourists' Urban Images with Geotagged Photos Using Convolutional Neural Networks." Spatial Information Research 28(2):241– 55. doi: 10.1007/s41324-019-00285-x.

[4] M. Parsafard, G. Chi, X. Qu, X. Li, and H. Wang, "Error Measures for Trajectory Estimations with Geo-Tagged Mobility Sample Data," IEEE Transactions on Intelligent Transportation Systems, vol. 20. 2019

[5] Ortigosa, Alvaro, José M. Martín, and Rosa M. Carro. 2014. "Sentiment Analysis in Facebook and Its Application to E-Learning." Computers in Human Behavior 31(1):527–41. doi: 10.1016/j.chb.2013.05.024.

[6] Gunawan, Billy, Helen Sasty, Pratiwi #2, Enda Esyudha, and Pratama #3. 2018. "JEPIN (Jurnal Edukasi Dan Penelitian Informatika) Sistem Analisis Sentimen Pada Ulasan Produk Menggunakan Metode Naive Bayes." 4(2):17–29.

[7] Feldman, Ronen, and James Sanger. 2007. The Text Mining Handbook : Advanced Approaches in Analyzing Unstructured Data. Cambridge University Press.

[8] Riyani, Ade, Muhammad Zidny Naf'an #2, and Auliya Burhanuddin. 2019. Penerapan Cosine Similarity Dan Pembobotan TF-IDF Untuk Mendeteksi Kemiripan Dokumen. Vol. 2.

[9] C. D. Manning, P. Raghavan, dan H. Schütze, An Introduction to Information Retrieval. Cambridge University Press, 2009.

[10] C. Meng, L. Zhou, and B. Liu, "A case study in credit fraud detection with SMOTE and

XGboost," Journal of Physics: Conference Series, vol. 1601. 2020

[11] Y. Yan, R. Liu, Z. Ding, X. Du, J. Chen, and Y. Zhang, "A parameter-free cleaning method for SMOTE in imbalanced classification," IEEE Access, vol. 7. 2019.

[12] Saritas, Mücahid Mustafa, and Ali Yasar. 2019. "International Journal of Intelligent Systems and Applications in Engineering Performance Analysis of ANN and Naive Bayes Classification Algorithm for Data Classification." Original Research Paper International Journal of Intelligent Systems and Applications in Engineering IJISAE 7(2):88–91. doi: 10.1039/b000000x.

[13] H. Tabrizchi, M. M. Javidi, and V. Amirzadeh, "Estimates of residential
building energy consumption using a multi-verse optimizer-based support
vector machine with k-fold cross-validation," Evol. Syst., vol. 12, no. 3, pp.
755–767, 2021, doi: 10.1007/s12530-019-09283-8.

[14] I. K. Nti, O. Nyarko-Boateng, and J. Aning, "Performance of Machine
Learning Algorithms with Different K Values in K-fold CrossValidation,"
Int. J. Inf. Technol. Comput. Sci., vol. 13, no. 6, pp. 61–71, 2021, doi:
10.5815/ijitcs.2021.06.05.

[15] M. Hasnain, M. F. Pasha, I. Ghani, M. Imran, M. Y. Alzahrani, and R.
Budiarto, "Evaluating Trust Prediction and Confusion Matrix Measures for
Web Services Ranking," IEEE Access, vol. 8, pp. 90847–90861, 2020, doi:
10.1109/ACCESS.2020.2994222.

[16] F. Rahmad, Y. Suryanto, and K. Ramli, "Performance Comparison of Anti-Spam Technology Using Confusion Matrix Classification," IOP Conf. Ser.
Mater. Sci. Eng., vol. 879, no. 1, 2020, doi: 10.1088/1757-
899X/879/1/012076.