KLASIFIKASI EMOSI PADA TWEET BERBAHASA INDONESIA MENGGUNAKAN TEKNIK SAMPLING ENN

I Gede Harsemadi[1)] I Komang Dharmendra[2)] I Made Pasek Pradnyana Wijaya[3)]
Program Studi Sistem Informasi[1)2)3)]
ITB STIKOM Bali[1)2)3)], Jalan Raya Puputan No 86 Renon Denpasar[1)2)3)]
harsemadi@stikom-bali.ac.id[1)], dharmendra@stikom-bali.ac.id[2)], pasek_pradnyana@stikom-bali.ac.id[3)]

ABSTRACT

Twitter serves as a primary focus due to its widespread popularity as a social media platform worldwide, with continually growing data. The potential for sentiment analysis and emotion detection from tweets is substantial. However, challenges arise due to the often imbalanced nature of tweet datasets across emotion classes. Therefore, this study examines the performance of four different classification algorithms using the Edited Nearest Neighbours (ENN) sampling technique on an imbalanced Indonesian tweet dataset with five emotion classes. The evaluation results reveal that the RandomForest model achieves the highest accuracy at approximately 62.55%, while the Neural Network excels in precision with a score of 67.23%. Although the SVM model exhibits high precision, low recall and F1-scores indicate limitations in correctly identifying positive classes. Thus, the use of ENN to address data imbalance in Indonesian tweet datasets demonstrates that the RandomForest and Neural Network models are better choices for the emotion classification task.

Keywords *Twitter, Emotion detection, Sampling technique, Classification algorithm*.

ABSTRAK

Twitter menjadi fokus utama sebagai platform media sosial yang sangat populer di seluruh dunia, dengan pertumbuhan data yang terus meningkat setiap hari. Potensinya untuk aplikasi analisis sentimen dan deteksi emosi dari tweet sangat besar. Namun, tantangan muncul karena dataset tweet sering kali tidak seimbang antara kelas emosi. Oleh karena itu, penelitian ini menguji empat algoritma klasifikasi yang berbeda dengan menggunakan teknik sampling Edited Nearest Neighbours (ENN) pada dataset tweet berbahasa Indonesia yang tidak seimbang dengan lima kelas emosi. Hasil evaluasi mengungkapkan bahwa model RandomForest memiliki akurasi tertinggi sekitar 62.55%, sementara Neural Network mendominasi dalam presisi dengan nilai 67.23%. Meskipun model SVM memiliki presisi yang tinggi, recall dan F1-score yang rendah menunjukkan keterbatasan dalam mengidentifikasi kelas positif dengan benar. Oleh karena itu, penggunaan ENN untuk mengatasi ketidakseimbangan data pada dataset tweet berbahasa Indonesia memperlihatkan bahwa model RandomForest dan Neural Network adalah pilihan yang lebih baik dalam tugas klasifikasi emosi.

Kata kunci: *Twitter, Deteksi emosi, Teknik sampling, Algoritma klasifikasi.,*

PENDAHULUAN

Twitter merupakan salah satu platform media sosial yang sangat populer dan sering digunakan oleh pengguna di seluruh dunia. Data dari Twitter terus bertambah setiap harinya dan terdapat potensi besar untuk memanfaatkannya dalam berbagai aplikasi. Salah satu aplikasi yang menarik adalah dalam analisis sentimen dan deteksi emosi dari pengguna Twitter. Namun, deteksi emosi pada data tweet dapat menjadi sulit karena masing-masing kelas memiliki jumlah data yang berbeda.

Untuk mengatasi masalah tersebut, banyak penelitian telah dilakukan untuk mengembangkan model klasifikasi emosi pada dataset tweet. Namun, sebagian besar penelitian tersebut belum memperhatikan ketidakseimbangan data dalam dataset tweet, yang menyebabkan deteksi emosi pada data tweet dapat menjadi sulit karena masing-masing kelas memiliki jumlah data yang berbeda. Oleh karena itu, klasifikasi emosi pada data tweet memerlukan teknik sampling yang tepat untuk mengatasi masalah ketidakseimbangan data pada setiap kelas dan bisa meningkatkan performa dari model klasifikasi emosi.

Oleh karena itu, penelitian ini bertujuan untuk menguji performa dari empat algoritma klasifikasi yang berbeda dengan menggunakan teknik sampling ENN pada dataset tweet berbahasa Indonesia dengan 5 kelas yang tidak seimbang. Ketiga algoritma klasifikasi yang digunakan adalah *Neural Network, Random Forest*, dan SVM karena keempat algoritma

tersebut telah terbukti efektif dalam penanganan masalah klasifikasi pada dataset tweet.

Diharapkan hasil dari penelitian ini dapat memberikan kontribusi dalam pengembangan teknik klasifikasi emosi pada data tweet yang tidak seimbang dan meningkatkan performa dari model klasifikasi emosi pada dataset tweet berbahasa Indonesia.

Kajian Pustaka

Emosi

Emosi adalah kompleksitas keadaan mental yang terjadi pada individu dan dipengaruhi oleh berbagai faktor seperti peristiwa eksternal, perubahan fisiologis, atau interaksi dengan orang lain. Emosi dapat menghasilkan respon yang beragam, seperti perubahan mood, perasaan, pikiran, dan perilaku. Dalam pandangan psikologis, emosi seringkali dianggap sebagai pengalaman subyektif yang melibatkan perasaan dan pikiran yang dirasakan oleh individu, serta memiliki keterkaitan dengan perasaan dan perilaku mereka[1].

Emosi terdiri dari dua dimensi utama, yaitu *arousal* dan *valensi*. Dimensi *arousal* merujuk pada sejauh mana emosi memengaruhi tingkat aktivitas fisiologis individu, seperti detak jantung, tingkat pernapasan, dan aktivitas otot. Sementara itu, dimensi *valensi* merujuk pada apakah emosi itu dianggap sebagai sesuatu yang positif atau negatif oleh individu[2].

Klasifikasi Emosi

Klasifikasi emosi adalah proses mengelompokkan emosi ke dalam kategori yang berbeda berdasarkan karakteristik dan dimensi yang dikenali oleh para ahli. Ada beberapa model klasifikasi emosi yang digunakan dalam psikologi dan neurosains, namun dua model yang paling umum digunakan adalah model *Rasch* dan model dimensional[3].

Model *Rasch*, yang diperkenalkan oleh psikolog Swiss, Max Rasch pada tahun 1960-an, mengelompokkan emosi ke dalam kategori yang diskrit, seperti kegembiraan, kecemasan, sedih, dan marah. Model Rasch memandang emosi sebagai kondisi yang bersifat stabilitas, yang dapat diukur dan diidentifikasi dengan tepat[4].

Twitter

Twitter adalah layanan jejaring sosial dan mikroblogging yang memungkinkan pengguna untuk mengirim dan membaca pesan berbasis teks hingga 280 karakter, yang dikenal sebagai "tweet". Layanan ini didirikan pada tahun 2006 dan telah menjadi salah satu platform media sosial paling populer di seluruh dunia, dengan jutaan pengguna aktif harian[5].

Twitter memungkinkan pengguna untuk membagikan informasi dan berkomunikasi dengan orang lain dengan cara yang cepat dan mudah. Pengguna dapat mengirim pesan singkat atau tweet, yang kemudian ditampilkan pada halaman profil mereka dan dapat diakses oleh pengikut mereka. Pengguna juga dapat mencari tweet yang berkaitan dengan topik atau hashtag tertentu, serta menandai atau membalas tweet orang lain.

Imbalance Dataset

Dataset yang tidak seimbang atau *imbalanced dataset* adalah kondisi di mana kelas-kelas dalam dataset memiliki jumlah data yang sangat tidak seimbang. Contohnya, jika dalam sebuah dataset terdapat 1000 data, dan hanya 10 data yang termasuk dalam kelas minoritas, maka dataset tersebut dapat dikategorikan sebagai dataset yang tidak seimbang[6].

Dalam praktiknya, dataset yang tidak seimbang seringkali ditemukan pada masalah klasifikasi[7], di mana kelas minoritas memiliki nilai yang lebih signifikan daripada kelas mayoritas.

Editing Nearest Neighbor (ENN)

Edited Nearest Neighbor (ENN) adalah salah satu metode *undersampling* yang digunakan untuk mengatasi dataset yang tidak seimbang atau *imbalanced dataset* pada masalah klasifikasi. Metode ini bekerja dengan menghapus sebagian data pada kelas mayoritas, sehingga jumlah data pada kelas minoritas dan mayoritas menjadi seimbang. Proses penghapusan data dilakukan dengan cara membandingkan setiap data pada kelas mayoritas dengan data pada kelas minoritas. Data yang memiliki label yang sama dengan tetangga terdekatnya dari kelas minoritas akan dihapus[8][9].

Metode ENN memiliki tiga varian, yaitu ENN1, ENN2, dan ENN3. Varian ENN1 bekerja dengan cara membandingkan setiap data pada kelas mayoritas dengan tetangga terdekatnya pada kelas minoritas. Data yang memiliki label yang sama dengan tetangga terdekatnya dari kelas minoritas akan dihapus. Varian ENN2 bekerja dengan cara membandingkan setiap data pada kelas mayoritas dengan dua tetangga terdekatnya pada kelas minoritas. Jika label dari dua tetangga terdekat berbeda, maka data akan dihapus. Varian ENN3 bekerja dengan cara membandingkan setiap data pada kelas mayoritas dengan tiga tetangga terdekatnya pada kelas minoritas. Data yang memiliki label yang

sama dengan tetangga terdekatnya dari kelas minoritas akan dihapus.

Neural Network

Neural network atau jaringan saraf tiruan adalah model matematis yang terinspirasi oleh cara kerja otak manusia dan terdiri dari kumpulan *neuron* yang saling terhubung. Tujuan dari neural network adalah untuk mempelajari pola dalam data dan membuat prediksi atau klasifikasi berdasarkan pola-pola tersebut[10].

Setiap *neuron* dalam jaringan saraf memiliki beberapa masukan, yang masing-masing memiliki bobot yang ditentukan secara acak pada awalnya. *Neuron* akan menghitung jumlah dari masukan-masukan yang diterima, kemudian menjalankan fungsi aktivasi untuk menghasilkan keluaran.

Random Forest

Random Forest adalah salah satu algoritma *ensemble learning* yang populer digunakan dalam *machine learning* untuk menangani masalah klasifikasi dan regresi. *Ensemble learning* sendiri merupakan teknik yang menggabungkan beberapa model *machine learning* ke dalam satu model yang lebih kompleks dan efektif dalam memprediksi target pada data baru[5].

Random Forest menerapkan konsep *bootstrap aggregating* atau biasa disebut *bagging*. *Bagging* memungkinkan *Random Forest* untuk menghasilkan model yang stabil dan akurat pada dataset yang besar atau noisy dengan membangun banyak pohon keputusan secara acak dan menggabungkan prediksi dari setiap pohon.

***Support Vector Machine* (SVM)**

Support Vector Machine (SVM) merupakan algoritma yang digunakan dalam *machine learning* untuk melakukan klasifikasi dan regresi. SVM adalah algoritma pembelajaran berbasis *kernel* yang mengambil dataset dan menghasilkan model yang optimal untuk memisahkan kelas data dengan margin terbesar. SVM dapat digunakan untuk dataset linier dan non-linier. Algoritma ini merupakan salah satu teknik pembelajaran mesin yang paling populer dan efektif dalam klasifikasi biner dan multi-kelas[11].

Konsep dasar dari SVM adalah memilih *hyperplane* (bidang pemisah) yang memiliki margin terbesar di antara kedua kelas. Margin adalah jarak antara *hyperplane* dan titik terdekat dari setiap kelas. SVM mencari hyperplane yang dapat memaksimalkan margin sehingga lebih mampu menggeneralisasi data yang belum pernah dilihat sebelumnya.

METODE PENELITIAN

Sumber Data

Penelitian akan menggunakan dataset twitter dari penelitian sebelumnya[12] yang terdiri dari 5 kelas emosi, yaitu : *anger* dengan 1101 data tweet, *happy* dengan 1017 data tweet, *sadness* dengan 997 data tweet, *fear* dengan 649 data tweet, *love* dengan 637 data tweet.

Alur Analisis

Alur analisis dari penelitian yang dilakukan adalah sebagai berikut:

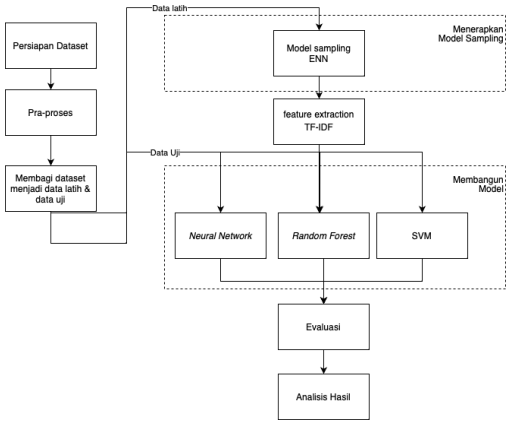
1. Mengumpulkan dataset
2. Melakukan praproses seperti menghilangkan karakter khusus, stopword removal, dan stemming
3. Melakukan eksplorasi data untuk memahami karakteristik dari dataset, termasuk jumlah data untuk setiap kelas emosi
4. Melakukan pengambilan sampel dengan teknik ENN untuk mengurangi ketidakseimbangan dataset.
5. Melakukan *feature extraction* menggunakan TF-IDF.
6. Train model klasifikasi *Neural Network*, *Random Forest*, dan *SVM* pada dataset train yang telah di-sampling dengan teknik ENN.

Evaluasi performa model pada dataset test menggunakan metrik evaluasi *seperti accuracy*, *precision*, *recall*, dan *f1-score*.

HASIL DAN PEMBAHASAN

Praproses

Proses praproses data teks adalah proses untuk membersihkan dan mempersiapkan data teks agar dapat digunakan untuk analisis data[13]. Proses ini biasanya dilakukan untuk meningkatkan akurasi dan kinerja model pembelajaran mesin.



Gambar 1. Alur Penelitian

Menghilangkan karakter khusus

Proses pertama yang dilakukan dalam praproses data teks adalah menghilangkan karakter khusus. Karakter khusus adalah karakter yang tidak memiliki arti seperti tanda baca, angka, simbol, dan sebagainya. Karakter-karakter ini dapat mengganggu proses analisis data karena dapat dianggap sebagai kata yang berbeda.

Stopword removal

Proses kedua yang dilakukan dalam praproses data teks adalah *stopword removal*. *Stopword* adalah kata-kata yang sering muncul dalam bahasa dan tidak memiliki arti penting. *Stopword removal* bertujuan untuk mengurangi ukuran data dan meningkatkan akurasi model pembelajaran mesin.

Stemming

Proses ketiga yang dilakukan dalam praproses data teks adalah *stemming*. *Stemming* adalah proses untuk merubah kata ke bentuk dasarnya. *Stemming* dapat membantu dalam proses analisis data karena dapat mengurangi jumlah kata yang berbeda.

Metode sampling dengan *ENN*

Dataset tweet yang tidak seimbang dapat menyebabkan bias pada model klasifikasi emosi. Hal ini karena model akan lebih cenderung untuk memprediksi kelas mayoritas.

Pengurangan sampel adalah teknik untuk mengurangi ukuran dataset dengan menghapus data dari kelas mayoritas. Teknik ini dapat membantu untuk mengurangi bias pada model klasifikasi. *Edited Nearest Neighbours* (ENN) adalah salah satu teknik pengurangan sampel yang populer. Teknik ini bekerja dengan menghapus data dari kelas mayoritas jika data tersebut memiliki tetangga terdekat dari kelas minoritas. Tabel 1 menunjukkan perubahan jumlah data sebelum menggunakan ENN dan setelah menggunakan ENN

Table 1 Perubahan Jumlah Data

		Jumlah Data
Sebelum dengan ENN	*Resampling*	4401
Setelah dengan ENN	*Resampling*	3484

Membagi Dataset

Setelah data diolah, langkah berikutnya adalah membagi data menjadi data latih dan data uji. Data latih digunakan untuk melatih model, sementara data uji digunakan untuk mengukur kinerja model yang telah dilatih.

Ekstraksi Fitur dengan TF-IDF

Kemudian, dilakukan penghitungan bobot untuk menilai signifikansi kata-kata dalam teks yang sedang dianalisis berdasarkan seberapa sering dan di mana kata-kata tersebut muncul dalam semua dokumen. Metode TF-IDF digunakan untuk mengekstraksi fitur-fitur kunci yang dapat memengaruhi keseluruhan sentimen dalam teks tersebut[14].

Membangun Model & Pengujian

Model-model klasifikasi yang telah diinisialisasi, seperti *Random Forest*, *Neural Network*, dan SVM, masing-masing dilatih dengan menggunakan data latih. Setelah melalui proses pelatihan, model-model tersebut dievaluasi menggunakan data uji untuk mengukur kinerja mereka dalam mengklasifikasikan emosi pada tweet.

Hasil evaluasi diukur dengan menggunakan sejumlah metrik evaluasi, seperti akurasi, presisi, *recall*, dan *F1-score*. Akurasi mengukur tingkat kebenaran secara umum, sedangkan presisi mengindikasikan kemampuan model dalam mengklasifikasikan dengan benar data positif secara proporsional terhadap hasil positif yang diberikan. *Recall* mengukur kemampuan model dalam mengenali kelas positif dengan benar dari seluruh kelas positif yang ada. *F1-score* menyatukan presisi dan *recall* dalam satu skor.

Model	Accuracy	Precision	Recall	F1-score
Random Forest	0,625538	0,720715	0,559137	0,574114
Neural Network	0,619799	0,672275	0,602246	0,625667
SVM	0,618364	0,76062	0,531937	0,549986

Random Forest

Model *Random Forest* memiliki akurasi sekitar 62.55%, yang mengindikasikan sejauh mana model ini dapat mengklasifikasikan data dengan benar secara keseluruhan. Nilai presisinya (72.07%) menunjukkan proporsi positif yang benar di antara semua hasil positif yang diberikan oleh model. *Recall* (55.91%) mengukur kemampuan model dalam mengidentifikasi dengan benar kelas positif dari seluruh kelas positif yang sebenarnya. *F1-score* (57.41%) mengukur keseimbangan antara presisi dan recall.

Neural Network

Model Neural Network memiliki akurasi sekitar 61.98%, yang menunjukkan tingkat kebenaran keseluruhan model. Presisi (67.23%) mengindikasikan seberapa baik model mengklasifikasikan dengan benar data positif secara proporsional terhadap hasil positif yang diberikan. *Recall* (60.22%) mengukur kemampuan model untuk mengenali kelas positif dengan benar dari seluruh kelas positif yang ada. *F1-score* (62.57%) menyatukan presisi dan *recall* dalam skor tunggal.

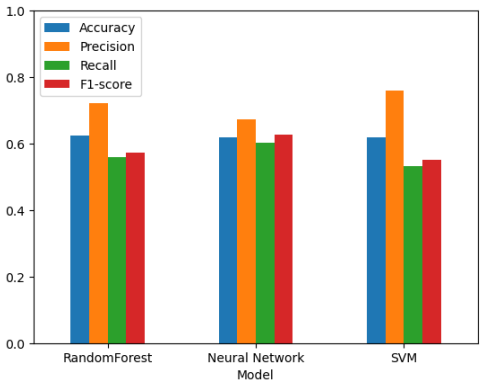
SVM

Model SVM memiliki akurasi sekitar 61.84%, mengukur tingkat kebenaran model secara umum. Presisi (76.06%) menunjukkan seberapa baik model dalam mengklasifikasikan dengan benar data positif secara proporsional terhadap hasil positif yang diberikan. *Recall* (53.19%) mengukur sejauh mana model mengidentifikasi kelas positif dengan benar dari seluruh kelas positif yang ada. *F1-score* (54.99%) mengukur keseimbangan antara presisi dan *recall*.

HASIL DAN PEMBAHASAN

Analisis ilmiah terkait hasil matriks evaluasi pada penggunaan *Edited Nearest Neighbours* (ENN) untuk pengurangan sampel data pada klasifikasi emosi dalam data tweet berbahasa Indonesia menunjukkan adanya perbandingan kinerja antara tiga model klasifikasi: *Random Forest*, *Neural Network*, dan SVM. Dalam penggunaan ENN untuk mengatasi masalah ketidakseimbangan kelas, hasil evaluasi mengungkapkan bahwa model *Random Forest* memiliki akurasi tertinggi dengan nilai sekitar 62.55%. Meskipun akurasi merupakan indikator umum untuk mengukur keberhasilan model, penting untuk melihat lebih dalam ke dalam metrik lainnya.

Dari segi presisi, Neural Network memiliki nilai tertinggi dengan 67.23%, menggambarkan kemampuan model ini dalam mengidentifikasi dengan benar kelas positif secara proporsional terhadap hasil positif yang dihasilkan. Namun, ini juga perlu dipertimbangkan bersamaan dengan nilai *recall*. Pada nilai *recall*, model *Random Forest* memiliki kinerja lebih baik dengan nilai 55.91%, menunjukkan kemampuan model dalam mengenali kelas positif dari total kelas positif yang ada. Dalam hal keseimbangan antara presisi dan *recall*, nilai F1-score juga menjadi pertimbangan penting. Model *Neural Network* mendominasi dalam hal ini, dengan F1-score sekitar 62.57%, menunjukkan keseimbangan yang lebih baik antara presisi dan *recall*.



Secara keseluruhan, penggunaan ENN untuk pengurangan sampel tampaknya memberikan efek positif dalam menghadapi masalah ketidakseimbangan kelas pada data tweet emosi berbahasa Indonesia. Model RandomForest dan Neural Network menunjukkan hasil yang kompetitif dalam berbagai metrik evaluasi, dengan model Neural Network menonjol dalam keseimbangan antara presisi dan *recall*. Meskipun model SVM memiliki nilai presisi yang tinggi (76.06%), rendahnya recall (53.19%) dan F1-score (54.99%) menunjukkan tantangan dalam mengenali kelas positif dengan benar. Oleh karena itu, dalam konteks penggunaan ENN untuk mengatasi ketidakseimbangan kelas, penggunaan model RandomForest dan Neural Network dapat dianggap sebagai pilihan yang lebih baik untuk tugas klasifikasi emosi pada data tweet berbahasa Indonesia.

SIMPULAN

Dalam pengujian klasifikasi emosi pada data tweet berbahasa Indonesia dengan pengurangan sampel menggunakan *Edited Nearest Neighbours* (ENN), terdapat perbandingan kinerja antara tiga model utama: RandomForest, Neural Network, dan SVM.

Meskipun model *Random Forest* memiliki akurasi tertinggi sekitar 62.55%, kinerja model harus dinilai lebih mendalam melalui metrik lainnya seperti presisi, *recall*, dan *F1-score*.

Dari berbagai metrik evaluasi, model *Neural Network* menonjol dengan nilai presisi tertinggi (67.23%) dan *F1-score* yang seimbang (62.57%), menjadikannya pilihan yang lebih baik dalam mengatasi masalah ketidakseimbangan kelas pada data tweet emosi berbahasa Indonesia dibandingkan dengan model *Random Forest* dan SVM.

DAFTAR PUSTAKA

[1] I. M. D. Ardiada, M. Sudarma, and D. Giriantari, 'Text Mining pada Sosial Media untuk Mendeteksi Emosi Pengguna Menggunakan Metode Support Vector Machine dan K-Nearest Neighbour', *Maj. Ilm. Teknol. Elektro*, vol. 18, no. 1, p. 55, May 2019, doi: 10.24843/mite.2019.v18i01.p08.

[2] F. A. Acheampong, C. Wenyu, and H. Nunoo-Mensah, 'Text-based emotion detection: Advances, challenges, and opportunities', *Eng. Rep.*, vol. 2, no. 7, p. e12189, 2020, doi: 10.1002/eng2.12189.

[3] I. D. Abirawa, A. B. Osmond, and C. Setianingsih, 'Klasifikasi Emosi Pada Lirik Lagu Menggunakan Metode Support Vector Machine', *EProceedings Eng.*, vol. 5, no. 3, Art. no. 3, Dec. 2018, Accessed: Dec. 04, 2022. [Online]. Available: https://openlibrarypublications.telkomuniversity.ac.id/index.php/engineering/article/view/8007

[4] P. Kumar and B. Raman, 'A BERT based dual-channel explainable text emotion recognition system', *Neural Netw.*, vol. 150, pp. 392–407, Jun. 2022, doi: 10.1016/j.neunet.2022.03.017.

[5] V. A. Fitri, R. Andreswari, and M. A. Hasibuan, 'Sentiment Analysis of Social Media Twitter with Case of Anti-LGBT Campaign in Indonesia using Naïve Bayes, Decision Tree, and Random Forest Algorithm', *Procedia Comput. Sci.*, vol. 161, pp. 765–772, Jan. 2019, doi: 10.1016/j.procs.2019.11.181.

[6] F. Abdulloh, A. Aminuddin, M. Rahardi, and S. Anggita, 'Observation of Imbalance Tracer Study Data for Graduates Employability Prediction in Indonesia', *Int. J. Adv. Comput. Sci. Appl.*, vol. 13, pp. 169–174, Sep. 2022, doi: 10.14569/IJACSA.2022.0130820.

[7] A. Nurhopipah and C. Magnolia, 'PERBANDINGAN METODE RESAMPLING PADA IMBALANCED DATASET UNTUK KLASIFIKASI KOMENTAR PROGRAM MBKM', *J. Publ. Ilmu Komput. Dan Multimed.*, vol. 2, no. 1, Art. no. 1, Jan. 2023, doi: 10.55606/jupikom.v2i1.862.

[8] I. Triguero, D. García-Gil, J. Maillo, J. Luengo, S. García, and F. Herrera, 'Transforming big data into smart data: An insight on the use of the k-nearest neighbors algorithm to obtain quality data', *WIREs Data Min. Knowl. Discov.*, vol. 9, no. 2, p. e1289, 2019, doi: 10.1002/widm.1289.

[9] Yennimar -, A. Rasid, and S. Kenedy, 'IMPLEMENTATION OF SUPPORT VECTOR MACHINE ALGORITHM WITH HYPER-TUNING RANDOMIZED SEARCH IN STROKE PREDICTION', *J. Sist. Inf. Dan Ilmu Komput. PrimaJUSIKOM PRIMA*, vol. 6, no. 2, Art. no. 2, Mar. 2023, doi: 10.34012/jurnalsisteminformasidanilmukomputer.v6i2.3479.

[10] A. Onan, 'Sentiment analysis on product reviews based on weighted word embeddings and deep neural networks', *Concurr. Comput. Pract. Exp.*, vol. 33, no. 23, p. e5909, 2021, doi: 10.1002/cpe.5909.

[11] E. M. O. N. Haryanto, A. K. A. Estetikha, and R. A. Setiawan, 'IMPLEMENTASI SMOTE UNTUK MENGATASI IMBALANCED DATA PADA SENTIMEN ANALISIS SENTIMEN HOTEL DI NUSA TENGGARA BARAT DENGAN MENGGUNAKAN ALGORITMA SVM', *Inf. Interaktif*, vol. 7, no. 1, Art. no. 1, Jan. 2022.

[12] M. S. Saputri, R. Mahendra, and M. Adriani, 'Emotion Classification on Indonesian Twitter Dataset', in *2018 International Conference on Asian Language Processing (IALP)*, Nov. 2018, pp. 90–95. doi: 10.1109/IALP.2018.8629262.

[13] S. M. Chamzah, M. Lestandy, N. Kasan, and A. Nugraha, 'Penerapan Synthetic Minority Oversampling Technique (SMOTE) untuk Imbalance Class pada

Data Text Menggunakan kNN', *Syntax J. Inform.*, vol. 11, no. 02, Art. no. 02, Nov. 2022, doi: 10.35706/syji.v11i02.6940.

[14] B. B. Baskoro, I. Susanto, and S. Khomsah, 'Analisis Sentimen Pelanggan Hotel di Purwokerto Menggunakan Metode Random Forest dan TF-IDF (Studi Kasus: Ulasan Pelanggan Pada Situs TRIPADVISOR)', *INISTA J. Inform. Inf. Syst. Softw. Eng. Appl.*, vol. 3, no. 2, Art. no. 2, Jun. 2021, doi: 10.20895/inista.v3i2.218.