Jicom : Journal Information & Computer
Volume 3, Nomor 2, Juli 2025, hlm.
DOI : 10.32493/jicomisc.v2i1.37666

P-ISSN : 3031-2779
E-ISSN : 3026-4197

Sentimen Opini Publik Terhadap Isu Pembangunan Tugu di Youtube Menggunakan Metode Random Forest dan SMOTE

Ardi al Ghifari[1], M.Kurniawan[2], Nur Rachmat[3]
[1,2,3] Informatika, Fakultas Ilmu Komputer dan Rekayasa, Universitas Multi Data Palembang
Email: * [1]ardialghifari_2226250113@mhs.mdp.ac.id, [2]mkurniawan_2226250069@mhs.mdp.ac.id, [3]nur.rachmat91@mdp.ac.id

Abstrak: *Pembangunan tugu sering kali menimbulkan pro dan kontra di kalangan masyarakat, terutama ketika proyek tersebut dianggap tidak sesuai dengan prioritas kebutuhan publik. Dalam era digital saat ini, opini masyarakat terhadap isu-isu seperti ini banyak disampaikan melalui media sosial dan platform berbagi video seperti YouTube.Pada penelitian ini, data dikumpulkan dari kolom komentar pada video yang membahas pembangunan tugu, kemudian dilakukan praproses teks sebelum dianalisis menggunakan metode Random Forest dengan SMOTE dan tanpa SMOTE.Hasil yang didapatkan bahwa, Algoritma Random Forest tanpa SMOTE didapatkan akurasi sebesar 0.97 sedangkan Algoritma Random Forest dengan SMOTE didapatkan akurasi 0.78. Terlihat juga nilai Precision tertinggi di dapatkan oleh Algoritma Random Forest tanpa SMOTE yaitu sebesar 1.00 untuk positif dan 0.97 untuk negatif, lalu nilai Recall tertinggi di dapatkan oleh Algoritma Random Forest tanpa SMOTE sebesar 0.53 untuk positif dan 1.00 untuk negatif, Dan Nilai F1 Score terbesar juga didapatkan oleh Algoritma Random Forest tanpa SMOTE sebesar 0.69 untuk positif dan 0.99 untuk negatif.Dari kesimpulan dari atas bahwa terdapat tidak kecocokannya penggunaan Metode Random Forest dan SMOTE, untuk penelitian ini disarankan lebih baik menggunakan Metode Random Forest Tanpa SMOTE.*

Kata Kunci – *Sentimen; Tugu; Random Forest; SMOTE; Youtube*

Abstract: *Monument construction often generates pros and cons among the public, especially when the project is deemed incompatible with the prioritization of public needs. In today's digital era, public opinion on issues like this is widely conveyed through social media and video sharing platforms such as YouTube.In this study, data was collected from the comments column on videos discussing monument construction, then text preprocessing was carried out before being analyzed using the Random Forest method with SMOTE and without SMOTE.The results obtained that, Random Forest Algorithm without SMOTE obtained an accuracy of 0.97 while Random Forest Algorithm with SMOTE obtained an accuracy of 0.78. It can also be seen that the highest Precision value is obtained by the Random Forest Algorithm without SMOTE which is 1.00 for positive and 0.97 for negative, then the highest Recall value is obtained by the Random Forest Algorithm without SMOTE of 0.53 for positive and 1.00 for negative, and the largest F1 Score value is also obtained by the Random Forest Algorithm without SMOTE of 0.69 for positive and 0.99 for negative.From the conclusion from above that there is a mismatch in the use of the Random Forest and SMOTE methods, for this research it is recommended that it is better to use the Random Forest method without SMOTE.*

Keywords – *Sentiment; Monument; Random Forest; SMOTE; Youtube*

1. PENDAHULUAN

Sudah sejak dulu tugu menjadi bagian dari perjalanan sejarah bangsa Indonesia. Dimulai sejak kerajaan untuk memperingati keberhasilan seorang raja atau dibangunla tugu dalam bentuk prasasti. Tidak beda dengan zaman sekarang tugu nyaris tidak pernah dilupakan dalam setiap peringatan yang berkaitan dengan peristiwa bersejarah. Namun, tugu kemudian terkikis kemunculannya bersama dengan zaman, ataupun tenggelam bersama dengan arus modernisasi. Hampir di setiap kota di Indonesia maupun di dunia memiliki tugu yang menjadi Ikon sebuah kota. Akan tetapi, tugu itu hanya benda yang berdiri tegak di tengah kota, persimpangan jalan tanpa ada diceritakan peristiwa yang dikandung dari sebuah tugu.

Platform youtube menawarkan berbagai bentuk permainan yang menarik dan bermanfaat sebagai bentuk hiburan dan juga sebagai forum sosial untuk memulai perdebatan mengenai isu yang menyangkut pembuatan Tugu, identitas kawasan atau peringatan. Tugu seringkali menjadi isu sentral yang menghadirkan pro dan kontra yang beragam. Dengan kata lain, Youtube menjadi wahana yang sangat baik untuk melakukan analisis data mengenai sentimen publik. Analisis sentimen terhadap teks, seperti mengidentifikasi dan mengklasifikasikan emosi atau opini sebagai positif, negatif, atau netral, adalah salah satu teknik pengolahan

data. Analisis sentimen memungkinkan pemahaman yang lebih baik tentang persepsi publik terhadap suatu isu tertentu dan dapat membantu dalam pengambilan keputusan strategis.Dengan menggunakan metode *Random Forest*,analisis sentimen diharapkan dapat menganalisis opini publik yang muncul terhadap berbagai isu pembangunan tugu yang diharapkan akan memberikan wawasan akan kepentingan mengenai dinamika opini publik[1]. *Random Forest* adalah salah satu metode klasifikasi yang baik untuk analisis sentimen. Metode ini dapat mengatasi data yang memiliki dimensi tinggi dan tetap melakukan ensemble dengan tingkat akurasi yang tinggi [2].

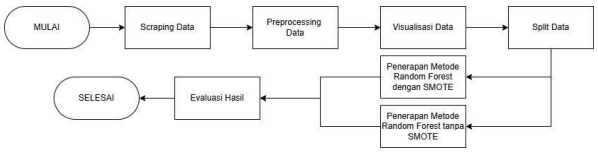
YouTube adalah platform media sosial terpopuler di Indonesia dengan jumlah pengguna sebanyak 170 juta orang atau sekitar 93.8% dari seluruh populasi negara yang mencapai 181.9 juta jiwa. Menurut laporan Hootsuite dan We Are Social pada bulan Januari 2021, anak muda usia 16–24 tahun mendominasi internet sebagai pengguna utama yang menghabiskan waktu mereka untuk menonton video online di YouTube. Artinya, mayoritas penduduk Indonesia menggunakan internet untuk mengakses video-video di platform tersebut.Fakta menyoroti potensi YouTube sebagai platform untuk penambangan data dan media untuk analisis sentimen pada topik tertentu. Banyak komentar dan pendapat dari pengguna membuat data yang tersedia secara bebas melimpah dan kaya informasi, yang memerlukan analisis lebih lanjut[3]. Proses ekstraksi data dalam bentuk teks biasanya dilakukan dengan mengambil sumber dari dokumen, dengan tujuan untuk mengidentifikasi kata-kata kunci yang dapat mewakili isi dokumen tersebut. Dengan begitu, hubungan atau keterkaitan antar dokumen dapat dianalisis lebih lanjut[4].

Sebuah penelitian lain[5] yang membahas opini terkait analisis sentimen pada aplikasi Duolingo dengan menerapkan metode Naive Bayes dan teknik *Synthetic Minority Over Sampling Technique* (*SMOTE*). Hasil penelitian menunjukkan bahwa penerapan kombinasi Naive Bayes dan *SMOTE* menghasilkan akurasi yang lebih tinggi, yaitu sebesar 91,95%, dibandingkan dengan penggunaan Naive Bayes tanpa *SMOTE* yang hanya mencapai akurasi 77,14%. Salah satu metode *Imbalance Class* yang paling sering digunakan adalah *Synthetic Minority Oversampling Technique* atau sering disebut *SMOTE*[6]. Proses ekstraksi data dalam bentuk teks biasanya dilakukan dengan mengambil sumber dari dokumen, dengan tujuan untuk mengidentifikasi kata-kata kunci yang dapat mewakili isi dokumen tersebut. Dengan begitu, hubungan atau keterkaitan antar dokumen dapat dianalisis lebih lanjut. Proses ekstraksi data dalam bentuk teks biasanya dilakukan dengan mengambil sumber dari dokumen, dengan tujuan untuk mengidentifikasi kata-kata kunci yang dapat mewakili isi dokumen tersebut. Dengan begitu, hubungan atau keterkaitan antar dokumen dapat dianalisis lebih lanjut.

Dari uraian diatas dapat penelitian ini sangat penting dilakukan untuk melihat hasil penggunaan SMOTE dan tanpa SMOTE pada Random Forest.

2. METODE PENELITIAN

Pada penelitian ini pengujian dilakukan dengan platform Google Collab sebagai IDE berbasis cloud dengan bahasa program *Python*. Pendekatan metode *machine learning* yang digunakan untuk melakukan klasifikasi sentimen, yaitu *Random Forest* dengan *SMOTE* yang digunakan untuk menyeimbangkan dataset. Pada Gambar 1. terdapat flowchart yang menunjukan beberapa tahapan akan dilakukan untuk menghasilkan hasil yang maksimal mulai dari *Scraping data*, *preprocessing data*, visualisasi data, *split data*, Penerapan metode *Random Forest*, dan evaluasi hasil



Gambar 1. *Flowchar*t Penelitian

2.1 Scraping Data

Scraping Data sentimen komen youtube menggunakan bahasa pemrograman *python* dengan IDE Google Colab. Dataset berisi teks berbahasa Indonesia yang diperoleh dari komentar pada aplikasi youtube dengan

Judul BINGUNG HABISIN ANGGARAN? Inilah Deretan Proyek Tugu Kontroversial di Indonesia, Bikin Heboh Natizen (link : https://www.youtube.com/watch?v=r__0QUfd9Zk).

Tahapan ini diperlukan untuk mendapatkan data teks yang digunakan sebagai *dataset*,dibutuhkan Tahap *labeling*,sebagai pengkategori agar dataset bisa diproses dengan benar,*labeling* yang dilakukan dengan dipecah menjadi 2 kategori, yaitu positif, dan negatif.Contoh *labeling* bisa dilihat pada Tabel 1.

Tabel 1.Labeling Komentar Youtube

Komentar	**Label**
Salam dari Jawa tengah	Positif
Orang pintar tidak di barengi dengan akhlak yang baik ya gini, bobrok	Negatif

2.2 Data Preprocessing

Preprocessing Merupakan proses mengubah data tak terstruktur menjadi data terstruktur sesuai dengan format yang dibutuhkan[7], tujuan dari proses ini adalah untuk memastikan data mudah dibaca [8]. Proses ini terdiri dari beberapa langkah kunci yang mengharuskan mendiskreditkan kebisingan dan ketidakteraturan dari data teks itu sendiri, Untuk memastikan proses ekstraksi data berlangsung secara tepat dan efisien, diperlukan tahapan pra-pemrosesan (*preprocessing*) yang sistematis Hasil dari proses *preprocessing* data dapat dilihat pada Tabel 2.Berdasarkan hal tersebut, langkah-langkah utama dalam proses *preprocessing* data adalah sebagai berikut

2.2.1 Text Cleaning

Tahap ini adalah proses membersihkan teks dari elemen-elemen yang tidak diinginkan atau tidak relevan. Proses ini biasanya melibatkan penghapusan simbol-simbol seperti tanda baca, angka, spasi berlebih, karakter khusus dan karakter non-alfabet [9].

2.2.2 Stopword Removal

Stopword merupakan kata-kata yang tidak deskriptif yang dapat dibuang [10] Stopword Removal merupakan proses untuk menghilangkan kata-kata yang dianggap tidak memiliki makna signifikan dalam analisis teks. *Stopword* adalah kata-kata yang sering muncul dalam teks, namun tidak memberikan kontribusi penting dalam proses *text mining*. Contoh dari *stopword* antara lain: "di", "ke", "akhirnya", "yang", "bagaimanapun", "dengan", "akan", dan sebagainya.

2.2.3 Tokenized

Tokenized adalah proses pemilahan data berupa kalimat atau frasa menjadi beberapa kata [11]. Kata-kata tersebut disebut sebagai token, di mana setiap token umumnya berupa kata atau frasa yang nantinya akan dianalisis lebih lanjut dalam tahap pemrosesan teks.

2.2.4 Stemming

Stemming adalah proses mengurangi kata-kata ke bentuk dasarnya atau akarnya. Tujuan dari proses stemming adalah menghilangkan imbuhan-imbuhan baik itu berupa prefiks, sufiks, maupun

konfiks yang ada pada setiap kata [12]. Proses untuk mengubah kata yang memiliki imbuhan menjadi bentuk dasarnya. Sebagai contoh, kata “makanan” akan diubah menjadi kata dasar “makan”.

Tabel 2.Proses Preprocessing Data

Sebelum Preprocessing	**Setelah Preprocessing**
Jangan menghina ngawi	jangan hina ngawi
Gilanya pemerintah konoha	gila pemerintah konoha

2.3 Visualisasi Data

Visualisasi data adalah sebuah visual yang menunjukan sentimen berdasarkan kelasnya. Pada proyek ini visualisasi berupa *Word Cloud* pada Gambar 2. yang menampilkan kata-kata sesuai dengan jumlah kemunculannya dalam sentimen positif ataupun negatif

Gambar 2. *Visualisasi Data berupa Word Cloud*

Word Cloud adalah gambaran visual berdasarkan frekuensi kemunculan kata-kata pada suatu kumpulan teks, dimana ukuran huruf menentukan frekuensi kemunculan sebuah kata yang artinya semakin besar ukuran huruf maka semakin besar kemunculan kata tersebut dan sebaliknya, semakin kecil huruf maka semakin kecil frekuensi kemunculan kata tersebut [13] yang berguna untuk menganalisis teks untuk membantu mengidentifikasi dan menyoroti kata-kata yang paling sering muncul dalam kumpulan data. Pada Visualisasi positif kata terbanyak di dapatkan oleh kata ”tugu” sedangkan visualisasi negatif kata terbanyak di dapatkan oleh kata ”korupsi”

2.4 Split Data

Pada tahap *split data*, data akan dibagi menjadi 2, yaitu data *training* sebesar 85% atau sebanyak 3865 data dan data *testing* sebanyak 15% atau sebanyak 580 data. Data *training* digunakan untuk melatih model dengan menggunakan Algoritma *Random Forest*, kemudian data *testing* digunakan untuk mengevaluasi kinerja model Algoritma *Random Forest* yang telah dibuat.

2.5 Penerapan Metode Random Forest dengan SMOTE dan tanpa SMOTE

Random Forest adalah algoritma machine learning yang menggunakan kombinasi pohon keputusan untuk membuat prediksi yang akurat guna menentukan cara yang lebih tepat dalam memproses data.Kelebihan *Random Forest* adalah dapat menangani kumpulan data yang besar dengan banyak fitur yang beragam untuk dapat mengolah data tersebut dengan baik serta mengatasi masalah *overfitting* yang dapat terjadi pada pohon hutan keputusan tunggal. Dan dapat menjaga stabilitas kinerja yang tinggi dan baik.[14].*SMOTE* merupakan teknik menyeimbangkan jumlah distribusi data sampel pada kelas minoritas dengan cara menyeleksi data sampel tersebut hingga jumlah data sampel menjadi seimbang dengan jumlah sampel pada kelas mayoritas.[15]

2.6 Evaluasi Hasil

Evaluasi hasil adalah tahap penting dalam pengembangan algoritma Random Forest. Evaluasi bertujuan untuk menilai kinerja model dalam melakukan tugas klasifikasi, memastikan model bekerja dengan baik pada data baru yang belum pernah dilihat sebelumnya.

Perbandingan kinerja antara *Random Forest* tanpa *SMOTE* dan *Random Forest* dengan *SMOTE* dilakukan menggunakan perhitungan *matrix accuracy*, *precision*, *recall*,dan *F1-score* performa masing-masing model. Berikut merupakan formula atau rumus persamaan untuk mendapatkan nilai dari *precision*, *recall*, *accuracy* dan *F1-score* [16] :

$$precision = \frac{TP}{TP + FP} \quad (1)$$

$$recall = \frac{TP}{TP + FN} \quad (2)$$

$$accuracy = \frac{TP + TN}{TP + TN + FN + FP} \quad (3)$$

$$F1 = \frac{2 \times precision \times recall}{precision + recall} \quad (4)$$

Keterangan :

TP = True Positive

TN = True Negative

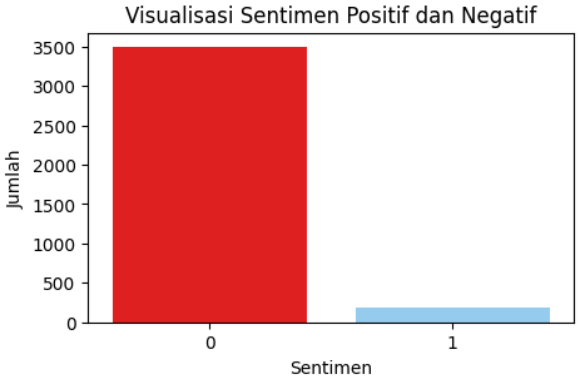
FP = False Positive

FN = False Negative

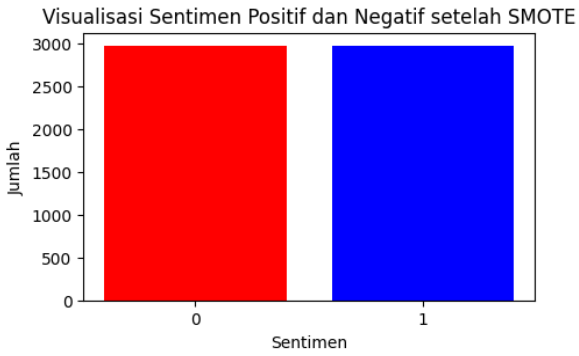
Dengan menggunakan hasil dari model *Random Forest*, kita bisa mendapatkan hasil klasifikasi yang baik berdasarkan karakteristik data dan tujuan analisis. Evaluasi hasil yang tepat akan membantu dalam memilih model yang memberikan keseimbangan terbaik antara kompleksitas dan kinerja prediktif.

3. HASIL DAN PEMBAHASAAN

Dalam penelitian ini *dataset* yang digunakan berjumlah 3865 komentar,setelah dilakukannya *preprocessing* maka didapatkan hasil yang tidak seimbang antara jumlah sentimen positif dan negatif,untuk menyeimbangkan dataset yang dihasilkan peneliti menggunakan *SMOTE* agar hasil sentimen positif dan negatif menjadi seimbang yang bisa dilihat pada Gambar 3 dan Gambar 4.



Gambar 3. Visualisasi Sentimen Tanpa SMOTE



Gambar 4. Visualisasi Sentimen dengan SMOTE

Berikut hasil dari pengujian dengan Model *Random Forest* tanpa *SMOTE* pada Tabel 3 dan Model *Random Forest* dengan *SMOTE* pada Tabel 4.

Tabel 3.Hasil Pengujian Model tanpa SMOTE

	Random Forest tanpa SMOTE	
	Positif	Negatif
Accuracy	0.97	
Precision	1.00	0.97
Recall	0.53	1.00
F1 Score	0.69	0.99

Tabel 4. Hasil Pengujian Model dengan SMOTE

	Random Forest dengan SMOTE	
	Positif	Negatif
Accuracy	0.78	
Precision	0.12	0.96
Recall	0.47	0.80
F1 Score	0.20	0.87

Berdasarkan dari Tabel 3 dan Tabel 4. terdapat perbedaan nilai *accuracy*, *precision*, dan *recall* pada kedua Metode. Pada *Random Forest* tanpa *SMOTE* didapatkan akurasi sebesar 0.97, sedangkan *Random Forest* dengan *SMOTE* didapatkan akurasi sebesar 0.78.

4. KESIMPULAN

Berdasarkan pengujian yang telah dilakukan, didapatkan kesimpulan bahwa Algoritma *Random Forest* tanpa *SMOTE* terbukti lebih baik dibandingkan Algoritma *Random Forest* dengan *SMOTE* untuk melakukan klasifikasi Sentimen Opini Publik Terhadap Isu Pembangunan Tugu di Platform Youtube Menggunakan Metode *Random Forest* dan *SMOTE*. Algoritma *Random Forest* tanpa *SMOTE* didapatkan akurasi sebesar 0.97 sedangkan Algoritma *Random Forest* dengan *SMOTE* didapatkan akurasi 0.78. Terlihat juga nilai Precision tertinggi di dapatkan oleh Algoritma *Random Forest* tanpa *SMOTE* yaitu sebesar 1.00 untuk positif dan 0.97 untuk negatif, lalu nilai Recall tertinggi di dapatkan oleh Algoritma *Random Forest* tanpa *SMOTE* sebesar 0.53 untuk positif dan 1.00 untuk negatif, Dan Nilai F1 Score terbesar juga didapatkan oleh Algoritma *Random Forest* tanpa *SMOTE* sebesar 0.69 untuk positif dan 0.99 untuk negatif.

DAFTAR PUSTAKA

[1] Medhat, W., Hassan, A., & Korashy, H. (2016). Sentiment analysis algorithms and applications: A survey. Ain Shams Engineering Journal, 5(4), 1093-1113. https://doi.org/10.1016/j.asej.2014.04.011.

[2] Rakhmawati, N., & Wahyuni, E. S. (2021). Perbandingan Metode Klasifikasi Random Forest dan SVM Pada Analisis Sentimen PSBB. Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi), 5(3), 504-511.

[3] R. Kurniawan, F. Lestari, A. S. Batubara, M. Z. A. Nazri, K. Rajab, and R. Munir, "Indonesian LexiconBased Sentiment Analysis of Online Religious Lectures Review," 2021.

[4] E. K. Putri and T. Setiadi, "Penerapan Text Mining Pada Sistem Klasifikasi Email Spam Menggunakan Naive Bayes," J. Sarj. Tek. Inform., vol. 2, no. 3, pp. 73–83, 2014.

[5] S. Chohan, A. Nugroho, A. Maezar Bayu Aji, W. Gata, and S. Nusa Mandiri, "Analisis sentimen aplikasi duolingo menggunakan metode naïve bayes dan synthetic minority over sampling technique," Paradigma –Jurnal Informatika dan Komputer, vol. 22, no. 2, pp. 139–144, 2020.

[6] R. B. Bahaweres, F. Agustian, I. Hermadi, A. I. Suroso, and Y. Arkeman, "Software defect prediction using neural network based smote," in International Conference on Electrical Engineering, Computer Science and Informatics (EECSI), Oct. 2020, vol. 2020-October, pp. 71–76. doi: 10.23919/EECSI50503.2020.9251874.

[7] Lasama, A. P. E. P, A. Prasetiadi, F. Teknologi, I. Teknologi, and T. Purwokerto, "Prediksi Tsunami Pada Gempa Menggunakan Random Forest Classifier," 2019.

[8] P. Mega, N. Dharmapatni, N. Luh, and P. Merawati, "Jurnal Bumigora Information Technology (BITe) Penerapan Algoritma Support Vector Machine Dalam Sentimen Analisis Terkait Kenaikan Tarif BPJS Kesehatan Jurnal Bumigora Information Technology (BITe) Jurnal Bumigora Information Technology (BITe) Jurnal," vol. 2, no. 2, pp. 105–112, 2020, doi: 10.30812/bite.v2i2.904.

[9] F. A. Nugraha, N. H. Harani, and R. Habibi, Analisis Sentimen Terhadap Pembatasan Sosial Menggunakan Deep Learning, Pertama. Pertama.Bandung: Kreatif Industri Nusantara, 2020

[10] S. Syafrizal, M. Afdal, and R. Novita, "Analisis Sentimen Ulasan Aplikasi PLN Mobile Menggunakan Algoritma Naïve Bayes Classifier dan K-Nearest Neighbor," MALCOM: Indonesian Journal of Machine Learning and Computer Science, vol. 4, no. 1, pp. 10–19, Dec. 2023, doi: 10.57152/malcom.v4i1.983.

[11] L. Aji Andika and P. Amalia Nur Azizah, "Analisis Sentimen Masyarakat terhadap Hasil Quick Count Pemilihan Presiden Indonesia 2019 pada Media Sosial Twitter Menggunakan Metode Naive Bayes Classifier," 2019.

[12] Y. P. Akbar, M. Sri Satyawati, and N. Putra Sastra, "Analisis Sentimen Kata Anjay pada Media Sosial Twitter Dalam Kajian Linguistik Komputasi," Stilistika : Journal of Indonesian Language and Literature, vol. 1, no. 2, p. 62, Apr. 2022, doi: 10.24843/stil.2022.v01.i02.p06.

[13] Z. Amrullah, A. Sofyan Anas, M. Adrian, and J. Hidayat, "Analisis Sentimen Movie Review Menggunakan Naive Bayes Classifier Dengan Seleksi Fitur Chi Square," Jurnal, vol. 2, no. 1, 2020, doi: 10.30812/bite.v2i1.804.

[14] J. Indri and Lindawati, "Analisis Sentimen Terhadap Sistem Informasi Akademik Mahasiswa Institut Teknologi Garut," 2022. [Online]. Available: https://jurnal.itg.ac.id/

[15] R. Siringoringo, “Klasifikasi Data Tidak Seimbang Menggunakan Algoritma SMOTE dan K-Nearest Neighbor,” J. ISD, vol. 3, no. 1, pp. 44–49, 2018.

[16] F. Caroline, R. G. S. Budi, dan M. E. A. Rivan, "Analisis Sentimen Masyarakat terhadap Kasus Korupsi PT. Timah Menggunakan Metode Support Vector Machine," *Jurnal Ilmu Komputer dan Informatika (JIKI)*, vol. 4, no. 1, pp. 43–50, Jun. 2024, doi: 10.54082/jiki.141.