Analisis Sentimen Era Kepelatihan Shin Tae Yong dengan Menggunakan SVM dan SMOTE

Putu Wahana MahaYoni[a1], I Gusti Ngurah Anom Cahyadi Putra[a2]

[a]Program Studi Informatika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Udayana
Jalan Raya Kampus Unud, Jimbaran, Bali, 80361, Indonesia
[1]yoni2308561110@student.unud.ac.id
[2]anom.cp@unud.ac.id

Abstract

This study explores the application of a combined approach using TF-IDF, SMOTE, and Support Vector Machine (SVM) to address sentiment classification on imbalanced text data. The dataset consists of 3,377 social media reviews categorized into three sentiment classes positive, negative, and neutral. Text features were extracted using TF-IDF, and class imbalance was handled using the SMOTE technique. The SVM model was trained and evaluated, achieving an accuracy of 90.82% and a weighted average F1-score of 0.91. The results demonstrate that the proposed method effectively improves sentiment classification performance, particularly in handling class imbalance.

Keywords: *Sentiment Analysis, TF-IDF, SMOTE, SVM, Text Classification.*

1. Pendahuluan

Media sosial kerap digunakan untuk menyampaikan opini terhadap suatu layanan ataupun ujaran kebencian yang mempengaruhi individu maupun masyarakat secara negatif maupun positif. Salah satu platform populer Twitter (X) kerap digunakan untuk membahas hal yang sedang hangat diperbincangkan, salah satunya performa atlet atau tim olahraga termasuk Timnas Indonesia. Era kepelatihan Shin Tae Yong menimbulkan banyak perdebatan publik yang dapat dianalisis melalui pendekatan analisis sentiment. Analisis sentimen adalah bagian dari *Natural Language Processing (NLP)* yang digunakan untuk mengidentifikasi konten dalam kumpulan data dalam bentuk opini atau pandangan tekstual (sentimen) tentang topik atau peristiwa positif, negatif atau netral [1]. Analisis sentimen dapat dilakukan dengan mengidentifikasi kata dan frasa positif, negatif, atau netral, serta dengan mencari pola penggunaan kata-kata tersebut. Dalam proses analisis, sering kali terjadi ketidakseimbangan distribusi label sentimen, di mana satu atau dua kelas lebih mendominasi pada jumlah data, sementara kelas lainnya memiliki representasi yang jauh lebih sedikit. Ketidakseimbangan ini dapat menyebabkan model belajar bias terhadap kelas mayoritas dan mengabaikan kelas minoritas.

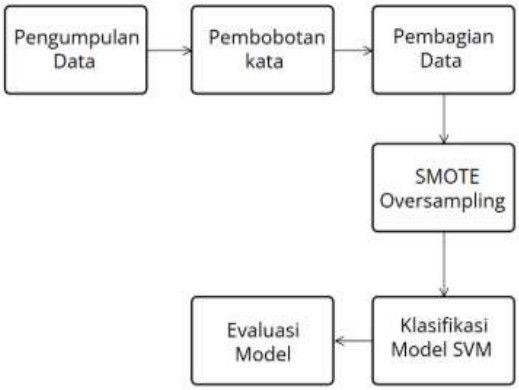
Penelitian ini membagun model kasifikasi sentimen berbagis algoritma SVM terhadap data teks dengan menerapkan teknik TF-IDF dalam ekstraksi fitur dan menggunakan teknik SMOTE untuk mengatasi permasalahan mengatasi ketidakseimbangan label pada dataset dengan menerapkan teknik SMOTE *(Synthetic Minority Over-sampling Technique)* sebagai metode oversampling. SMOTE bekerja dengan menghasilkan sampel sintetik dari kelas minoritas melalui interpolasi linier antara setiap titik data dan tetangga terdekatnya dalam kelas yang sama, sehingga distribusi kelas menjadi lebih seimbang [2].

Dalam penelitian saputra et al.(2025) melalui jurnal Instek, perbandingan performa algoritma Naive Bayes dan SVM dalam analisis data media sosial dengan pendekatan Word2Vec dan SMOTE. Hasil penelitian menunjukkan bahwa SVM yang dipadukan dengan SMOTE mampu meningkatkan performa klasifikasi, terutama dalam mengurangi bias pada mayoritas kelas. [1]

Dalam penelitian Reni Nursyanti (2025) melalui jurnal *AJIEE* membandingkan algoritma SVM dan Bi-LSTM untuk klasifikasi sentimen terhadap isu pelestarian hewan liar. Walaupun Bi-LSTM memberikan akurasi tertinggi sebesar 84%, model SVM yang dikombinasikan dengan TF-IDF dan SMOTE tetap menunjukkan performa kompetitif dengan akurasi sekitar 83% [3].

2. Metode Penelitian

Penelitian ini disusun secara sistematis untuk mencapai tujuan utama yaitu membangun model klasifikasi sentimen menggunakan SVM (*Algoritma support vector machine*) dan menggunakan SMOTE (*Synthetic Minority Over-sampling Technique*) untuk menyeimbangkan data sentimen. Penelitian ini menggunakan pendekatan kuantitatif dengan metode eksperimen dimana data dikumpulkan, diproses dan dianalisis secara terstruktur melalui beberapa tahapan yang dimulai dari pengumpulan data, pembobotan kata, pembagian data, ekstrasksi penyeimgang kelas SMOTE, pelatihan menggunakan algoritma SVM hingga evaluasi performa model.



Gambar 1. Metode Penelitian

2.1. Pengumpulan Data

Data yang digunakan pada penelitian ini merupakan open dataset bahasa indonesia yang bersumber dari platform Kaggle, dengan nama file "Tweets Data_STY_label_imblnc.csv" yang membahas opini publik terhadap kinerja Tim Nasional Indonesia di bawah pelatih Shin Tae Yong. Dataset ini terdiri dari 3.377 data yang dimana datanya sudah dibersihkan dan diberikan label. Data ini memiliki dua atribut utama, yaitu text dan Label. Atribut teks berisi teks komentar yang ditulis oleh pengguna mengenai Shin Tae Yong, sementara atribut label menunjukkan klasifikasi sentimen komentar tersebut, yaitu Positif Negatif dan netral. Dari total 3.377 komentar, 1458 merupakan komentar netral 1.355 negative dan 564 komentar positif.

2.2. Pembobotan Kata

Ekstraksi fitur TF IDF dilakukan dengan mengkonversi data teks menjadi bentuk vektor agar dapat mempertimbangkan urutan kata yang tepat. Setiap kata dalam korpus berkorelasi dengan angka oleh TF-IDF yang menunjukkan seberapa signifikan setiap kata untuk korpus [4]. TF-IDF digunakan untuk mengubah teks menjadi representasi numerik berbobot, di mana bobot dihitung berdasarkan frekuensi kata di suatu dokumen dibandingkan dengan seluruh korpus. Dalam dokumen.

2.3. Pembagian Data

Pada tahap ini data akan dibagi menjadi dua kategori yaitu data training dan testing. Data training akan digunakan untuk melatih model, lalu model akan dievaluasi pada data baru yaitu data testing. Data akan dibagi 80% digunakan untuk data training dan 20% akan digunakan untuk evaluasi model.

2.4. SMOTE Synthetic Minority Over-sampling Technique

Pada tahap ini dilakukan penyeimbangan jumlah data dengan membuat data baru yang lebih mirip dan identik dengan minoritas yang sudah ada. SMOTE adalah teknik oversampling yang dugunakan untuk menangani ketidaksembangan kelas dalam ataset ketika sutu kelas lebih sedikit daripada kelas yang lainnya. Ini mencegah masalah seperti model yang terlalu fokus pada data yang sedikit dan menghindari penghilangan informasi penting seperti yang terjadi pada cara mengurangi data dari kelompok mayoritas [1].

2.5. Klasifikasi Model

Model klasifikasi yang digunakan adalah Support Vector Machine (SVM) dengan kernel linear. SVM adalah algoritma pembelajaran mesin yang bekerja berdasarkan prinsip *Structural Risk Minimization* (SRM) dengan tujuan untuk menemukan *hyperplane* optimal yang memisahkan dua kelas dalam ruang input [5].

2.6. Evaluasi Model

Model yang telah dilatih kemudian diuji pada data testing. Evaluasi dilakukan dengan melihat performa dengan Akurasi, Prescision, recal dan F1-score. Akurasi yang dimaksud adalah proporsi prediksi yang benar terhadap total data, precision adalah ketepatan prediksi untuk masing-masing kelas, recal tingkat keberhasilan model dalam menemukan data dari masing-masing kelas dan F1-score adalah harmonisasi precision dan recall.

3. Hasil dan Diskusi

3.1. Pengumpulan Data

Data yang digunakan dalam penelitian ini terdiri dari 3.377 data ulasan pengguna yang telah dibersihkan dan diberikan label sentimen positive, negative dan neutral. Data bersumber dari platform kaggle. Data yang diperoleh menunjukan adanya perbedaan bentuk bahasa dan struktur kalimat. Hal ini menunjukan pentingnya pengolahan data teks agar dapat diekstraksi ke model pembelajaran mesin.

Tabel 1. Data Ulasan Pengguna

No	Review	Sentiment
1	dipecat pssi sty join psht	Negative
2	tetap percaya sty stystay	Positive
3	pemecatan sty ini pengalihan isu apa menurut kalian	Negative
4	selamat sty sudah lepas dari vederasi gatau di untung	Negative
5	sty adalah kita bedanya cyma di jumlah gaji aja	Neutral

3.2. Pembobotan Kata

Untuk mengubah data teks menjadi numerik agar dapat diproses oleh algoritma machine learning digunakan metode TF-IDF (*Term Frequency - Inverse Document Frequency*). Metode ini memberikan bobot pada setiap kata berdasarkan kemunculannya. Jumlah ekstraksi dibatasi hingga 3.000 kata terpenting. Setiap ulasan dikonversi menjadi vektor berdimensi 3000. Dimensi hasil TF-IDF (3377, 3000). Contoh 10 fitur pertama yaitu ['aaaaa', 'aaaarghhhh', 'abadi', 'abaikan', 'abangabangan', 'abduh', 'abidal', 'abidzar', 'ability', 'abis']. Dengan hasil ini, setiap ulasan memiliki representasi digital yang unik dan siap untuk digunakan dalam proses pelatihan model klasifikasi.

3.3. Pembagian Data

Data dibagi menjadi dua subset data training dan testing. Data training sebesar 80% dari total data yaitu 2.701 data dan data testing sebesar 20% dari total data yaitu 676 data. Dapat terlihat bahwa data mengalami ketidakseimbangan kelas, di mana kelas positive jauh lebih sedikit dibandingkan kelas lainnya.

Tabel 2. Pembagian Data

Sentimen	Data training	Data testing
Neutral	1.166	292
Negative	1.084	271
Positive	451	131

3.4. SMOTE

Distribusi label pada data pelatihan menunjukkan adanya ketidakseimbangan, di mana kelas positive hanya berjumlah 451 data, jauh lebih sedikit dibandingkan kelas negative dan neutral. Hal ini dapat menyebabkan model cenderung bias terhadap kelas mayoritas. Oleh karena itu, dilakukan penyeimbangan kelas pada data latih menggunakan teknik SMOTE (*Synthetic Minority Over-sampling Technique*) untuk meningkatkan kinerja klasifikasi, terutama pada kelas minoritas. Jumlah data training sebelum smote adalah 2.701, setelah ekstraksi teknik smote menjadi 3.498.

Tabel 3. Penyeimbang Kelas SMOTE

Sentimen	Sebelum	Sesudah
Neutral	1.166	1.166
Negative	1.084	1.166
Positive	451	1.166

3.5. SVM

Model klasifikasi yang digunakan dalam penelitian ini adalah Support Vector Machine (SVM) dengan kernel linear. Model SVM dipilih karena kemampuannya yang baik dalam menangani data berdimensi tinggi dan cocok untuk teks yang telah ditransformasi menjadi vektor TF-IDF. Model dilatih menggunakan data hasil SMOTE dan proses pelatihan berhasil dilakukan tanpa kesalahan.

3.6. Evaluasi

Model SVM yang dilatih pada data training hasil oversampling SMOTE menunjukkan performa klasifikasi yang sangat baik. Evaluasi dilakukan pada data testing sebanyak 676 data dengan hasil akurasi keseluruhan 90,82% dengan pengukuran sebagai berikut.

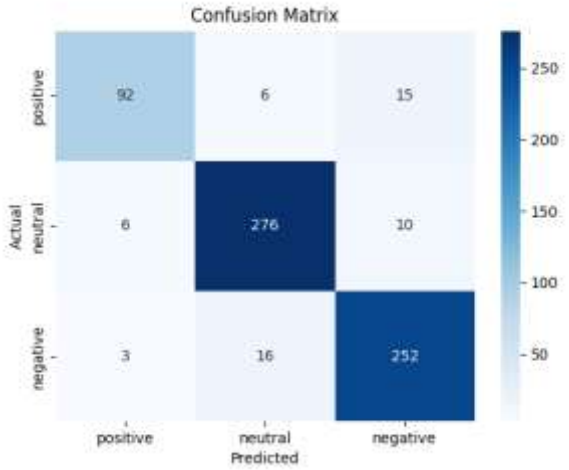
Tabel 4. Hasil Evaluasi

Sentimen	Precision	Recal	F1-score	Jumlah Data
Neutral	0.90	0.93	0.91	271
Negative	0.92	0.93	0.93	292
Positive	0.98	0.80	0.84	113

Rata rata avg F1-score adalah 0.89 dan Weighted avg FI-score adalah 0.91 hasil ini menunjukan bahwa model berkerja dengan baik meskipun kelas positive lebih rendah dibandingkan kelas lainnya.

3.7. Analisis Confusion Matrix

Gambar *Confusion Matrix* menunjukkan hasil prediksi model SVM terhadap data uji sebanyak 676 data. Masing-masing baris merepresentasikan label aktual, dan masing-masing kolom menunjukkan label prediksi.



Gambar 2. Confusion Matrix

Dari gambar *cofusion matix* diatas menunjukan bahwa kelas sentimen neutral memiliki *precision* dan *recaly* yang baik, dimana hal ini menunjukan bahwa model mampu mengidentifikasi sentimen neutral dengan sangat akurat. Kelas positive memiliki nilai recal yang sedikit lebih rendah mengindikasikan bahwa beberapa data positif salah diklasifikasikan ke dalam kelas lain, khususnya di kelas negative. Nilai rata-rata makro dan rata-rata tertimbang F1-score masing-masing mencapai **0.91** dan **0.92**, yang menunjukkan model memiliki kemampuan klasifikasi yang seimbang di seluruh kelas. *Confusion matrix* memperlihatkan distribusi prediksi yang seimbang dan tidak terlalu banyak kesalahan klasifikasi antar kelas. Hal ini menandakan bahwa strategi preprocessing, ekstraksi fitur dengan TF-IDF, dan penyeimbangan data menggunakan SMOTE berhasil meningkatkan kemampuan generalisasi model SVM terhadap data uji.

3.8. Pembahasan

Hasil eksperimen menunjukkan bahwa pendekatan yang digunakan sangat efektif untuk klasifikasi sentimen pada data teks ulasan. Model SVM menunjukkan keunggulan dalam klasifikasi teks karena kemampuannya membentuk garis batas keputusan (*decision boundary*) yang optimal dalam ruang berdimensi tinggi. Dengan menggunakan kernel linear, model menjadi lebih efisien dan dapat dikontrol dengan baik untuk menghindari overfitting. Dalam penelitian ini, SVM mampu mengklasifikasikan dengan tepat sebagian besar ulasan dengan sentimen neutral dan negative. Sementara itu, performa pada sentimen positive masih sedikit menurun karena kompleksitas bahasa positif yang lebih variatif dan lebih jarang muncul dalam data. Secara keseluruhan, kombinasi antara TF-IDF, SMOTE, dan SVM merupakan pendekatan yang baik dan dapat diandalkan untuk tugas analisis sentimen pada data teks tidak seimbang.

4. Kesimpulan

Berdasarkan hasil pengujian yang telah dilakkan terdapat kesimpulan mengenai efektifitas model klasifikasi sentimen mengunakan kombinasi metode TF-IDF, SMOTE dan SVM. dimana

p-ISSN: 2986-3929
e-ISSN: 3032-1948

Algoritma berhasil mencapai akurasi sebesar 90,82% dalam mengklasifikasikan teks ke dalam 3 katagori positive neutral dan negative. Metode TF-IDF mampu memnghasilkan informasi dalam bentuk vektor numerik berdimensi 3000 sehingga memungkinkan model macine learning untuk memahami makna dari kana penting. SMOTE berhasil menangani mesalah ketidak seimbangan data pada kelas minoritas (positive) dengan menghasilkan data sintetis yang relevan dengan demikian model dapat berjalan secara merata dan tidak contong terhadap kelas mayoritas. Kelas sentimen neutral memiliki performa terbaik yang terbukti dengan nilai precision dan recal yang tinggi sehingga menandakan model mampu membedakan karakterisitik kelas dengan baik.

Secara keseluruhan, pendekatan yang digunakan dalam penelitian ini dapat dijadikan acuan untuk pengembangan sistem klasifikasi sentimen, terutama ketika menghadapi dataset yang tidak seimbang. Integrasi TF-IDF, SMOTE, dan SVM terbukti menjadi kombinasi yang cocok dalam menangani tantangan pada data ulasan sosial media. Namun adapun bebrapa pengembangan yang dapat dilakukan untuk menjadi fokus penelitian lanjutan seperti Eksplorasi teknik representasi semantik berbasis deep learning, seperti Word2Vec, GloVe, atau BERT untuk menangkap konteks makna kata secara lebih mendalam. Selain itu meningkatkan kualitas dataset, baik dari sisi jumlah data maupun keragaman topik perlu dilakuakan agar model lebih stabil terhadap perubahan data, noise, error, atau kondisi ekstrem lainnya.

Daftar Pustaka

[1] J. Saputra, L. Maryani, R. Rahmaddeni, D. Wulandari, and W. Eka, “Analisis Performa Naive Bayes dan SVM terhadap Sentimen Teks Media Sosial dengan Word2Vec dan SMOTE,” *Instek: Jurnal Penelitian Sains dan Teknologi*, vol. 10, no. 1, pp. 143–155, Apr. 2025. [Online].
Available: https://journal.uin-alauddin.ac.id/index.php/instek/article/view/54889

[2] S. Rabbani, D. Safitri, N. Rahmadhani, A. A. F. Sani, and M. K. Anam, “Comparative Evaluation of SVM Kernels for Sentiment Classification in Fuel Price Increase Analysis,” *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, vol. 3, no. 2, pp. 153–160, Oct. 2023. [Online]. Available: https://doi.org/10.57152/malcom.v3i2.897

[3] R. Nursyanti, N. Alamsyah, and T. M. Yusuf, “Perbandingan dan Efektivitas Kinerja Algoritma SVM dan Bi-LSTM dengan TF-IDF dan SMOTE untuk Analisis Sentimen Kelestarian Hewan Liar,” *Aisyah Journal of Informatics and Electrical Engineering*, vol. 7, no. 1, 2025. [Online]. Available: http://jti.aisyahuniversity.ac.id/index.php/AJIEE

[4] V. Rupapara, F. Rustam, H. F. Shahzad, A. Mehmood, I. Ashraf, and G. S. Choi, “Impact of SMOTE on imbalanced text features for toxic comments classification using RVVC model,” IEEE Access, vol. 9, pp. 78621–78634, 2021.
doi: 10.1109/ACCESS.2021.3083845

[5] O. I. Gifari, M. Adha, I. R. Hendrawan, and F. F. S. Durrand, “Analisis Sentimen Review Film Menggunakan TF-IDF dan Support Vector Machine,” *IFOTECH (Journal of Information Technology)*, vol. 2, no. 1, pp. 36–42, Mar. 2022.

[6] T. A. Anastasya, A. D. P. Saka, M. J. D. Deke, and A. M. Rizki, “Optimasi Algoritma SVM dengan PSO untuk Analisis Sentimen pada Media Sosial X terhadap Kinerja Timnas di Era Shin Tae-Yong,” *JATI (Jurnal Mahasiswa Teknik Informatika)*, vol. 9, no. 1, pp. xx–xx, Feb. 2025.