



SKEWNESS

Jurnal Statistika, Aktuaria dan Sains Data

Volume 2, No. 2, April 2025

Penggunaan Algoritma *Support Vector Machine* untuk Menganalisis Sentimen Mengenai Generasi Z di Dunia Kerja pada Pengguna Media Sosial X

Tria Yuliani¹, Mutia Nur Estri^{2*} & Indra Herdiana³

^{1,2,3} Program Studi Matematika, Universitas Jenderal Soedirman, Purwokerto, Indonesia

E-mail Koresponden: mutia.estri@unsoed.ac.id

Abstrak. Generasi Z dalam beberapa tahun terakhir menjadi salah satu topik yang ramai diperbincangkan di media sosial, terutama di platform X. Banyak unggahan yang membahas generasi Z di dunia kerja, dengan sentimen yang beragam, baik positif maupun negatif. Penelitian ini bertujuan untuk mengidentifikasi polaritas sentimen tersebut dan menguji kinerja algoritma *Support Vector Machine* (SVM) dalam analisis sentimen dengan berbagai pembagian data. Proses klasifikasi meliputi tahapan *text preprocessing*, ekstraksi fitur menggunakan *Term Frequency-Inverse Document Frequency* (TF-IDF), *resampling dataset* dengan metode SMOTETomek, dan selanjutnya klasifikasi menggunakan SVM. Penelitian dilakukan dengan tiga perbandingan data training dan testing, yaitu 70:30, 80:20, dan 90:10. Algoritma SVM diuji dengan empat jenis kernel: *linear*, *polynomial*, *Radial Basis Function* (RBF), dan *sigmoid*, serta tiga nilai *parameter cost* (C), yaitu 0,1; 1; dan 10. Hasil penelitian menunjukkan bahwa sentimen negatif lebih dominan dibandingkan sentimen positif. Akurasi tertinggi, yaitu 0,931, diperoleh pada rasio 90:10 dengan kernel linear dan $C = 1$.

Kata kunci: analisis sentimen, generasi z, support vector machine, text preprocessing

1 Pendahuluan

Seiring berjalannya waktu, generasi Z mulai berperan dalam dunia kerja bersama dengan generasi X, generasi Y, dan generasi boomers. Hal ini memungkinkan adanya empat generasi yang berbeda dalam suatu perusahaan, yaitu sebagian kecil generasi boomers yang akan pensiun, generasi X dan generasi Y yang mendominasi, dan generasi Z yang baru memasuki dunia kerja. Keberagaman generasi dalam lingkungan kerja berpotensi menimbulkan kesenjangan yang dapat mengakibatkan ketidakharmonisan dalam hubungan kerja dan berdampak pada kinerja di perusahaan [1].

Generasi Z memiliki karakteristik yang membedakannya dari generasi sebelumnya. Karakteristik tersebut terbentuk dari dinamika perkembangan yang terjadi selama rentang waktu antara 1990-an hingga 2010-an [2]. Generasi ini tumbuh diiringi dengan pesatnya kemajuan teknologi, seperti internet, perangkat pintar, dan beragam inovasi teknologi lainnya. Kondisi tersebut berpengaruh terhadap cara mereka belajar dan berinteraksi dengan lingkungan. Selain itu, berbagai perubahan sosial, seperti isu kesetaraan gender, keadilan, dan kesehatan mental turut membentuk pola pikir serta perspektif hidup generasi ini.

Dalam beberapa tahun terakhir, karakteristik yang dimiliki generasi Z menjadi sorotan di media sosial, khususnya X. Berbagai sentimen mengenai generasi Z dalam konteks pekerjaan, baik sentimen positif maupun sentimen negatif banyak dibagikan oleh pengguna X. Banyaknya pengguna yang membahas mengenai hal ini, menciptakan berbagai stereotip negatif yang cukup memengaruhi penilaian perusahaan terhadap kinerja generasi Z. Dengan demikian, dilakukan analisis polaritas sentimen positif dan negatif pengguna X mengenai generasi Z di dunia kerja.

Analisis sentimen (*opinion mining*) adalah cara untuk mengklasifikasikan sentimen ke dalam kategori positif, negatif, atau netral [3]. Algoritma *machine learning* yang banyak digunakan untuk melakukan sentimen analisis adalah algoritma *Support Vector machine* (SVM). SVM digunakan untuk melakukan klasifikasi data, baik data yang dapat diklasifikasikan secara linear maupun nonlinear. Tujuan utama SVM adalah menemukan pemisah atau disebut *hyperplane* yang secara efektif dapat memisahkan dua kelas data [4]. Berdasarkan survei terkait penelitian analisis sentimen, diketahui bahwa SVM merupakan salah satu algoritma yang banyak digunakan untuk mengklasifikasikan kelas positif dan negatif karena akurasi dan kecepatannya [5].

Penelitian yang dilakukan oleh Rizki Maulana dan Putri di tahun 2024 [6] mengenai analisis sentimen terhadap Komisi Pemilihan Umum (KPU) 2024 di Indonesia menggunakan SVM dengan kernel linear mendapat akurasi sebesar 73%. Selanjutnya, analisis sentimen mengenai kepuasan pengguna terkait promosi aplikasi Gojek menggunakan SVM dengan kernel linear juga pernah dilakukan Haryani dan mendapat akurasi sebesar 93%. Data penelitian tersebut diambil dari *tweet* pengguna X dengan beberapa kata kunci [7]. Selain itu, Rahman juga pernah melakukan klasifikasi ujaran kebencian pada media sosial X dengan membandingkan tiga kernel SVM, yaitu kernel linear, sigmoid, dan *Radian Basis Function* (RBF). Penelitian tersebut memperoleh rata-rata akurasi sebesar 92% [8].

Berdasarkan penelitian-penelitian tersebut, dapat diketahui bahwa SVM dapat digunakan untuk melakukan analisis sentimen. Penelitian ini secara khusus melakukan analisis sentimen pengguna media sosial X mengenai generasi Z di dunia kerja dengan algoritma SVM menggunakan empat jenis kernel berbeda, yaitu linear, *polynomial*, RBF, dan sigmoid. Analisis sentimen diklasifikasikan ke dalam dua kelas, yaitu sentimen positif dan sentimen negatif.

2 Metodologi

Metode yang digunakan dalam penelitian ini adalah studi kasus dengan menerapkan algoritma klasifikasi SVM untuk melakukan analisis sentimen. Data yang digunakan adalah *tweet* berbahasa Indonesia pengguna media sosial X pada tanggal 1 Agustus 2023–31 Desember 2024 yang berhubungan dengan generasi Z di dunia kerja. Penelitian ini dilakukan dengan Google Colaboratory. Algoritma SVM untuk melakukan klasifikasi adalah sebagai berikut.

1. Mempersiapkan data *training*

Dalam mempersiapkan data *training*, terdapat beberapa tahapan yang dilakukan, yaitu:

a. *Text preprocessing*

Bahasa yang digunakan di sosial media X cenderung menggunakan bahasa informal dan tidak mengikuti kaidah kebahasaan yang baku. Selain itu, terdapat komponen, seperti penggunaan tagar, simbol, ataupun mention dalam *tweet* yang diunggah. Tahapan yang digunakan dalam *text preprocessing*, yaitu *case folding*, *text cleaning*, *tokenization*, *normalization*, dan *stopword removal*.

b. Pelabelan *dataset*

pelabelan data dilakukan secara manual oleh penulis dengan membaca dan menilai satu per satu isi *tweet* berdasarkan konteks serta makna yang terkandung di dalamnya. Data diklasifikasikan ke dalam dua kategori setimen, yaitu sentimen positif dan sentimen negatif. *Tweet* yang mengandung pernyataan memuji, mengapresiasi, atau memberikan penilaian positif atas kinerja generasi Z dikategorikan sebagai sentimen positif. Sebaliknya, *tweet* yang berisi kritik, keluhan, sindiran, atau memberikan penilaian negatif dengan nada merendahkan atas kinerja generasi Z dikategorikan sebagai sentimen negatif.

c. Ekstraksi Fitur TF-IDF

TF-IDF (*Term Frequency-Inverse Document Frequency*) adalah metode statistik yang digunakan untuk menentukan seberapa relevan suatu kata di setiap dokumen dalam kumpulan dokumen (*corpus*). Dengan mengubah teks dalam dokumen menjadi vektor, isi teks tersebut diwakilkan oleh angka-angka dalam vektor tersebut. Pengubahan kata menjadi nilai dalam vektor dapat

dilakukan dengan mengalikan nilai TF (*Term Frequency*) dengan nilai IDF (*Inverse Document Frequency*), yaitu

$$TFIDF(t, m) = TF(t, m) \times IDF(t). \quad (1)$$

Nilai TF diperoleh dari persamaan

$$TF(t, m) = \frac{f(t, m)}{N_m}. \quad (2)$$

Nilai IDF diperoleh dari persamaan

$$IDF(t) = \log\left(\frac{N}{m_{f_t}}\right). \quad (3)$$

Menurut Hardle dkk., dalam [9], setiap pengamatan ke- i pada SVM, terdiri atas sepasang prediktor $\mathbf{x}_i^T = (x_{i1}, x_{i2}, \dots, x_{ip}) \in \mathbb{R}^p$, $i = 1, 2, \dots, n$, dan $y_i = \{-1, +1\}$. Jika \mathbf{x}_i adalah anggota kelas $(+1)$, maka \mathbf{x}_i dipasangkan dengan label $y_i = +1$. Jika \mathbf{x}_i adalah anggota kelas (-1) , maka \mathbf{x}_i dipasangkan dengan label $y_i = -1$. Dengan demikian, himpunan data *training* yang akan diklasifikasikan dinyatakan dalam persamaan

$$D_n = \{(\mathbf{x}_i, y_i), \dots, (\mathbf{x}_n, y_n)\} = \{(\mathbf{x}_i, y_i) \mid i = 1, 2, \dots, n\}. \quad (4)$$

2. Menghitung $K(\mathbf{x}_i, \mathbf{x}_j)$

Fungsi kernel yang digunakan dalam SVM, yaitu:

a. Kernel linear

Kernel linear adalah fungsi kernel paling sederhana dengan menghitung *dot product* antara dua titik data tanpa transformasi ke ruang fitur yang lebih tinggi.

Persamaan kernel linear, yaitu:

$$K(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i^T \mathbf{x}_j \quad (5)$$

untuk setiap i dan j .

b. Kernel *polynomial*

Kernel *polynomial* adalah fungsi kernel yang menghitung *dot product* antara dua titik data ke dalam bentuk derajat polinomial tertentu. Persamaan kernel *polynomial*, yaitu:

$$K(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i^T \mathbf{x}_j + 1)^m. \quad (6)$$

dengan $m \in \mathbb{N}$.

c. Kernel *Radian Basis Function* (RBF) atau Gauss

Kernel RBF adalah fungsi kernel yang mengukur kesamaan antara dua titik data berdasarkan jarak Euclid. Persamaan kernel RBF, yaitu

$$K(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2\sigma^2}\right). \quad (7)$$

dengan parameter σ^2 yang diestimasi dan $\sigma^2 \in \mathbb{R}$.

d. Kernel sigmoid

Kernel sigmoid adalah fungsi kernel yang dikembangkan dari jaringan saraf tiruan. Persamaan kernel sigmoid, yaitu

$$K(\mathbf{x}_i, \mathbf{x}_j) = \tanh(\alpha(\mathbf{x}_i^T \mathbf{x}_j) + \beta). \quad (8)$$

dengan $\alpha, \beta \in \mathbb{R}$.

3. Mengoptimalkan persamaan $L_D(\alpha)$

Persamaan yang dioptimalkan, yaitu

$$L_D(\alpha) = \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \alpha_i y_i \alpha_j y_j K(\mathbf{x}_i, \mathbf{x}_j). \quad (9)$$

4. Menentukan klasifikasi data *testing*

Pengklasifikasian objek baru dilakukan dengan persamaan

$$g(\mathbf{x}) = \text{sgn}\left(\sum_{i=1}^n \alpha_i y_i K(\mathbf{x}_i, \mathbf{x}) + b\right). \quad (10)$$

5. Evaluasi kinerja algoritma SVM

Salah satu cara mengevaluasi model klasifikasi adalah dengan *confusion matrix* untuk melihat klasifikasi yang benar dan klasifikasi yang salah. Berdasarkan *confusion matrix*, dapat dievaluasi kinerja model klasifikasi dengan beberapa metrik, seperti *accuracy*, *precision*, *recall*, dan *F1 Score* [10].

a. *Accuracy*

Nilai *accuracy* adalah salah satu ukuran untuk menilai kinerja model klasifikasi yang didefinisikan sebagai proporsi keseluruhan dari prediksi yang benar berdasarkan model yang dibuat. Nilai *accuracy* dihitung dengan persamaan:

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN}. \quad (11)$$

b. *Precision*

Precision atau dikenal sebagai nilai prediktif positif digunakan untuk mengukur seberapa tepat prediksi positif yang dibuat oleh model. Nilai *precision* dihitung dengan persamaan

$$Precision = \frac{TP}{TP + FP}. \quad (12)$$

c. *Recall*

Recall didefinisikan sebagai jumlah kasus dari kelas positif yang berhasil diprediksi dengan benar. Nilai *recall* dapat dihitung dengan persamaan

$$Recall = \frac{TP}{TP + FN}. \quad (13)$$

d. *F1 Score*

F1 Score adalah cara untuk mengevaluasi kinerja model dengan mempertimbangkan *precision* dan *recall*. Nilai *F1 Score* dapat dihitung dengan persamaan

$$F1\ Score = \frac{2 \times Precision \times Recall}{Precision + Recall}. \quad (14)$$

3 Hasil dan Pembahasan

3.1 Pengumpulan *Dataset*

Pengumpulan *dataset* dilakukan dengan melakukan *crawling data*. Terdapat sepuluh kata kunci yang digunakan dalam melakukan *crawling data*, yaitu “kinerja gen Z”, “gen Z kalo kerja”, “kerja gen Z”, “kerja sama gen Z”, “gen Z di kerjaan”, “gen Z di kantor”, “gen Z di dunia kerja”, “cara kerja gen Z”, “staf gen Z”, dan “pegawai gen Z”. Berdasarkan kata kunci tersebut, diperoleh 5.105 *tweet* mengenai generasi Z.

3.2 Seleksi Data

Proses seleksi data dilakukan secara manual dengan membaca satu per satu *tweet* yang diperoleh dan menghapus *tweet* yang dinilai tidak relevan. Setelah dilakukan seleksi data berdasarkan relevansi topik, diperoleh 1.020 *tweet* yang relevan dari 5.105 *tweet* hasil *crawling data*. **Tabel 1** menyajikan contoh *tweet* yang telah diseleksi dan digunakan sebagai *dataset*.

Tabel 1 *Dataset*

| <i>Tweet</i> |
|---|
| <p>“makin kesini makin banyak kerja sama gen z cepet cepet banget kerjanya pake script dll yg mudahin kerja mereka... keren lah #yeaks”</p> <p>“: Anak2 gen z kok pada lemah ya. dibentak dikit soal kerjaan nangis. . . . https://t.co/ywFSqKHxZ6”</p> |

3.3 *Text Preprocessing*

1) *Case Folding*

Case folding, yaitu mengubah seluruh huruf dalam teks menjadi huruf kecil. Hal ini karena kata “kerja”, “Kerja”, dan “KERJA” dianggap berbeda entitas. Oleh karena itu, perlu dilakukan *case folding* untuk mengurangi variasi kata yang disebabkan oleh perbedaan huruf kapital.

2) *Text Cleaning*

Text cleaning dilakukan untuk menghapus elemen yang tidak perlu, yaitu tanda baca, tautan, *mention*, angka, dan tagar. Tahapan ini membantu menghasilkan teks yang lebih bersih dan fokus pada konten utama *tweet*.

3) *Tokenization*

Tokenization adalah proses memisahkan gabungan kata menjadi unit-unit kata yang disebut token. Setiap kata dalam *tweet* dipisahkan dan diproses sebagai satuan tersendiri.

4) *Normalization*

Normalization bertujuan untuk mengatasi kata slang, singkatan, *typo* atau kata yang berlebihan sesuai dengan bentuk standarnya. Proses *normalization* dilakukan dengan tiga langkah, yaitu mereduksi pengulangan huruf, menghubungkan dengan API KBBI, dan menggunakan kamus normalisasi yang dibuat oleh penulis.

5) *Stopword Removal*

Stopword removal, yaitu menghapus kata-kata umum yang tidak memiliki makna penting dalam analisis sentimen. Penulis membuat daftar kata yang dinilai tidak memiliki makna penting pada *dataset*. Daftar tersebut disusun berdasarkan kata yang tidak ditemukan dalam KBBI dan dimodifikasi dengan menambahkan kata-kata yang sering muncul dalam *tweet* tetapi tidak mengandung sentimen.

Berikut adalah contoh hasil *text preprocessing* yang ditunjukkan dalam **Tabel 2**.

Tabel 2 *Text preprocessing*

| <i>Text Preprocessing</i> | Hasil |
|---------------------------|--|
| <i>Dataset</i> | : Anak2 gen z kok pada lemah ya. dibentak dikit soal kerjaan nangis. ... https://t.co/ywFSqKHxZ6 |
| <i>Case folding</i> | : anak2 gen z kok pada lemah ya. dibentak dikit soal kerjaan nangis. ... https://t.co/ywfsqkhxz6 |
| <i>Text cleaning</i> | anak gen z kok pada lemah ya dibentak dikit soal kerjaan nangis |
| <i>Tokenization</i> | ['anak', 'gen', 'z', 'kok', 'pada', 'lemah', 'ya', 'dibentak', 'dikit', 'soal', 'kerjaan', 'nangis'] |
| <i>Normalization</i> | ['anak', 'generasi', 'z', 'kok', 'pada', 'lemah', 'ya', 'dibentak', 'dikit', 'soal', 'kerjaan', 'nangis'] |
| <i>Stopword removal</i> | ['generasi', 'z', 'lemah', 'dibentak', 'dikit', 'soal', 'kerjaan', 'nangis'] |

3.4 Pelabelan *Dataset*

Berdasarkan 1.020 *tweet* yang dikumpulkan, sebanyak 675 *tweet* (66,18%) termasuk dalam sentimen negatif, sedangkan 345 *tweet* (33,82%) termasuk dalam sentimen positif. Proporsi ini menunjukkan bahwa pengguna media sosial X cenderung memberikan penilaian negatif terhadap generasi Z di dunia kerja.



Gambar 1 *Word Cloud*

Berdasarkan **Gambar 1**, kata-kata yang muncul dalam sentimen positif, seperti “pintar”, “cerdas”, dan “kritis” menunjukkan pengakuan atas intelektual generasi Z, sementara kata “cepat”, “rajin”, “baik”, “ambisius”, dan “berani” mencerminkan pujian atas sikap generasi Z di lingkungan kerja. Selain itu, terdapat kata “bagus”, “keren”, dan “kreatif” yang menunjukkan apresiasi atas kinerja generasi Z. Ungkapan kata “suka”, “senang”, dan “*alhamdulillah*” menunjukkan penilaian positif secara emosional. Tiga topik

utama dalam sentimen positif berdasarkan kata yang sering disebutkan adalah ide, teknologi, dan etos kerja.

Sementara itu, dalam sentimen negatif muncul kata “ngeluh”, “lemah”, dan “malas” yang menunjukkan kritik terhadap generasi Z di dunia kerja. Ungkapan kata “capek”, “kesal”, dan “anjing” mencerminkan bentuk penilaian negatif secara emosional. Topik yang dibicarakan dalam sentimen negatif, yaitu *mental health*, *attitude*, *mood* dalam bekerja, dan inisiatif di lingkungan kerja.

3.5 Ekstraksi Fitur TF-IDF

Pada penelitian ini, jumlah kata unik dalam *dataset* adalah 2.829, yang artinya vektor setiap *tweet* berdimensi 2.829. Bentuk vektor dari hasil TF-IDF adalah:

$$\mathbf{x}_i = [TFIDF(\text{absurd}, D_i), TFIDF(\text{abuse}, D_i), \dots, TFIDF(\text{zoom}, D_i)],$$

sehingga vektor yang merepresentasikan D_1 dapat ditulis menjadi $\mathbf{x}_1 = [0, 0, \dots, 0, 0]$. Hasil TF-IDF dari *dataset* dapat dilihat pada **Tabel 3**.

Tabel 3 Hasil TF-IDF

| Kata | Nilai TF-IDF | | | | | | | |
|--------|--------------|-------|-------|-------|-------|-----|------------|------------|
| | D_1 | D_2 | D_3 | D_4 | D_5 | ... | D_{1019} | D_{1020} |
| absurd | 0 | 0 | 0 | 0 | 0 | ... | 0 | 0 |
| abuse | 0 | 0 | 0 | 0 | 0 | ... | 0 | 0 |
| acara | 0 | 0 | 0 | 0 | 0 | ... | 0 | 0 |
| action | 0 | 0 | 0 | 0 | 0 | ... | 0 | 0 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... |
| sukses | 0 | 0 | 0 | 0 | 0 | ... | 0 | 0 |
| sulit | 0 | 0 | 0 | 0 | 0 | ... | 0 | 0 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... |
| zene | 0 | 0 | 0 | 0 | 0 | ... | 0 | 0 |
| zonk | 0 | 0 | 0 | 0 | 0 | ... | 0 | 0 |
| zoom | 0 | 0 | 0 | 0 | 0 | ... | 0 | 0 |

3.6 Pembagian *Dataset*

Dilakukan tiga simulasi rasio pembagian *dataset* menjadi data *training* dan data *testing*. Simulasi rasio yang dilakukan, yaitu 70: 30, 80: 20, dan 90: 10. Hasil pembagian *dataset* berdasarkan simulasi rasio dapat dilihat pada **Tabel 4**.

Tabel 4 Pembagian *Dataset*

| Rasio | Sentimen | Data <i>Training</i> | Data <i>Testing</i> | Total |
|--------------|----------|----------------------|---------------------|--------------|
| 70: 30 | Positif | 242 | 103 | 345 |
| | Negatif | 472 | 203 | 675 |
| Total | | 714 | 306 | 1.020 |
| 80: 20 | Positif | 276 | 69 | 345 |
| | Negatif | 540 | 135 | 675 |
| Total | | 816 | 204 | 1.020 |
| 90: 10 | Positif | 311 | 34 | 345 |
| | Negatif | 607 | 68 | 675 |
| Total | | 918 | 102 | 1.020 |

Berdasarkan pembagian *dataset* pada **Tabel 4**, diketahui bahwa adanya ketidakseimbangan jumlah data *training* antara sentimen negatif dan sentimen positif. Jumlah sentimen negatif lebih dominan dibandingkan dengan sentimen positif. Oleh karena itu, dilakukan *resampling* agar model dapat mengklasifikasikan kedua sentimen dengan baik. *Resampling* yang digunakan dalam penelitian ini adalah gabungan antara SMOTE *over-sampling* dan Tomek Links *under-sampling* (SMOTETomek). **Tabel 5** menunjukkan perbandingan data *training* sebelum dan sesudah *resampling*.

Tabel 5 Data *Training* Sebelum dan Setelah *Resampling*

| Rasio | Sentimen | Data <i>Training</i> Sebelum <i>Resampling</i> | Data <i>Training</i> Setelah <i>Resampling</i> |
|--------|----------|---|---|
| 70: 30 | Positif | 242 | 472 |
| | Negatif | 472 | 472 |
| 80: 20 | Positif | 276 | 538 |
| | Negatif | 540 | 538 |
| 90: 10 | Positif | 311 | 605 |
| | Negatif | 607 | 605 |

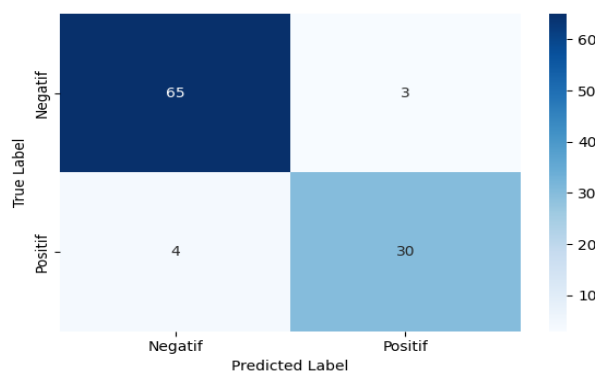
3.7 Klasifikasi Sentimen dengan SVM

Fungsi kernel yang digunakan dalam melakukan analisis sentimen dengan SVM, yaitu kernel linear, kernel RBF, kernel *polynomial*, dan kernel sigmoid. Adapun nilai parameter *cost* (*C*) yang digunakan pada setiap kernel adalah 0,1, 1, dan 10. hasil *accuracy* terbaik dari setiap rasio dapat dilihat pada **Tabel 6**.

Tabel 6 Perbandingan Evaluasi Kerja Terbaik Setiap Rasio Pembagian *Dataset*

| Rasio | Kernel | <i>Accuracy</i> | <i>Precision</i> | <i>Recall</i> | <i>F1-Score</i> |
|-------|---------------------|-----------------|------------------|---------------|-----------------|
| 70:30 | Linear ($C = 1$) | 0,892 | 0,88 | 0,87 | 0,88 |
| 80:20 | Sigmoid ($C = 1$) | 0,907 | 0,90 | 0,89 | 0,90 |
| 90:10 | Linear ($C = 1$) | 0,931 | 0,93 | 0,92 | 0,92 |

Tabel 6 menunjukkan bahwa rasio pembagian *dataset* dapat memengaruhi hasil *accuracy*, *precision*, *recall*, dan *f1-score* klasifikasi sentimen menggunakan algoritma SVM. Hasil simulasi memperlihatkan bahwa semakin besar proporsi data *training*, *accuracy* klasifikasi cenderung meningkat. Peningkatan jumlah data *training* memungkinkan model untuk mempelajari pola sentimen lebih mendalam dan akurat. Sementara itu, dari empat fungsi kernel yang digunakan, kernel linear menunjukkan hasil yang konsisten dan unggul pada rasio 70:30 dan 90:10. Secara keseluruhan, kombinasi rasio 90:10 dengan kernel linear ($C = 1$) menghasilkan *accuracy* terbaik untuk klasifikasi sentimen pada *dataset* ini dengan *accuracy* sebesar 0,931.



Gambar 2 *Confusion Matrix* Kernel Linear ($C = 1$) pada Rasio 90:10

Confusion matrix pada **Gambar 2** menunjukkan bahwa dari 102 data *testing*, kernel linear dengan $C = 1$ dapat mengklasifikasikan 65 *tweet* dengan benar sebagai sentimen negatif dan 30 *tweet* dengan benar sebagai sentimen positif. Sementara itu, terdapat 3 *tweet* bersentimen negatif yang diklasifikasikan sebagai sentimen positif (*false positive*) dan 4 *tweet* bersentimen positif yang diklasifikasikan sebagai sentimen negatif (*false negative*). Hal ini menunjukkan bahwa model memiliki kemampuan yang baik dalam membedakan sentimen negatif dan positif berdasarkan pola kalimat dan kata-kata yang termuat dalam setiap *tweet*. Terbukti bahwa *tweet* dengan sentimen negatif lebih mendominasi daripada sentimen positif.

4 Kesimpulan dan Saran

Berdasarkan analisis sentimen dari 1.020 *tweet* pengguna media sosial X mengenai generasi Z di dunia kerja, terdapat 675 *tweet* (66,18%) bersentimen negatif dan 345 *tweet* (33,82%) bersentimen positif. Hal ini mengindikasikan bahwa sebagian besar pengguna media sosial X memiliki persepsi negatif terhadap generasi Z di dunia kerja. Selanjutnya, kombinasi rasio 90: 10 dengan kernel linear ($C = 1$) menghasilkan *accuracy* terbaik untuk klasifikasi sentimen dengan *accuracy* sebesar 0,931. Adapun saran untuk penelitian selanjutnya adalah menggunakan N-gram pada TF-IDF untuk mengatasi frasa negasi, seperti “enggak pintar” dan melakukan ekstraksi fitur dengan *word embedding*, seperti *Word2Vec* atau *BERT embedding*.

5 Nomenklatur

| | |
|-----------------|--|
| $f(t, m)$ | : Frekuensi kata t dalam dokumen m |
| N_m | : Total kata dalam dokumen m |
| N | : Total jumlah dokumen dalam <i>corpus</i> |
| mf_t | : Jumlah dokumen yang mengandung kata t |
| D_n | : Himpunan data <i>training</i> |
| \mathbf{x}_i | : Vektor prediktor ke- i |
| y_i | : Label atau target ke- i |
| b | : Bias |
| L_D | : Persamaan Lagrange |
| $g(\mathbf{x})$ | : Fungsi keputusan |

Referensi

- [1] D. Rachmawati, “Welcoming gen Z in job world (Selamat datang generasi Z di dunia kerja),” *Proceeding Indones. Carr. Cent. Netw. Summit 2019*, vol. 1, no. 1, hal. 21–24, 2019.
- [2] D. Sunyoto, *Mengasah Generasi Z*, 1 ed. Kabupaten Purbalingga: EUREKA MEDIA AKSARA, 2024.
- [3] B. S. Lakshmi, P. S. Raj, dan R. R. Vikram, “Sentiment analysis using deep learning technique CNN with KMeans,” *Int. J. pure Appl. Math.*, vol. 114, no. 11, hal. 47–57, 2017.
- [4] P. Kurniawati, R. Y. Fa’rifah, dan D. Witarsyah, “Sentiment Analysis of Maxim Online Transportation App Reviews using Support Vector Machine (SVM) Algorithm,” *Build. Informatics, Technol. Sci.*, vol. 5, no. 2, hal. 466–475, 2023, doi: 10.47065/bits.v5i2.4265.
- [5] P. Kumar, T. Choudhury, S. Rawat, dan S. Jayaraman, “Analysis of Various

- Machine Learning Algorithms for Enhanced Opinion Mining Using Twitter Data Streams,” in *2016 International Conference on Micro-Electronics and Telecommunication Engineering (ICMETE)*, 2016, hal. 265–270. doi: 10.1109/ICMETE.2016.19.
- [6] M. Rizki Maulana dan R. A. Putri, “Sistemasi: Jurnal Sistem Informasi Analisis Sentimen terhadap Komisi Pemilihan Umum 2024 di Indonesia melalui Twiter menggunakan Algoritma Support Vector Machine (SVM) Sentiment Analysis of the 2024 General Election Commission in Indonesia through Twiter ,” vol. 13, hal. 1226–1242, 2024, [Daring]. Tersedia pada: <http://sistemasi.ftik.unisi.ac.id>
- [7] C. A. Haryani, A. E. Widjaja, H. Hery, dan F. V. Ferdinand, “Sentiment Analysis of User Satisfaction Towards Sales Promotion of Gojek Application Service Using Support Vector Machine (SVM),” *Ultim. InfoSys J. Ilmu Sist. Inf.*, vol. 14, no. 2, hal. 66–70, 2023, doi: 10.31937/si.v14i2.3398.
- [8] O. H. Rahman, G. Abdillah, dan A. Komarudin, “Klasifikasi Ujaran Kebencian pada Media Sosial Twitter Menggunakan Support Vector Machine,” *J. RESTI(rekayasa Sist. dan Teknol. informasi)*, vol. 5, no. 1, hal. 17–23, 2021, doi: <https://doi.org/10.29207/resti.v5i1.2700>.
- [9] W. K. Härdle, D. D. Prastyo, dan C. M. Hafner, “Support Vector Machines with Evolutionary Model Selection for Default Prediction,” in *The Oxford Handbook of Applied Nonparametric and Semiparametric Econometrics and Statistics*, J. S. Racine, L. Su, dan A. Ullah, Ed., Oxford University Press, 2014, hal. 346–373. doi: 10.1093/oxfordhb/9780199857944.013.011.
- [10] D. Sarkar, R. Bali, dan T. Sharma, *Practical Machine Learning with Python*, 1 ed. Karnataka: Apress, 2018. doi: 10.1007/978-1-4842-5121-8.