SIGMA: JURNAL PENDIDIKAN MATEMATIKA
p-ISSN: 2085-3610, e-ISSN: 2746-7503
Volume 18 Nomor 1, 2026, Halaman 465 - 479
https://journal.unismuh.ac.id/index.php/sigma/index

PENERAPAN SMOTE DAN *CLUSTER-BASED UNDERSAMPLING TECHNIQUE* DALAM KLASIFIKASI OPINI PUBLIK BERBASIS *SUPPORT VECTOR MACHINE*

Dina Zulfiana Matiyeni*, Djihad Wungguli, Siti Nurmardia Abdussamad
Statistika, Universitas Negeri Gorontalo, Indonesia
sitinurmardiaabd@gmail.com

Informasi Artikel

Submitted: April 1, 2026
Revised: April 26, 2026
Accepted: May 14, 2026

Kata Kunci
Analisis Sentimen;
SVM;
SMOTE:
Cluster-Based Undersampling;
Imbalanced Data.

Abstrak

Tujuan: Penelitian ini bertujuan untuk menerapkan metode *hybrid* yang menggabungkan SMOTE dan *Cluster-Based Undersampling Technique* guna mengatasi ketidakseimbangan data dalam klasifikasi sentimen terhadap Rancangan Undang-Undang Perampasan Aset menggunakan *Support Vector Machine* (SVM).
Metode: Penelitian ini menggunakan pendekatan kuantitatif dengan rancangan eksperimental komparatif. Data dikumpulkan dari media sosial X terkait Rancangan Undang-Undang Perampasan Aset, dilanjutkan dengan *preprocessing*, pelabelan, ekstraksi fitur, serta pemisahan data latih dan data uji. Ketidakseimbangan data diatasi dengan menggabungkan metode SMOTE dan *Cluster-Based Undersampling Technique* pada data latih. Selanjutnya, klasifikasi sentimen dilakukan menggunakan *Support Vector Machine* (SVM).
Hasil: Hasil penelitian menunjukkan bahwa model SVM tanpa penyeimbangan data menghasilkan akurasi 70,10%, presisi 62%, *recall* 46%, dan *F1-score* 47%, dengan *recall* kelas negatif yang sangat rendah sebesar 8%. Setelah penerapan metode *resampling hybrid* SMOTE dan *Cluster-Based Undersampling Technique*, performa model meningkat signifikan dengan akurasi 82%, presisi 84%, *recall* 82%, dan *F1-score* 82%, yang mengindikasikan bahwa metode *hybrid* mampu mengatasi dominasi kelas mayoritas dan meningkatkan sensitivitas model secara merata pada seluruh kelas sentimen.
Simpulan: Temuan penelitian ini mengindikasikan bahwa penerapan metode SMOTE dan *Cluster-Based Undersampling Technique* berkontribusi signifikan dalam meningkatkan keadilan prediksi model SVM pada data yang tidak seimbang. Oleh karena itu, kombinasi kedua metode tersebut dapat dijadikan solusi yang efektif dalam pengembangan sistem klasifikasi sentimen opini publik, khususnya pada kasus dengan distribusi kelas yang tidak proporsional.

Keywords
Sentiment Analysis;
SVM;
SMOTE;
Cluster-Based Undersampling;
Imbalanced Data.

Abstrack

Purpose: This study aims to implement the hybrid SMOTE and *Cluster-Based Undersampling Technique* methods in handling *imbalanced data* in sentiment classification of the Asset Confiscation Bill (RUU Perampasan Aset) using *Support Vector Machine* (SVM).
Method: This study employs a quantitative approach with a comparative experimental research design. Data were collected from platform X related to the Asset Confiscation Bill, followed by *pre-processing*, labeling, feature extraction, and data splitting into training and testing sets. *Imbalanced data* was addressed using a combination of SMOTE and *Cluster-Based Undersampling Technique* on the training data. Subsequently, sentiment classification was performed using *Support Vector Machine* (SVM).
Results: The results indicate that the SVM model without data balancing achieved an accuracy of 70.10%, precision of 62%, *recall* of 46%, and *F1-score* of 47%, with the *recall* of the negative class being critically low at 8%. After applying the hybrid *resampling* method of SMOTE and *Cluster-Based Undersampling Technique*, the model performance improved significantly with an accuracy of 82%, precision of 84%, *recall* of 82%, and *F1-score* of 82%, indicating that the hybrid method effectively overcame class majority dominance and improved model sensitivity across all sentiment classes.

Conclusion: The findings indicate that the hybrid *resampling* method of SMOTE and *Cluster-Based Undersampling Technique* significantly contributes to improving the reliability and prediction fairness of the SVM model on imbalanced data. Therefore, the combination of both methods can serve as an effective solution in developing public opinion sentiment classification systems, particularly in cases with disproportionate class distribution.

PENDAHULUAN

Kemajuan teknologi informasi telah menjadikan media sosial sebagai salah satu pendorong utama interaksi sosial masyarakat modern. Di Indonesia, media sosial kini berfungsi bukan sekadar alat komunikasi, melainkan sebagai infrastruktur politik digital yang memiliki pengaruh besar terhadap pembentukan sikap dan preferensi masyarakat terhadap kebijakan pemerintah (Sumartias dkk., 2023). Hubungan erat antara dinamika kebijakan dan respons masyarakat ini sangat terlihat pada media sosial yang memfasilitasi diskusi terbuka secara *real-time*, di mana informasi dapat menyebar luas tanpa hambatan hierarkis (Rachmawati & Nugraha, 2025).Dalam sistem digital tersebut, media sosial X (sebelumnya Twitter) menempati posisi unik sebagai ruang publik digital paling dinamis karena karakteristiknya yang berbasis teks pendek dan penggunaan *hashtag* yang mampu mengagregasi ribuan opini secara instan (Hasan dkk., 2024). Karakteristik media sosiaal ini memungkinkan terjadinya perdebatan intens mengenai isu strategis nasional, menjadikannya alat yang akurat untuk mengukur sentimen kolektif masyarakat terhadap tindakan otoritas publik (Rachmawati & Nugraha, 2025). Atas peran strategisnya sebagai jembatan antara aspirasi warga dan diskursus kebijakan, media sosial X dipilih dalam penelitian ini untuk membedah persepsi publik terhadap rancangan undang-undang perampasan aset yang memunculkan beragam opini publik.

Opini publik merupakan cerminan dari pandangan kolektif masyarakat terhadap isu-isu sosial, politik, dan ekonomi yang dapat memengaruhi arah kebijakan publik serta tindakan pemerintah (Tafana Destiana Larassetya dkk., 2024). Pemahaman terhadap opini publik menjadi langkah penting untuk mengetahui bagaimana masyarakat menilai efektivitas kebijakan pemerintah, khususnya dalam upaya pemberantasan korupsi dan penegakan hukum. Analisis terhadap opini ini tidak hanya memberikan gambaran mengenai tingkat penerimaan masyarakat, tetapi juga membantu mengidentifikasi persepsi, kepercayaan, dan sentimen yang berkembang di ruang publik (Akbari dkk., 2017). Oleh karena itu, diperlukan pendekatan analisis data yang efektif untuk mengidentifikasi serta mengklasifikasikan opini masyarakat, salah satunya melalui analisis sentimen.

Analisis sentimen adalah proses otomatis untuk memahami, mengekstrak, dan mengolah data teks guna memperoleh informasi mengenai penilaian, evaluasi, emosi, tingkat kepercayaan, dan tingkat kepuasan (Akbari dkk., 2017). Proses ini menghadapi berbagai tantangan, terutama karena karakteristik data teks di media sosial yang tidak terstruktur. *Tweet* yang mengandung singkatan, emotikon, atau bahasa tidak baku dapat menyebabkan sistem sulit mengenali makna sebenarnya dari suatu opini. Dalam mengatasi tantangan ini, *Support Vector Machine* (SVM) diakui efektif untuk klasifikasi data dalam analisis sentimen berkat kemampuannya menemukan *hyperplane* pemisah dengan jarak maksimum, sehingga memaksimalkan selisih antara kelas-kelas (Pritama dkk., 2024).

Penelitian sebelumnya menunjukkan bahwa SVM bekerja lebih efektif dibandingkan *Naive Bayes* pada analisis sentimen di X, dengan nilai akurasi sebesar 70,82 % berbanding

63,02 % (Ningsih dkk., 2024). Namun, pada penerapan SVM, seringkali ditemui permasalahan *imbalanced data*, yaitu kondisi di mana kelas mayoritas mendominasi sehingga mesin cenderung memprediksi kelas mayoritas dan mengakibatkan akurasi prediksi kelas minoritas yang rendah (Khausari, 2018). Untuk mengatasi hal ini, dikembangkan metode *resampling* yang terdiri atas *oversampling* dan *undersampling*. Namun, kedua teknik tersebut memiliki keterbatasan apabila dilakukan secara terpisah untuk klasifikasi *multi-class*, karena *oversampling* berpotensi menghasilkan *noisy instances*, sementara *undersampling* berisiko menghilangkan sub-konsep penting dari kelas mayoritas (Indrawati, 2021).

Untuk mengatasi kelemahan tersebut, penelitian ini mengusulkan metode *hybrid* yang mengombinasikan *Synthetic Minority Oversampling Technique* (SMOTE) dan *Cluster-Based Undersampling Technique*. Penerapan SMOTE terbukti dapat meningkatan akurasi SVM pada kelas minoritas yaitu sentimen positif dari 38% menjadi 84% sehingga menjadi seimbang dengan data mayoritas yaitu sentimen negatif dengan nilai akurasi 94% (Rahman Fauzan dkk., 2023). Sementara itu, *Cluster-Based Undersampling Technique* memberikan kinerja yang lebih baik dibandingkan metode *undersampling* lain seperti *ENN, Tomek Link*, dan *NearMiss* (Bach dkk., 2023). Kombinasi kedua metode tersebut juga terbukti efektif dalam menyeimbangkan data KSA yang menunjukkan bahwa metode ini mampu menghasilkan data sintetik yang memiliki karakteristik serupa dengan dataset asli serta mempertahankan kesamaan nilai rata-rata pada data numerik dengan tingkat ketidakseimbangan yang ekstrem (Sondriva dkk., 2024). Meskipun berbagai penelitian telah mengkaji penggunaan SMOTE dan teknik undersampling dalam menangani imbalanced data, sebagian besar masih menerapkan metode tersebut secara terpisah dan belum banyak mengeksplorasi kombinasi keduanya pada data teks dalam konteks analisis sentimen menggunakan SVM. Hal ini menunjukkan adanya celah penelitian terkait penerapan metode hybrid pada data teks yang tidak terstruktur.

Dengan demikian, penelitian ini difokuskan pada penerapan kombinasi SMOTE dan *Cluster-Based Undersampling Technique* untuk menangani *imbalanced data* pada analisis sentimen isu Rancangan Undang-Undang (RUU) Perampasan Aset menggunakan algoritma SVM. Isu ini dipilih karena data sentimen yang diperoleh dari media sosial X menunjukkan ketidakseimbangan yang signifikan antara kelas positif negatif dan netral, yang memiliki jumlah data pada salah satu kelas jauh mendominasi kelas lainnya. Kondisi ini menjadikan *dataset* tersebut sangat relevan untuk diterapkan pada penelitian mengenai metode penyeimbangan data. Melalui penerapan metode kombinasi ini, diharapkan dapat dihasilkan *dataset* yang lebih seimbang tanpa menghilangkan karakteristik data asli, sehingga model SVM mampu mempelajari pola sentimen masyarakat secara lebih akurat. Penelitian ini memberikan kontribusi nyata dalam pengembangan metode penanganan imbalanced data pada analisis sentimen dengan mengimplementasikan kombinasi SMOTE dan Cluster-Based Undersampling Technique pada SVM serta mengevaluasi peningkatan kinerja model secara empiris.

METODE PENELITIAN

Desain Penelitian

Penelitian ini bersifat kuantitatif dan menggunakan rancangan eksperimental komparatif. Pendekatan eksperimental digunakan untuk mengevaluasi penerapan metode *resampling* hibrida yang menggabungkan SMOTE dan *Cluster-Based Undersampling Technique* guna mengatasi ketidakseimbangan data, sedangkan pendekatan komparatif digunakan untuk

membandingkan kinerja model SVM sebelum dan sesudah penyeimbangan data, berdasarkan metrik evaluasi seperti akurasi, presisi, *recall*, dan *skor F1*.

Subjek

Subjek penelitian berupa 1.018 data *tweet* berbahasa Indonesia yang diperoleh dari media sosial X menggunakan teknik *scrapping* dengan alat bantu *Tweet Harvest* pada Python. Data dikumpulkan dengan kata kunci "RUU Perampasan Aset" dan ”Perampasan Aset” dalam periode waktu 1 Januari – 30 September 2025. Seluruh data kemudian dilabeli secara manual ke dalam tiga kategori sentimen, yaitu positif, negatif, dan netral.

Analisis Data

Analisis data dilakukan melalui serangkaian langkah komputasi menggunakan Python, yang mencakup langkah-langkah berikut:

Pengumpulan data

Pengumpulan data dilakukan dengan *scrapping* melalui *package tweet-harvest* pada Python. kata kunci yang digunakan yaitu "RUU Perampasan Aset" dan "Perampasan Aset" dalam rentang waktu 1 Januari hingga 30 September 2025. Proses *scrapping* menghasilkan 1018 *tweet.*

Preprocessing Data

Preprocessing data merupakan langkah penting dalam mempersiapkan data mentah untuk diproses oleh model SVM. Proses ini mencakup lima tahap utama berikut:

1. *Case Folding*, yaitu mengubah semua huruf dalam *tweet* menjadi huruf kecil untuk memastikan konsistensi teks;
2. *Cleaning*, yaitu menghapus data duplikat dan elemen yang tidak relevan seperti sebutan, tagar, dan tautan;
3. *Tokenizing*, yang melibatkan pemecahan teks menjadi unit kata-kata individual sehingga setiap bagian dapat dianalisis secara terpisah;
4. *Normalization*, yang melibatkan konversi kata-kata non-standar menjadi bentuk standar sesuai dengan aturan linguistik; dan
5. *Stemming*, yang melibatkan penghapusan afiks kata untuk mengembalikannya ke bentuk dasarnya (Fajriyah dkk., 2025).

TF-IDF

Dalam penelitian ini, ekstraksi fitur dilakukan dengan menggunakan metode TF-IDF *(Term Frequency–Inverse Document Frequency)*, yaitu teknik untuk memberi bobot pada kata-kata dalam suatu dokumen guna mengukur tingkat kepentingan suatu kata relatif terhadap keseluruhan korpus (Rifut Nur Mustaqim dkk., n.d.). Metode TF-IDF terdiri dari dua komponen utama *Term Frequency* (TF), yang menghitung frekuensi kemunculan suatu kata dalam satu dokumen, dan *Inverse Document Frequency* (IDF), yang menilai relevansi suatu kata berdasarkan distribusinya di seluruh dokumen. Nilai TF-IDF diperoleh dengan mengalikan kedua komponen ini, di mana semakin tinggi nilainya, semakin besar kontribusi kata tersebut terhadap proses klasifikasi sentimen (Putra., 2023) Pada penelitian ini digunakan skema N-gram (*unigram* dan *bigram*) dengan menghasilkan 5.000 fitur terbaik.

Pelabelan Data

Pelabelan data dilakukan secara manual oleh para penulis, yang mengelompokkan sentimen ke dalam tiga kategori: positif (dilabeli 1), negatif (dilabeli -1), dan netral (dilabeli 0). Sentimen positif diberikan pada cuitan yang mengandung dukungan terhadap RUU, sentimen negatif pada cuitan yang menunjukkan penolakan atau kritik, dan sentimen netral pada cuitan yang berisi informasi atau perkembangan berita. Hasil pelabelan selanjutnya divalidasi oleh

Dosen Bahasa Indonesia selaku ahli bahasa untuk memastikan kesesuaian makna dan keakuratan klasifikasi sentimen.

Split Data

Kumpulan data tersebut dibagi menjadi dua bagian utama dengan perbandingan 80:20, di mana 80% dialokasikan untuk kumpulan data *training* dan 20% untuk kumpulan data *testing*. Perbandingan ini dipilih karena dianggap optimal untuk mencapai keseimbangan antara pelatihan model yang komprehensif dan evaluasi yang akurat.

SMOTE dan *Cluster-Based Undersampling Technique*

SMOTE dan *Cluster-Based Undersampling Technique* merupakan metode *hybrid* yang dirancang untuk mengatasi ketidakseimbangan data dengan menyeimbangkan jumlah instansi setiap kelas ke nilai rata-rata (m) dari keseluruhan *dataset* (Fred, 2016). SMOTE bekerja dengan membangkitkan data sintetis pada kelas minoritas melalui interpolasi antara suatu instansi dan tetangga terdekatnya, sedangkan *Cluster-Based Undersampling Technique* mereduksi kelas mayoritas menggunakan pendekatan *clustering* agar setiap subkelompok tetap terwakili secara proporsional (Salehi & Khedmati, 2024).

Nilai rata-rata (m) dihitung sebagai:

$$\mathrm{m} = \frac{1}{k}\sum_{i=1}^{k} ni$$

dengan n_i= jumlah *instance* pada kelas ke-i dan k= jumlah total kelas. Proses penyeimbangan dilakukan dalam dua tahap:

1. *Oversampling* Kelas Minoritas
   Setiap kelas dengan $n_i < m$ ditambah menggunakan SMOTE hingga mencapai m. Data sintetis dibangkitkan melalui interpolasi:

   $$x_{new} = x_i + \delta \times (x_{nn} - x_i)$$

   dengan x_i= instansi minoritas ke-i, x_{nn}= *nearest neighbor* dari x_i, dan $\delta \in [0,1]$= bilangan acak penentu posisi interpolasi.
2. *Undersampling* Kelas Mayoritas
   Setiap kelas dengan $n_i > m$ dikurangi menggunakan *Cluster-Based Undersampling Technique* melalui tahapan berikut:

a. Kelas mayoritas dikelompokkan menggunakan algoritma *Expectation-Maximization* (EM) dengan jumlah cluster k=3k = 3k=3, iterasi maksimum 100, dan toleransi konvergensi sebesar 10−510^{-5}10−5, sehingga diperoleh klaster {C1,C2,…,Ck}\{C_1, C_2, \ldots, C_k\}{C1,C2,…,Ck}.

b. Proporsi tiap klaster dihitung sebagai:

$$p_j = \frac{n_j}{\sum_{l=1}^{k} n_l}$$

c. Jumlah instansi yang diambil dari klaster ke-j:

$$s_j = \lfloor p_j \times m \rfloor$$

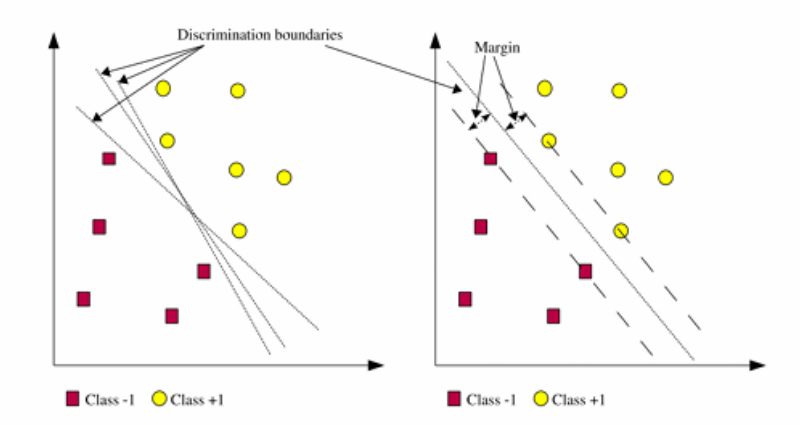
d. Sebanyak s_j instansi diambil secara acak dari setiap klaster dan digabungkan:

$$D_i' = \bigcup_{j=1}^{k} RandomSample(C_j, s_j)$$

Metode ini memastikan setiap subkelompok dalam kelas mayoritas tetap terwakili sehingga karakteristik penting data tidak hilang selama proses *undersampling* (Fred, 2016).

Support Vector Machine

Support Vector Machine (SVM) merupakan salah satu algoritma klasifikasi yang banyak digunakan dalam analisis sentimen karena kemampuannya dalam membedakan data ke dalam dua atau lebih kelas dengan tingkat akurasi yang tinggi. SVM bekerja dengan cara mencari *hyperplane* optimal yang berfungsi sebagai batas pemisah antara kelas-kelas dalam ruang fitur. *Hyperplane* ini menjadi garis keputusan yang membedakan satu kelas dari kelas lainnya (Fajriyah dkk., 2025). Prinsip dasar SVM terletak pada upaya memaksimalkan *margin* atau jarak antara *hyperplane* dan titik data terdekat dari masing-masing kelas. Titik data yang paling dekat dengan *hyperplane* disebut vektor pendukung, yang memainkan peran penting dalam menentukan posisi dan orientasi *hyperplane* pemisah. *Hyperplane* yang optimal adalah *hyperplane* yang terletak tepat di antara kedua kelas tersebut, sehingga menghasilkan batas klasifikasi yang paling efektif.



Gambar 1. SVM Temukan *Hyperplane* Terbaik

SVM bekerja dengan mencari *hyperplane* optimal yang memisahkan kelas-kelas data dengan margin terluas. Seperti yang ditunjukkan pada **Gambar 1**, sisi kiri menampilkan berbagai kemungkinan garis pemisah antara kedua kelas, sedangkan sisi kanan menggambarkan *hyperplane* optimal yang dipilih oleh SVM garis yang terletak tepat di antara kedua kelas dengan *margin* terluas. Titik-titik data yang paling dekat dengan *hyperplane* ini disebut vektor pendukung, yang memainkan peran penting dalam menentukan posisi pemisahan.

Namun, tidak semua data dapat dipisahkan secara linier. Untuk menangani pola data yang kompleks, SVM menggunakan fungsi kernel untuk memetakan data ke ruang fitur berdimensi lebih tinggi. (Fajriyah dkk., 2025):

1. Kernel Linear

   Kernel linier digunakan ketika data dapat dipisahkan secara linier tanpa perlu mentransformasikannya ke ruang berdimensi lebih tinggi. Kernel ini bekerja dengan menghitung hasil kali skalar antara dua vektor data dan sangat cocok untuk kumpulan data berdimensi tinggi yang menunjukkan sifat-sifat linier.

   $$K(x_i, x_j) = x_i \cdot x_j$$

2. Kernel Polinomial

   Kernel polinomial memudahkan pemetaan data ke ruang berdimensi lebih tinggi melalui fungsi berbasis pangkat. Kernel ini sangat cocok untuk data yang tidak sepenuhnya linier, namun menunjukkan pola interaksi antar fitur yang dapat dimodelkan menggunakan transformasi polinomial.

   $$K(x_i, x_j) = (x_i \cdot x_j + 1)^p$$

3. Kernel RBF (*Radial Basis Function)*
   Kernel RBF adalah kernel nonlinier yang memperhitungkan jarak *euclidean* antara dua titik data. Fungsi ini sangat fleksibel dan sering digunakan karena dapat menangani data yang memiliki pola nonlinier yang kompleks. Parameter γ mengontrol lebar kernel:

- Nilai γ besar →model lebih kompleks, rentan *overfitting*.
- Nilai γ kecil → model lebih halus, generalisasi lebih baik.

$$K(x_i, x_j) = \exp\left(-\gamma \|x_i - x_j\|^2\right)$$

4. Kernel Sigmoitd
   Kernel sigmoid menggunakan fungsi aktivasi *hyperbolic* tangent *(tanh)* yang mirip dengan yang digunakan pada jaringan saraf tiruan. Kernel ini cocok untuk data nonlinier dan memberikan fleksibilitas dalam membentuk batas keputusan yang lebih adaptif.

$$K(x_i, x_j) = tanh(\eta x_i \cdot x_j + v)$$

   Dengan:
   - η = parameter skala
   - v = konstanta bias

   Dalam penelitian ini, kernel *Radial Basis Function* (RBF) dipilih karena kemampuannya dalam mengolah data nonlinier melalui pemetaan ke ruang berdimensi tinggi. Penelitian yang dilakukan oleh Fajriyah dkk. (2025) menunjukkan bahwa kernel RBF mencapai akurasi tertinggi sebesar 93%, mengungguli kernel linier, polinomial, dan sigmoid.

Untuk menangani klasifikasi *multi-class*, digunakan metode SVM *multi-class* dengan pendekatan *One Vs Rest* (OVR). Metode ini membangun model biner sebanyak jumlah kelas, kemudian menggabungkannya menjadi satu model utama. Pemilihan OVR didukung oleh penelitian Dasriani dkk. (2024) yang menunjukkan performa OVR lebih baik dibandingkan OVO.

Evaluasi Model Klasifikasi

Evaluasi model adalah proses penilaian terhadap kemampuan model dalam mengenali pola dan memprediksi kelas data dengan tingkat akurasi yang tinggi. Proses ini sangat penting untuk memastikan bahwa algoritma yang digunakan menghasilkan prediksi yang optimal. Dalam penelitian ini, evaluasi dilakukan dengan menggunakan matriks kebingungan, yaitu metode yang membandingkan prediksi model dengan label sebenarnya pada data uji. Dalam matriks kebingungan, baris mewakili kelas sebenarnya, sedangkan kolom mencerminkan kelas yang diprediksi oleh model. Nilai *True Positive* (TP) menunjukkan jumlah titik data kelas positif yang diklasifikasikan dengan benar, *True Negative* (TN) untuk titik data kelas negatif yang diprediksi dengan benar, dan *True Neutral* (TNt) untuk titik data kelas netral yang diklasifikasikan dengan benar. Sebaliknya, *False Positive* (FP) mengacu pada data kelas positif

yang salah diklasifikasikan, *False Negative* (FN) untuk data kelas negatif yang salah diprediksi, dan *False Neutral* (FNt) untuk data kelas netral yang tidak sesuai dengan kelas aslinya.

Tabel 1. *Confusion Matrix*

Actual	Klasifikasi		
	Positif	**Negatif**	**Netral**
Positif	*True Positif* (TP)	*False Negatif* (FN)	*False Neutral* (FNt)
Negatif	*False Positif* (FP)	*True Negatif* (TN)	*False Neutral* (FNt)
Netral	*FalsePositif* (FP)	*False Negatif* (FN)	*TrueNeutral* (TNt)

Berdasarkan hasil *confusion matrix* diperoleh empat metrik evaluasi sebagai berikut:

1. *Accuracy*

   Hitung rasio prediksi yang benar terhadap jumlah total titik data uji.

$$Accuracy = \frac{(TP + TN + TNt)}{(TP + TN + TNt + FP + FN + FNt)} \times 100\%$$

2. *Precision*

   Menunjukkan ketepatan model dalam memprediksi kelas positif.

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

3. *Recall*

   Menilai kemampuan model dalam mengidentifikasi semua data positif secara akurat.

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

4. *F1-Score*

   Menujukkan keseimbangan anata presisi dan *recall*.

$$F1 - Score = 2 \times \frac{(Precision \times Recall)}{(Precision + Recall)}$$

HASIL DAN PEMBAHASAN

Hasil Penelitian

Pengumpulan data

Data penelitian ini berupa *tweet* dari pengguna media sosial X yang dipublikasikan dalam rentang waktu 1 Januari hingga 30 September 2025. Pengumpulan data dilakukan menggunakan *package tweet-harvest* pada Python dengan kata kunci "RUU Perampasan Aset" dan "Perampasan Aset". Proses pemungutan data menghasilkan total 1.018 cuitan yang kemudian disimpan dalam format Microsoft Excel. Berikut merupakan hasil pengambilan data.

Tabel 2. Contoh Hasul Pengumpulan Data

	Cuitan
1.	@MoodNetizen dulu tu emng berapa lama sih baru beberapa bulan bikin aturan ini juga... gila ya ngerampas hak rakyat buru-buru amat uu perampasan aset tak kunjung keliatan
2.	@psi_id Sepakat gak kata Wamen bukan UU perampasan aset tapi UU pemulihan aset

3.	@txtdrimedia Di sini ga takut mati cuma takut miskin. Perampasan aset cuma omon omon

Preprocessing Data

Hasil dari *preprocessing* data sebagai berikut

Tabel 3. Hasil Masing-masing Tahapan *Preprocessing* data

Proses	*Input*	*Output*
Case Folding	@MoodNetizen dulu tu emng berapa lama sih baru beberapa bulan bikin aturan ini juga... gila ya ngerampas hak rakyat buru2 amat uu perampasan asset tak kunjung keliatan	@moodnetizen dulu tu emng berapa lama sih baru beberapa bulan bikin aturan ini juga... gila ya ngerampas hak rakyat buru2 amat uu perampasan asset tak kunjung keliatan
Cleaning	@moodnetizen dulu tu emng berapa lama sih baru beberapa bulan bikin aturan ini juga... gila ya ngerampas hak rakyat buru2 amat uu perampasan asset tak kunjung keliatan	dulu tu emng berapa lama sih baru beberapa bulan bikin aturan ini juga gila ya ngerampas hak \ rakyat buru amat uu perampasan asset tak kunjung keliatan
Tokenizing	dulu tu emng berapa lama sih baru beberapa bulan bikin aturan ini juga gila ya ngerampas hak \ rakyat buru amat uu perampasan asset tak kunjung keliatan	[’dulu’, ’tu’, ’emng’, ’berapa’, ’lama’, ’sih’, ’baru’, ’beberapa’, ’bulan’, ’bikin’, ’aturan’, ’ini’, ’juga’, ’gila’, ’ya’, ’ngerampas’, ’hak’, ’rakyat’, ’buru’, ’amat’, ’uu’, ’perampasan’, ’aset’, ’tak’, ’kunjung’, ’keliatan’]
Normalization	[’dulu’, ’tu’, ’emng’, ’berapa’, ’lama’, ’sih’, ’baru’, ’beberapa’, ’bulan’, ’bikin’, ’aturan’, ’ini’, ’juga’, ’gila’, ’ya’, ’ngerampas’, ’hak’, ’rakyat’, ’buru’, ’amat’, ’uu’, ’perampasan’, ’aset’, ’tak’, ’kunjung’, ’keliatan’]	[’dulu’, ’tu’, ’memang’, ’berapa’, ’lama’, ’sih’, ’baru’, ’beberapa’, ’bulan’, ’bikin’, ’aturan’, ’ini’, ’juga’, ’gila’, ’iya’, ’ngerampas’, ’hak’, ’rakyat’, ’buru’, ’amat’, ’uu’, ’perampasan’, ’aset’, ’tidak’, ’kunjung’, ’kelihatan’]
Stemming	[’dulu’, ’tu’, ’memang’, ’berapa’, ’lama’, ’sih’, ’baru’, ’beberapa’, ’bulan’, ’bikin’, ’aturan’, ’ini’, ’juga’, ’gila’, ’iya’, ’ngerampas’, ’hak’, ’rakyat’, ’buru’, ’amat’, ’uu’, ’perampasan’, ’aset’, ’tidak’, ’kunjung’, ’kelihatan’]	[’dulu’, ’tu’, ’memang’, ’berapa’, ’lama’, ’sih’, ’baru’, ’beberapa’, ’bulan’, ’bikin’, ’atur’, ’ini’, ’juga’, ’gila’, ’iya’, ’ngerampas’, ’hak’, ’rakyat’, ’buru’, ’amat’, ’uu’, ’ampas’, ’aset’, ’tidak’, ’kunjung’, ’lihat’]

Pelabelan Data

Berdasarkan hasil pelabelan manual yang telah divalidasi oleh validator ahli Bahasa Indonesia, diperoleh distribusi label sebagaimana disajikan pada Tabel 4.

Tabel 4. Contoh Hasil Pelaelan dan Validasi Data

Cuitan	Label Awal	Label Validasi	Catatan
dulu tu memangberapamanasih baru beberapa bulan bikin atur ini juga gila iya ngerampas hak rakyat buru amat uu rampasaset tidak kunjung lihat	-1	-1	-
sepakat tidak kata wamen bukan uurampasaset tapi uu pulih aset	1	-1	Berubah karena cuitan merujuk

			pada dukungan pemulihan aset bukan perampasan aset
masalah bayar bayar untuk adil itu sudah jadi rahasia umum cumajumlahnya yang tidak umummakanyaadaistilah maia adil tidak terbayang jk uu rampas aset di sahkan mk akanjadi ladang kerja baru untuk maia adil	1	-1	Berubah karena cuitan mengandung maknayang merujuk pada penolakan

Berikut merupakan hasil distribusi label sentimen hasil validasi

Tabel 5. Distribusi Label Sentimen Hasil Validasi

Label Sentimen	Jumlah Data	Persentase	Keterangan
Positif	662	65.03%	Mayoritas
Negatif	160	15.72%	Minoritas
Netral	196	19.25%	Minoritas

Berdasarkan **Tabel 5**, label positif mendominasi *dataset* dengan 662 cuitan (65,03%), diikuti netral 196 cuitan (19,25%), dan negatif 160 cuitan (15,72%). Ketimpangan distribusi antar kelas ini mengindikasikan adanya ketidakseimbangan data yang berpotensi membiaskan model klasifikasi ke arah kelas mayoritas.

TF-IDF

NO.	abai anggota	abal	abal punya	....	zalim lenyap	zonauang	zonauang bioremicellarxbt
1	0	0	0	....	0	0	0
2	0	0	0	....	0	0	0
3	0	0	0	....	0	0	0
4	0	0	0	....	0	0	0
5	0	0	0	....	0	0	0
6	0	0	0	....	0	0	0
...	.....	.....	.....	.....	.....	.....	.....
...	.....	.....	.....	.....	.....	.....	.....
1001	0	0	0	....	0	0	0
1002	0	0	0	....	0	0	0
1003	0	0	0	....	0	0	0
1004	0	0	0	....	0	0	0

Gambar 2. Hasil Ekstraksi Fitur TF-IDF

Pembagian Data

Setelah ekstraksi fitur, dataset dibagi menggunakan rasio 80:20 di mana 80% data (814 cuitan) digunakan sebagai data latih dan 20% data (204 cuitan) digunakan sebagai data uji.

Penyeimbangan Data dengan SMOTE dan *Cluster-Based Undersampling Technique*

Penerapan metode *hybrid* SMOTE dan *Cluster-Based Undersampling Technique* pada data latih (814 sampel) menghasilkan distribusi kelas yang seragam sebagaimana disajikan pada Tabel 6

Tabel 6. Hasil Penyeimbangan Data

Label	Sebelum	Metode	Sesudah
Positif	523	*Cluster-Based Undersampling*	271
Negatif	124	SMOTE *(Oversampling)*	271
Netral	167	SMOTE *(Oversampling)*	271
Total	**814**		**813**

Klasifikasi *Support Vector Machine*

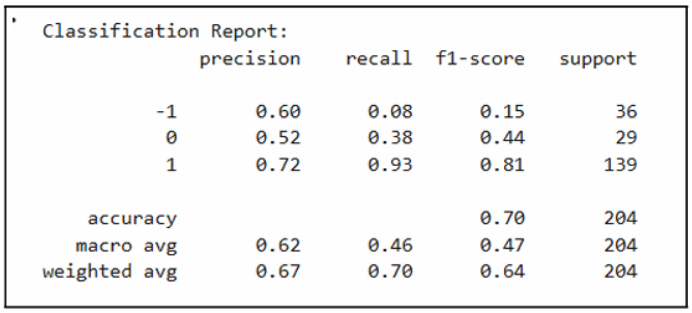
Klasifikasi sentimen dilakukan menggunakan algoritma SVM dengan fungsi kernel RBF dan pendekatan *multi-class One-vs-Rest* (OVR). Pengujian dilakukan dalam dua skema untuk membandingkan performa: (1) klasifikasi menggunakan *dataset* asli tanpa penyeimbangan data, dan (2) klasifikasi menggunakan dataset yang telah diseimbangkan dengan metode *hybrid* SMOTE dan *Cluster-Based Undersampling Technique*.

1. ***Confusion Matrix* SVM Tanpa SMOTE dan *Cluster-Based Undersampling Technique***

Tabel 7. *Confusion Matrix* SVM Tanpa Penyeimbangan Data

Label	Negatif	Netral	Positif
Negatif	3	2	31
Netral	0	11	18
Positif	2	8	129

Berdasarkan Tabel 7, nilai-nilai diperoleh dan digunakan untuk menghitung metrik evaluasi model. Hasil evaluasi kinerja model SVM tanpa SMOTE dan *Cluster-Based Undersampling Technique* berdasarkan nilai akurasi, presisi, *recall*, dan *F1-score* disajikan pada Gambar 3.



Classification Report:

	precision	recall	f1-score	support
-1	0.60	0.08	0.15	36
0	0.52	0.38	0.44	29
1	0.72	0.93	0.81	139
accuracy			0.70	204
macro avg	0.62	0.46	0.47	204
weighted avg	0.67	0.70	0.64	204

Gambar 3. Hasil Evaluasi Klasifikasi SVM tanpa SMOTE dan *Cluster-Based Undersampling Technique*

2. ***Confusion Matrix* SVM dengan SMOTE dan *Cluster-Based Undersampling Technique***

Tabel 8. *Confusion Matrix* SVM Dengan Penyeimbangan Data

Label	Negatif	Netral	Positif
Negatif	53	4	12
Netral	1	48	13
Positif	2	5	66

Berdasarkan Tabel 8, nilai-nilai diperoleh dan digunakan untuk menghitung metrik evaluasi model. Hasil evaluasi kinerja model SVM dengan SMOTE dan *Cluster-Based Undersampling Technique* berdasarkan nilai akurasi, presisi, *recall*, dan *F1-score* disajikan pada Gambar 4.

Laporan Detail:

	precision	recall	f1-score	support
-1	0.95	0.77	0.85	69
0	0.84	0.77	0.81	62
1	0.73	0.90	0.80	73
accuracy			0.82	204
macro avg	0.84	0.82	0.82	204
weighted avg	0.84	0.82	0.82	204

Gambar 4. Hasil Evaluasi Klasi ikasi SVM dengan SMOTE dan *Cluster-Based Undersampling Technique*

3. **Perbandingan Kinerja Model**

Tabel 9. Perbandingan Kinerja Model SVM

Model	Akurasi	Presisi	*Recall*	*F1-Score*
SVM tanpa SMOTE dan *Cluster-Based Undersampling Technique*	70%	62%	46%	47%
SVM dengan SMOTE dan *Cluster-Based Undersampling Technique*	82%	84%	82%	82%

Berdasarkan **Tabel 9**, model SVM tanpa penyeimbangan data menghasilkan akurasi 70% dengan presisi 62%, *recall* 46%, dan *F1-score* 47%. Nilai *recall* yang rendah, khususnya pada kelas negatif yang hanya mencapai 8%, menunjukkan bahwa model gagal mendeteksi sebagian besar sampel minoritas.

Setelah penerapan SMOTE dan *Cluster-Bsed Undersampling Technique*, akurasi meningkat menjadi 82%, presisi 84%, *recall* 82%, dan *F1-score* 82%. Peningkatan recall kelas negatif dari 8% menjadi 77% mengindikasikan bahwa model mengalami penurunan bias terhadap kelas mayoritas. Hal ini menunjukkan bahwa penerapan metode hybrid SMOTE dan Cluster-Based Undersampling Technique mampu meningkatkan sensitivitas model dalam mengenali kelas minoritas tanpa menurunkan performa keseluruhan secara signifikan.

Pembahasan

Hasil penelitian menunjukkan bahwa dominasi sentimen positif (65,03%) mencerminkan bahwa sebagian besar pengguna media sosial X memberikan dukungan terhadap isu RUU Perampasan Aset selama periode Januari hingga September 2025. Temuan ini konsisten dengan studi sebelumnya dalam analisis sentimen media sosial yang menunjukkan bahwa kebijakan publik yang berkaitan dengan pemberantasan korupsi cenderung memperoleh respons positif dari masyarakat. Di sisi lain, keberadaan sentimen negatif sebesar 15,72% mengindikasikan adanya kelompok yang menolak atau meragukan efektivitas regulasi tersebut, khususnya terkait kekhawatiran mengenai potensi penyalahgunaan wewenang oleh aparat penegak hukum. Hal ini menunjukkan bahwa distribusi sentimen tidak sepenuhnya homogen dan mencerminkan adanya perbedaan persepsi publik terhadap kebijakan yang diusulkan.

Perbandingan kedua skema membuktikan bahwa, pada skema pertama tanpa model penyeimbangan data SMOTE dan *Cluster-Based Undersampling Technique* model SVM mengahsilkan akurasi sebesar 70%. Namun, metrik evaluasi lainnya menunjukkan performa yang kurang memuaskan dengan presisi 62% *recall* 46% dan *f1-score* 47%. Rendahnya perolehan *recall*, khusunya pada kategori negatif yang hanya menyentuh angka 8%, mengindikasikan bahwa model gagal melakukan generalisasi pada kelas-kelas dnegan sampel terbatas. Kondisi ini memperlihatkan bahwa meskipun nilai aurasi terlihat mencukupi, model sebenarnya terjebak pada kecenderungan untuk memprediksi label positif secara berlebihan, sehingga mengabaikan variansi sentimen lainnya.

Sebaliknya, pada pengujian skema kedua dengan penerapan SMOTE dan *Cluster-Based Undersampling Technique* menunjukkan eskalasi performa yang signifikan, dimana akurasi meningkat menjadi 82%. Capaian presisi sebesar 84% dan *recall* sebesar 82% menandakan bahwa model kini memiliki sensitivitas yang seimbang dalam mendeteksi tiap kategori sentimen. Peningkatan *f1-score* menjadi 82% menajdi bukti bahwa model tidak lagi terpengaruh oleh dominasi kelas tertentu. Hal ini mencerminkan bahwa SVM dan metode penyimbangan data *hybrid* SMOTE dan *Cluster-Based Undersampling Technique* mampu menghasilkan sistem klasifikasi yang lebih presisi dan memiliki keampuan deteksi yang konsisten pada seluruh kelas sentimen.

Implikasi

Temuan Penelitian ini dapat memberikan wawasan mengenai penerapan metode SMOTE dan *Cluster-Based Undersampling Technique* dalam menangani *imbalanced data* pada klasifikasi sentimen menggunakan SVM, sekaligus memperkaya kajian ilmiah di bidang data mining dan analisis sentimen terkait kombinasi teknik *oversampling* dan *undersampling* untuk meningkatkan performa model berbasis teks. Secara praktis, temuan penelitian ini dapat menjadi acuan bagi para peneliti, lembaga penelitian, dan instansi pemerintah untuk memahami persepsi masyarakat terhadap isu-isu kebijakan publik, khususnya rancangan undang-undang perampasan aset. Selain itu, temuan ini juga dapat menjadi panduan bagi para praktisi ilmu data dalam mengatasi ketidakseimbangan data pada analisis teks media sosial, sehingga klasifikasi yang dihasilkan menjadi lebih akurat dan representatif.

Keterbatasan dan Rekomendasi Penelitian Lanjutan

Penelitian ini memiliki keterbatasan yaitu data yang digunakan terbatas pada cuitan dari rentang waktu 1 Januari hingga 30 September 2025 dengan jumlah 1.018 sampel, sehingga generalisasi model terhadap dinamika opini publik pada periode atau isu kebijakan lain perlu dilakukan secara hati-hati. Berdasarkan keterbatasan tersebut, penelitian selanjutnya disarankan untuk memperluas rentang waktu dan jumlah data agar generalisasi model terhadap isu-isu nasional menjadi lebih kuat.

SIMPULAN

Berdasarkan hasil penelitian dan pembahasan, dapat ditarik tiga kesimpulan. Pertama, metode hybrid SMOTE dan Cluster-Based Undersampling Technique menunjukkan kinerja yang lebih baik dalam menangani imbalanced data pada dataset penelitian ini dengan menyetarakan distribusi sampel berdasarkan nilai rata-rata sebesar 227 data per kelas. Pendekatan ini

berkontribusi terhadap peningkatan performa model SVM, khususnya dalam meningkatkan kemampuan model dalam mengenali kelas minoritas secara lebih seimbang. Kedua, setelah penyeimbangan data, algoritma SVM dengan kernel RBF menghasilkan performa yang stabil dan objektif dengan akurasi 82%, presisi 84%, *recall* 82%, dan *F1-score* 82%, mengindikasikan adanya peningkatan kemampuan model dalam mengenali kelas minoritas setelah penerapan metode penyeimbangan data. Sementara itu, stabilitas performa pada metrik agregat serta peningkatan nilai macro average menunjukkan bahwa distribusi kinerja antar kelas menjadi lebih seimbang. Temuan ini mengindikasikan bahwa metode hybrid SMOTE dan Cluster-Based Undersampling Technique mampu meningkatkan sensitivitas model terhadap kelas minoritas tanpa menurunkan performa pada kelas mayoritas secara signifikan. Ketiga, perbandingan performa menunjukkan peningkatan signifikan setelah penerapan metode penyeimbangan, yakni akurasi dari 70% menjadi 82%, presisi dari 62% menjadi 84%, *recall* dari 46% menjadi 82%, dan *F1-score* dari 47% menjadi 82%, dengan peningkatan *recall* kelas negatif dari 8% menjadi 83% secara *macro average* membuktikan bahwa kombinasi SMOTE dan *Cluster-Based Undersampling Technique* berkontribusi besar dalam meningkatkan reliabilitas model untuk mendeteksi seluruh kategori sentimen pada isu RUU Perampasan Aset.

DAFTAR PUSTAKA

Bach, M., Trofimiak, P., Kostrzewa, D., & Werner, A. (2023). CLEANSE—Cluster-based undersampling method. *Procedia Computer Science, 225*, 4541–4550. https://doi.org/10.1016/j.procs.2023.10.452

Fajriyah, N., Lapatta, N. T., Nugraha, D. W., & Laila, R. (2025). Implementasi SVM dan SMOTE pada analisis sentimen media sosial X terhadap pelantikan Agus Harimurti Yudhoyono. *JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika), 10*(2), 1359–1370. https://doi.org/10.29100/jipi.v10i2.6246

Fred, A. (Ed.). (2016). *IC3K 2015: Proceedings of the 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (Vol. 3, KMIS): Lisbon, Portugal, November 12–14, 2015*. SCITEPRESS - Science and Technology Publications.

Handayani, A., & Zufria, I. (2023). Analisis sentimen terhadap bakal capres RI 2024 di Twitter menggunakan algoritma SVM. *Journal of Information System Research (JOSH), 5*(1), 53–63. https://doi.org/10.47065/josh.v5i1.4379

Indra, M., Arsya, H., Akbari, D., Novianty, A., & Setianingsih, C. (n.d.). Analisis sentimen menggunakan metode learning vector quantization.

Indrawati, A. (2021). Penerapan teknik kombinasi oversampling dan undersampling untuk mengatasi permasalahan imbalanced dataset. *Jurnal Informatika dan Komputer, 4*(1). https://doi.org/10.33387/jiko

Larassetya, T. D., Suryasuciramdhan, A., Salsa, N. U., & Aeni, I. S. (2024). Analisis opini publik terhadap Pemilu 2024 pada media sosial X. *TUTURAN: Jurnal Ilmu Komunikasi, Sosial dan Humaniora, 2*(2), 292–301. https://doi.org/10.47861/tuturan.v2i2.994

Mujilahwati, S. (2016). Pre-processing text mining pada data Twitter. Dalam *Prosiding Seminar Nasional Teknologi Informasi dan Komunikasi*.

Mustaqim, E. R. N., Pagalay, U., & Crysdian, C. (n.d.). Prediksi tingkat kepercayaan masyarakat terhadap Pilpres 2024 menggunakan TF-IDF dan BoW menggunakan metode SVM.

Ningsih, W., Alfianda, B., Rahmaddeni, R., & Wulandari, D. (2024). Perbandingan algoritma SVM dan Naïve Bayes dalam analisis sentimen Twitter pada penggunaan mobil listrik di Indonesia. *MALCOM: Indonesian Journal of Machine Learning and Computer Science, 4*(2), 556–562. https://doi.org/10.57152/malcom.v4i2.1253

Pritama, F., Leluni, E. R. D., Parhusip, J., & Universitas Palangka Raya. (2024). Analisis distribusi kinerja SVM dan KNN berdasarkan rata-rata simpangan baku dan stabilitas. *Jurnal Ilmiah Informatika dan Komputer, 1*(2), 170–174.

Putra, K. T. (2023). *Analisis feature extraction pada text processing untuk analisis sentimen*.

Rahman Fauzan, M., Oktafia Lingga Wijaya, H., & Karman, J. (2023). Analisis sentimen masyarakat terhadap kenaikan harga BBM di media sosial Twitter menggunakan metode support vector machine. Dalam *Seminar Riset Mahasiswa-Computer & Electrical (SERIMA-CE), 1*(1).

Salehi, A. R., & Khedmati, M. (2024). A cluster-based SMOTE both-sampling (CSBBoost) ensemble algorithm for classifying imbalanced data. *Scientific Reports, 14*(1). https://doi.org/10.1038/s41598-024-55598-1

Sondriva, W., Kurniawati, Y., Amalita, N., & Salma, A. (2024). Penanganan ketidakseimbangan multikelas pada dataset survei kerangka sampel area menggunakan metode SCUT. *UNP Journal of Statistics and Data Science, 2*(2), 159–164. https://doi.org/10.24036/ujsds/vol2-iss2/163