Institut Riset dan Publikasi Indonesia (IRPI)
MALCOM: Indonesian Journal of Machine Learning and Computer Science
Journal Homepage: https://journal.irpi.or.id/index.php/malcom
Vol. 5 Iss. 3 July 2025, pp: 981-989
ISSN(P): 2797-2313 | ISSN(E): 2775-8575

Comparison of Naïve Bayes and Support Vector Machine in Sentiment Analysis of Confiscation of Corrupt Assets on Twitter

Perbandingan *Naïve Bayes* dan *Support Vector Machine* Dalam Analisa Sentimen Tentang Penyitaan Aset Koruptor di Twitter

Ananda Sholekhah[1*], Muntahanah[2]

[1,2]Program Studi Teknik Informatika, Universitas Muhammadiyah Bengkulu, Indonesia

E-Mail: [1]anandasholekhah1@gmail.com, [2]muntahanah@umb.ac.id

Received Apr 28th 2025; Revised Jul 17th 2025; Accepted Jul 21th 2025; Available Online Jul 31th 2025, Published Jul 31th 2025
Corresponding Author: Ananda Sholekhah

Abstract

Sentiment analysis is one approach used to understand public perception of social issues and government policies through textual data. This study examines Indonesian public opinion toward President Prabowo Subianto's statement regarding the confiscation of corruptors' assets, which questioned, "Is it fair for the children to suffer?". The dataset was collected from Twitter, comprising 1,561 tweets gathered between April 9 and April 25, 2025, using relevant keywords. The analysis involved preprocessing, TF-IDF vectorization, and classification using Naïve Bayes and Support Vector Machine (SVM) algorithms. Model performance was evaluated using a confusion matrix and four evaluation metrics: accuracy, precision, recall, and F1-score. The results indicate that SVM outperformed Naïve Bayes, achieving an accuracy of 70.51% and an F1-score of 0.69, while Naïve Bayes recorded 66.34% accuracy and 0.66 F1-score. Most of the sentiments were classified as positive, reflecting majority public support for the asset confiscation policy despite its impact on the perpetrators' families. This research demonstrates the effectiveness of machine learning-based classification in mapping public opinion on controversial policy issues through social media platforms.

Keyword: Asset Forfeiture, Corruptors, Naïve Bayes, Sentiment Analysis, Support Vector Machine

Abstrak

Analisis sentimen merupakan salah satu pendekatan dalam memahami persepsi publik terhadap isu-isu sosial dan kebijakan pemerintah melalui data teks. Penelitian ini mengkaji opini masyarakat Indonesia terhadap pernyataan Presiden Prabowo Subianto mengenai penyitaan aset koruptor yang berbunyi "Apakah adil anaknya menderita?". Data dikumpulkan dari Twitter sebanyak 1.561 *tweet* dalam rentang waktu 9 hingga 25 April 2025 dengan menggunakan kata kunci yang relevan. Proses analisis dilakukan melalui tahap prapemrosesan, pembobotan *TF-IDF*, dan klasifikasi menggunakan algoritma *Naïve Bayes* dan *Support Vector Machine* (*SVM*). Evaluasi performa model dilakukan menggunakan *confusion matrix* serta empat metrik evaluasi, yaitu akurasi, presisi, *recall*, dan *F1-score*. Hasil menunjukkan bahwa *SVM* unggul dengan akurasi 70,51% dan *F1-score* 0,69, sedangkan *Naïve Bayes* memperoleh akurasi 66,34% dan *F1-score* 0,66. Sentimen terbanyak berasal dari kelas positif, mengindikasikan mayoritas publik mendukung penyitaan aset koruptor meskipun berdampak pada keluarganya. Penelitian ini memperlihatkan efektivitas pendekatan *machine learning* dalam memetakan opini publik terhadap isu kebijakan kontroversial di media sosial.

Kata Kunci: Analisis Sentimen, Koruptor, *Naïve Bayes*, Penyitaan Aset, *Support Vector Machine*.

1. PENDAHULUAN

Metode yang digunakan bagi mengetahui opini publik terhadap suatu topik, seperti isu-isu sosial, kebijakan pemerintah, maupun layanan publik ialah analisis sentimen yang dimanfaatkan sebagai sarana evaluasi rancangan kebijakan yang telah diusulkan oleh presiden Prabowo Subianto terhadap penyitaan *asset* koruptor, dengan cara mengumpulkan pandangan publik lewat platform daring terutama pada platform Twitter [1]. Twitter merupakan satu diantara platform daring sangat sering yang dipakai oleh warga Indonesia bagi menyampaikan pendapat politik mereka secara terbuka dan *real-time*.

Pernyataan Prabowo mengenai penyitaan aset koruptor “Apakah adil anaknya menderita?” menimbulkan perdebatan luas pada platform digital. Penelitian saat ini dilakukan demi menganalisis sentimen publik mengenai isu yang dimaksud dengan menerapkan metode analisis sentimen berbasis *machine learning*, sebagaimana *Naïve Bayes* serta *Support Vector Machine (SVM)*. Data diperoleh lewat Twitter selama periode 9 hingga 25 April 2025 menggunakan kata kunci yang relevan, dan diproses melalui tahapan prapemrosesan dan representasi *TF-IDF* [2].

Penelitian terdahulu menunjukkan bahwa analisis sentimen terhadap isu-isu politik di media sosial mampu merepresentasikan persepsi masyarakat dan memberikan indikasi terhadap kemungkinan pengaruhnya terhadap arah kebijakan pemerintah [3]. Namun, hingga saat ini belum ada penelitian komprehensif yang menganalisis sentimen masyarakat khusus terhadap dilema moral antara penegakan hukum dan dampak tidak langsung terhadap anggota keluarga yang tidak bersalah, terutama dalam konteks pernyataan Prabowo tersebut.

Berbagai studi telah mengimplementasikan penilaian sentimen terkait isu sosial melalui metode *Naïve Bayes* dan *SVM* Raniya dan Nuri [4]. Dalam penelitiannya terhadap opini publik mengenai kinerja pemerintahan memakai model pembelajaran mesin *Naïve Bayes* dan *SVM* menunjukkan bahwa tingkat akurasi mencapai 82% dimiliki oleh *SVM* yang merupakan akurasi tertinggi. Sedangkan *Naive Bayes* memperoleh akurasi sebesar 80%.

Studi serupa oleh Rima et al [5] terkait dengan isu vaksinasi *COVID-19* di Twitter, hasil penelitian juga Menandakan bahwa algoritma *SVM* memberikan hasil terbaik dari segi akurasi 83,33% jika dibandingkan dengan *Naïve Bayes* yang memiliki akurasi 70,00%. Penelitian oleh Muhammet et al. [6] juga menyimpulkan bahwa *SVM* menghasilkan akurasi lebih tinggi dalam analisis sentimen media sosial secara *real-time*. Sementara itu, studi oleh Indraet al. [7] pada data *Twitter Marketplace* menemukan bahwa *SVM* lebih efektif daripada *Naïve Bayes*, khususnya terkait menyelesaikan penyaluran kelas yang mana belum proporsional.

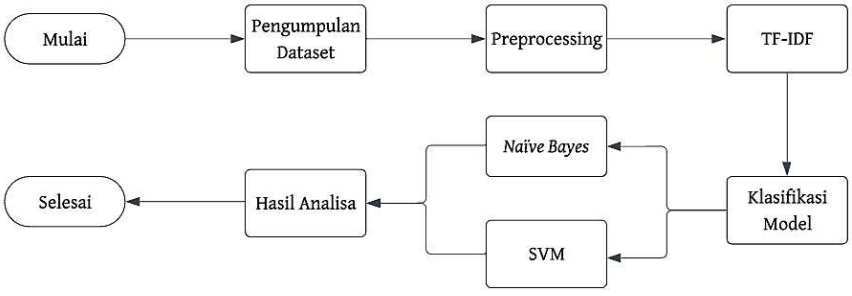
Meskipun demikian, masih sangat sedikit penelitian yang menerapkan kedua algoritma ini dalam konteks isu yang memiliki dimensi etis serta moral yang kompleks, seperti penyitaan aset koruptor. Berdasarkan kesenjangan penelitian sebelumya, tujuan dari studi ini bagi mengevaluasi sentimen (positif, negatif, netral) terhadap pernyataan Prabowo terkait penyitaan aset koruptor serta dampaknya terhadap keluarga koruptor, memanfaatkan algoritma *Naïve Bayes* serta *SVM* sebagai Algoritma klasifikasi .

Dalam Riset ini, informasi dikumpulkan dari media sosial Twitter memakai kata kunci seperti "aset koruptor", "anak koruptor", serta "penyitaan aset", dalam rentang waktu 9 April 2025 hingga 25 April 2025. Data ini mencerminkan opini masyarakat terhadap pernyataan Prabowo serta memungkinkan analisis sentimen dilakukan secara sistematis.

Penelitian ini diarahkan bagi mendalami aspek-aspek terkait perasan masyarakat terkait politik di platform Twitter, dengan memperhatikan aspek-aspek seperti *accuracy, precision, recall, F1-Score* serta *confusion matrix*, kajian ini diharapkan sanggup memjadi bantuan bagi penelitian selanjutnya dalam memilih klasifikasi yang paling efektif antara *Naïve Bayes* serta *SVM* Setiap algoritma juga dioptimalkan melalui *SMOTE* serta *hyperparameter tuning* bagi mencapai performa yang terbaik. Hasil studi ini dapat berfungsi sebagai referensi bagi pemilihan algoritma yang paling sesuai bagi menganalisis opini publik dalam konteks kebijakan politik yang bersifat sensitif secara sosial serta moral.

2. BAHAN DAN METODOLOGI PENELITIAN

Penelitian ini memakai metode analisis sentimen berbasis *machine learning* bagi menganalisis sentimen publik terhadap pernyataan Presiden Prabowo Subianto terkait penyitaan aset koruptor melalui penggunaan proses komputasi *Naïve Bayes* serta *SVM*. Strategi ini diambil sebab kecakapannya terkait melakukan analisis basis data teks dalam skala besar dengan cara yang objektif serta terukur. Data dikumpulkan melalui *API Twitter* memakai pustaka *Python Tweepy*, dengan pencarian berdasarkan kata kunci seperti “aset koruptor”, “anak koruptor”, serta “penyitaan aset”. Dalam rentang waktu pengambilan data ditetapkan dari tanggal 9 April hingga 25 April 2025.



Gambar 1. Langkah-langkah Analisis Sentiment

Pada Gambar 1 proses penelitian ini melibatkan beberapa fase utama, dimulai dengan pengumpulan *dataset* melalui *Twitter API* memakai pustaka *Python Tweepy*, dengan memakai pencarian kata kunci yang relevan. Data yang diperoleh akan melewati prapemrosesan, mencakup *case folding, cleansing, tokenizing, stopword removal,* serta *stemming* bagi meningkatkan kualitas teks bagi analisis [8]. Selanjutnya, teks yang telah disempurnakan diubah menjadi format numerik memakai pendekatan *Term Frequency-Inverse Document Frequency (TF-IDF)*, yang menelaah ukuran kontribusi frasa terkait salah satu teks [9]. Setelah itu, data dikategorikan memakai dua *machine learning* yakni *Naïve Bayes* serta SVM, yang keduanya telah Menyatakan keefektifannya dalam analisis sentiment [10].

Hasil dari klasifikasi masing-masing model kemudian divisualisasikan melalui grafik bagi menggambarkan distribusi sentimen serta kinerja masing-masing algoritma. Fase terakhir melibatkan analisis hasil, membandingkan performa kedua model berdasarkan ukuran penilaian seperti *accuracy, precision, recall, F1-score,* serta *confusion matrix,* yang berujung pada suatu kesimpulan.

2.1. Pengambilan Data/*Crawling Data*

Tweepy Python digunakan sebagai modul, data dikumpulkan dari platform Twitter melalui *Twitter API*. *Dataset* dikumpulkan dengan memanfaatkan kata kunci yang terkait dengan pernyataan Presiden Prabowo Subianto, termasuk "aset koruptor", "anak koruptor", serta "penyitaan aset". Sebanyak 1.561 *tweet* berhasil dikumpulkan selama proses pengambilan data, yang berlangsung antara 9 April serta 25 April 2025. *Tweet-tweet* ini menjadi sumber informasi utama bagi prosedur analisis sentiment yang dilakukan pada penelitian ini.

2.2. *Preprocessing* Data

Setelah semua dataset terkumpul, tahap selanjutnya melakukan pengolahan terhadap data yang telah diperoleh. Tahap ini bertujuan untuk menata serta memperjelas informasi agar dapat diolah oleh algoritma *machine learning* secara lebih efisien dan akurat. *Dataset* awal terdiri dari 1.584 *tweet* yang dikumpulkan melalui *API Twitter*. Setelah melalui proses *preprocessing*, sebanyak 1.561 *tweet* berhasil dipertahankan dan siap digunakan untuk proses klasifikasi. Proses pengolahan awal mencakup *cleansing, case folding, stopword removal, stemming* serta *normalization* [8]. Dengan proses ini, dimensi data menjadi lebih terstruktur serta kata-kata yang mempunyai arti yang sama dapat dikelompokkan bersama, maka dapat meningkatkan efektivitas dalam proses klasifikasi.

2.3. *TF-IDF*

Teknik ekstraksi fitur *TF-IDF* digunakan bagi Menilai setiap kata Frekuensinya pada arsip dan seluruh korpus. Kata yang dominan terkait sebuah berkas namun minim terlihat pada berkas lainnya cenderung memperoleh bobot lebih besar karena merepresentasikan kekhasan isi dokumen [9]. Proses ini diperlukan karena istilah dapat berupa kata atau frasa. Agar dokumen atau teks dapat dikenali dalam konteks sistem, maka dokumen atau teks tersebut harus diberi bobot dalam bentuk nilai biner.

Pembobotan *TF-IDF* dapat diterapkan secara efektif melalui kelas *TfidfVectorizer* yang tersedia dalam pustaka *Sklearn* di Python.

$$Wij=tfij \times log\left(\frac{n}{df}\right) \quad (1)$$

Wij merupakan bobot *TF-IDF* bagi term ke-*I* terkait berkas ke-*j,* sedangkan *tfij* menyatakan Frekuensi term i dalam arsip *j.* Nilai n mempresentasikan jumlah total serta *dfi* ialah jumlah dokumen yang memuat term *i.*

2.4. *Naïve Bayes*

Modeling yang pertama kali dilakukan dengan memakai algoritma *Naïve Bayes* dengan pendekatan *Multinomial Naïve Bayes*, yang paling umum dimanfaatkan dalam klasifikasi teks. Dasar dari algoritma ini ialah *Teorema Bayes*, yang memakai distribusi frekuensi dari atribut bagi menentukan kemungkinan data termasuk dalam kelas tertentu.

Multinomial Naïve Bayes akan menentukan kemungkinan sebuah *tweet* Terhitung dalam salah satu dari tiga kategori sentimen seperti positif, negatif, serta netral berdasarkan distribusi kata pada dokumen pelatihan ketika berhadapan dengan data teks yang telah direpresentasikan dengan *TF-IDF*. Kemudahan penggunaan model, waktu pelatihan yang cepat, serta kinerja kategorisasi teks yang kuat menyebabkan model ini dipilih [11].

Agar menangani ketimpangan kelas pada *dataset*, Teknik *Synthetic Minority Over-sampling Technique (SMOTE)* ini diimplementasikan pada data pelatihan. *SMOTE* beroperasi bersama mendistribusikan data kelas minoritas agar menjadi lebih seimbang [12].

$$P(Ci \mid x) = \frac{P(x \mid Ci) \cdot P(Ci)}{P(x)} \quad (2)$$

Di mana $P(Ci \mid x)$ adalah probabilitas jika dokumen x bagian dari pada kelas Ci. Nilai $P(Ci \mid x)$ merupakan probabilitas dokumen x muncul dalam kelas Ci, sedangkan $P(Ci)$ menunjukkan probabilitas awal (prior) berdasarkan kelas Ci yaitu seberapa besar kemungkinan suatu dokumen berasal dari kelas tersebut sebelum melihat fitur pada dokumen. Sementara itu, $P(x)$ ialah probabilitas keseluruhan berdasarkan dokumen x muncul pada seluruh data. Persamaan ini menjadi dasar dalam menentukan kelas paling mungkin bagi suatu dokumen teks berdasarkan distribusi data pelatihan yang tersedia.

2.5. *Support Vector Machine (SVM)*

Dilakukan pemodelan pada metode tersebut memakai metode *SVM* melalui karnel linear yang dianggap paling sesuai bagi data berdimensi tinggi seperti representasi teks dari *TF-IDF* [13]. *SVM* menentukan bidang pemisah efektif demi mengelompokkan data menuju kategori jenis kelas yang variatif secara maksimal.

Pemodelan ini dipilih karena kemampuannya mengatasi data kompleks serta memberikan akurasi tinggi dalam klasifikasi teks. Tujuan utama *SVM* ialah memaksimalkan margin antara kelas yang berbeda agar prediksi lebih optimal. Pada metode ini *hyperparameter tuning* diterapkan bagi mengoptimalkan performa model.

$$f(x)=sign(sum_{i}^{n} =1 y_i a_i K(x_i, x)+b) \quad (3)$$

f(x) adalah fungsi prediksi yang menentukan kelas positif, negatif atau netral berdasarkan hasil perhitungan. Variabel y_i merupakan label kelas dari data latih ke-*i*, a_i merupakan bobot hasil optimasi $K(x_i, x)$ adalah peran kernel yang mengkalkulasi kemiripan pada kedua data latih serta data uji, serta *b* adalah bias. Hasil dari fungsi ini menentukan apakah data diklasifikasikan ke dalam kelas positif atau negatif berdasarkan tanda dari outputnya.

Penggunaan *kernel linear* dipilih lantaran sejalan melalui karakter data teks berdimensi signifikan serta representasi *sparse* seperti hasil ekstraksi fitur dari *TF-IDF*.

2.6. *Confusion Matrix*

Dalam penilaian model dilakukan bagi menilai kinerja algoritma klasifikasi terhadap data uji dan data testing. Evaluasi *confusion matrix* dimanfaatkan dalam penelitian ini, tabel yang menyajikan perbandingan antara hasil prediksi model dan label aktual dari data. *Confusion matrix* menyajikan pemahaman komprehensif seputar keseluruhan perkiraan yang akurat serta keliru dari masing-masing kelas [14].

Pada klasifikasi multikelas dengan tiga kelas, yaitu 1 (positif), 0 (negatif), dan 2 (netral), *confusion matrix* berbentuk matriks 3×3 yang memetakan jumlah prediksi benar dan salah antar kelas. Elemen diagonal dari matriks merepresentasikan prediksi yang akurat untuk masing-masing kelas, sedangkan elemen non-diagonal menunjukkan kesalahan dalam pengelompokan. Secara umum, *True Positive* (*TP*) adalah jumlah data dari suatu kelas yang diprediksi dengan benar, *False Positive* (*FP*) adalah jumlah data dari kelas lain yang salah diprediksi sebagai kelas tersebut, *False Negative* (*FN*) adalah jumlah data dari kelas tersebut yang salah diprediksi sebagai kelas lain, dan *True Negative* (*TN*) adalah jumlah data dari kelas lain yang diprediksi dengan benar bukan sebagai kelas tersebut.

Berdasarkan nilai-nilai *confusion matrix* tersebut, berikut formula matrix *accuracy, precision, recall,* dan *F1-score* dihitung berikut ini:

2.6.1 *Accuracy*

Menghitung proporsi total prediksi yang benar terhadap semua dataset. Metrik ini memberikan gambaran umum kinerja model, namun bisa menjadi kurang representatif jika data tidak seimbang.

$$Accuracy = \frac{TP+TN}{TP+FP+FN+TN} \quad (4)$$

2.6.2 *Precision*

Mengukur proporsi data yang diprediksi positif dan yakni seberapa banyak dari data yang diprediksi positif ternyata benar-benar progresif.

$$Precision = \frac{TP}{TP+FP} \quad (5)$$

2.6.3 *Recall*

Atau sensitivitas menunjukkan kapasitas model saat mengidentifikasi seluruh data aktual positif.

$$Recall = \frac{TP}{TP+FN} \quad (6)$$

2.6.4 *F1-Score*

Ukuran kinerja yang menyatukan *precision* dan *recall* dalam satu metrik harmonis, yang memberikan penilaian seimbang antara keduanya. Cocok digunakan ketika distribusi data tidak seimbang.

$$F1\text{-}Score = 2\frac{Precision \times Recall}{Precision + Recall} \quad (7)$$

3. HASIL DAN PEMBAHASAN

3.1. Pengumpulan Dataset

Pendataan terkait isu penyitaan aset koruptor dilakukan melalui media sosial Twitter dengan memanfaatkan *API* resmi dari platform tersebut. Proses ini dijalankan memakai pustaka *Python Tweepy* di lingkungan Google Colab guna mempermudah proses akses data serta penyimpanan secara otomatis.

Potongan kode pada Gambar 2 menunjukkan proses autentikasi ke *Twitter API* serta pengambilan *tweet* berdasarkan kata kunci tertentu:

```python
import tweepy
import pandas as pd

# Autentikasi ke Twitter API
auth = tweepy.OAuthHandler("API_KEY", "API_SECRET")
auth.set_access_token("ACCESS_TOKEN", "ACCESS_TOKEN_SECRET")
api = tweepy.API(auth, wait_on_rate_limit=True)

# Query pencarian
query = "aset koruptor OR anak koruptor OR penyitaan aset"

# Ambil tweet
tweets = tweepy.Cursor(api.search_tweets,
                       q=query,
                       lang="id",
                       since="2025-04-09",
                       until="2025-04-25",
                       tweet_mode='extended').items(1561)

# Simpan ke dalam list
data = [[tweet.conversation_id_str, tweet.created_at, tweet.favorite_count, tweet.full_text,
tweet.id_str, tweet.image_url, tweet.in_reply_to_screen_name, tweet.lang, tweet.location,
tweet.quote_count, tweet.reply_count, tweet.retweet_count, tweet.tweet_url, tweet.user_id_str,
tweet.username] for tweet in tweets]

# Konversi ke DataFrame dan simpan sebagai CSV
df = pd.DataFrame(data, columns=["Tanggal", "Username", "Isi_Tweet"])
df.to_csv("prabowo.csv", index=False)
```

Gambar 2. Potongan *Source Code* Pengambilan *Dataset*

Yang kemudian dataset tersebut dikumpulkan serta disimpan dalam file prabowo.csv agar mempermudah proses pengolahan lebih lanjut. Data yang tersimpan kemudian digunakan sebagai dasar bagi tahap *preprocessing* dan analisis sentimen.

Hasil dari potongan *source code* dapat dilihat pada Tabel 1 yang merupakan tampilan *Excel* dengan format *CSV*.

Tabel 1. Tampilan *Dataset* Twitter

No	Tanggal	*Username*	*Full Text*
1.	Apr 12	Kyoutachibana	@txtdrimedia Gemes banget dah. Tulil ny di batasin bisa? tulil mulu perasaan
2.	Apr 12	kakakUly	@txtdrimedia Weleehh..... pertimbantidakn macam apaan iniii...??
3.	Apr 12	ferdisijabat	@txtdrimedia Memang kehidupan anak para koruptor bakalan lebih miris daripada kehidupan orang fakir miskin..ya
4.	Apr 12	LennaLenno	@udamandi @txtdrimedia Karena lo bukan ahli hukum dan bukan bagian dari org yang terlibat makanya lu tidakmpang ngomong gt. Lu kira hukum itu sesimple emosi lu yang dangkal? Lu kira ah gua marah nih hukum mati dia! Gt doang bikin UU? Tolol ih
5.	Apr 12	LennaLenno	@BukanKaumMuna @txtdrimedia Ini pola pikir yang harus diciptakan yang belum korupsi nah yang udah korup? Ini kan lg ngomongin yang udh korup. Ginana sih lu
6.	Apr 12	LennaLenno	@Yudhamaru @txtdrimedia Klo sudut pandang ini setuju. PR nya ahli2 hukum buat merumuskan UU nya dan mentidaktur sistem perampasan asetnya ntidaktur penyelidikannya jutidak. PR besar bgt ini. Klo mau tidakmpang emg dihukum mati aja tapi kan apa yakin dia b

No	Tanggal	*Username*	*Full Text*
7.	Apr 12	LennaLenno	@geminta03254661 @txtdrimedia Itulah PR Besarnya ahli2 HUKUM untuk merumuskan UU Perampasan aset ini. Emang tidakk mudah bikin UU yang bisa adil ke semua pihak. Makanya korupsi tuh kejahatan yang ribet bgt diHukum gue bukan bela koruptor. Tapi ini realita
8.	Apr 12	LennaLenno	@al_khalid212 @txtdrimedia Kalimat lo framing bgt bego
9.	Apr 12	LennaLenno	@OBlank31599 @txtdrimedia Aduhh RIP literasi wakakkak Perampasan aset yang dimaksud itu adalah aset yang didapat hasil korupsi. Tapi gmna kalo kecampur? Itulah yang jadi PR para ahli hukum untuk merumuskan nya supaya adil. Paham tidakkkklk??
10.	Apr 12	PrakosoEdi17795	@txtdrimedia Bapake presiden kepiye iki. Klo orang tuane urung ketauan koripsi anak anakke hidup mewah orang lain menderita iku ana filme dari @Gelaper125866 . Sita bae hasil korupsi kasi ke orang2 yang membutuhkan.
1584	Apr 9	LovingRetidakl	@txtdrimedia enaknya jad anak koruptor

3.2. *Preprocessing*

Data yang telah dikumpulkan dari Twitter melalui file Prabowo.csv selanjutnya melewati beberapa Langkah *preprocessing* agar membersihkan serta menyederhanakan dataset sebelum memulai klasifikasi. Langkah pertama adalah *cleaning text*, artinya menghilangkan simnol tidak berkaitan misalnya angka, tanda baca, simbol, *URL, mention*, serta *hashtag* untuk menghilangkan *noise* dalam data [15].

Setelah tahap pembersihan teks, proses selanjutnya ialah mengonversi seluruh huruf kapital menjadi huruf kecil (*lowercase*). Tujuan dari langkah ini ialah supaya mencegah terjadinya penggandaan kata yang memiliki makna sama namun berbeda dalam penulisan kapital, seperti “Aset” dan “aset” [16], kemudian dilakukan *stopword removal*, yaitu menyingkirkan kata yang sering muncul seperti “yang”, “dan”, “di”, atau “itu” yang tidak memiliki nilai informasi signifikan terhadap sentimen teks [17].

Selanjutnya dilakukan tahapan membagi teks dipecah ke dalam satuan kata yang berskala kecil yang disebut *token*. Proses ini menjadi dasar dalam pembobotan kata dan ekstraksi fitur karena *token* merupakan input utama dalam representasi teks [18]. Dalam penelitian ini, *tokenisasi* dilakukan memakai metode berbasis spasi dan tanda baca, dibantu oleh *library* pemrosesan bahasa Python seperti *nltk*. Setelah itu dilakukan *stemming*, yaitu mengubah kata ke bentuk dasar menggunakan algoritma *Sastrawi,* misalnya kata “menderita” menjadi “derita” [19]. Proses terakhir yaitu mengubah kata tidak baku atau kata *slang* ke dalam bentuk baku yang sesuai dengan kamus besar bahsa Indonesia (KBBI). Seperti kata “duit” menjadi “uang” [20]. Tahapan *preprocessing* dapat dilihat pada Tabel 2.

Tabel 2. Tahapan *Preprocessing*

Tahapan	Hasil
Data Awal	1909890000000000000Fri Apr 11 13:06:58 +0000 20250@txtdrimedia @txtdrimedia Adil pak @prabowo selama mreka bahagia makan duit korup bapaknya dan banyak tidakya anakÂ² indonesia banyak yang menderita. Jd sdh selayaknya kekayaan hasil korupsi itu d sita netidakra. 1910680000000000000txtdrimediain000 https://x.com/ksatria_17/status/ 191068097610955990613731900000000 00000ksatria_17
Cleaning Text	Adil pak selama mreka bahagia makan duit korup bapaknya dan banyak tidakya anak indonesia banyak yang menderita. Jd sdh selayaknya kekayaan hasil korupsi itu d sita netidakra.
Case Folding	adil pak selama mreka bahagia makan duit korup bapaknya dan banyak tidakya anak indonesia banyak yang menderita. Jd sdh selayaknya kekayaan hasil korupsi itu d sita netidakra.
Stopword Removal	adil mreka bahagia makan duit korup bapaknya banyak anak indonesia banyak menderita jd sdh selayaknya kekayaan hasil korupsi sita netidakra
Tokenizing	[‘adil’, ‘mreka’, ‘bahagia’, ‘makan’, ‘duit’, ‘korup’, ‘bapaknya’, ‘banyak’, ‘gaya’, ‘anak’, ‘indonesia’, ‘banyak’, ‘menderita’, ‘jd’, ‘sdh’, ‘selayaknya’, ‘kekayaan’, ‘hasil’, ‘korupsi’, ‘sita’, ‘netidakra’]
Stemming	adil mreka bahagia makan duit korup bapak banyak gaya anak indonesia banyak derita jadi sudah layak kaya hasil korup sita netidakra
Normalization	adil mereka bahagia makan uang korupsi bapak banyak gaya anak indonesia banyak derita jadi sudah layak kekayaan hasil korupsi sita negara

3.3. *Naïve Bayes*

Model *Naïve Bayes* diuji pada data uji yang telah direpresentasikan dalam bentuk vektor *TF-IDF*. Data yang digunakan terdiri dari 1.561 komentar, ada 1.247 data latih serta 312 data uji yang telah melalui pembagian dataset rasio 80:20 antara pelatihan dan pengujian.

Supaya mengatasi permasalahan data imbalance kelas dalam data latih, diterapkan *SMOTE* diterapkan untuk menyeimbangkan distribusi data dengan menambahkan sampel pada kelas minoritas. Jumlah data sebelum penerapan *SMOTE* adalah 1.247, kemudian meningkat menjadi 1.668 setelah proses *oversampling*. Model kemudian diuji terhadap 313 data uji, menghasilkan performa seperti pada Tabel 3.

Tabel 3. Hasil Evaluasi *Naïve Bayes*

Kelas	*Precision*	*Recall*	*F1-Score*
0	0.63	0.60	0.62
1	0.77	0.85	0.81
2	0.42	0.36	0.39
Accuraacy			0.66

Berdasarkan hasil pengujian yang ditampilkan pada tabel 3, diperoleh akurasi sebesar 66,34%. Dari *classification report*, nilai *precision* tertinggi diperoleh pada kelas 1 (positif) dengan skor 0,77, diikuti oleh kelas 0 (negatif) sebesar 0,63, dan kelas 2 (netral) sebesar 0,42. Hal ini mengindikasikan bahwa model lebih mampu mengenali kelas yang masuk ke dalam kategori kelas positif, yang merupakan kelas mayoritas dalam distribusi data.

Confusion matrix menunjukkan adanya kesalahan klasifikasi antar kelas, khususnya pada komentar negatif yang banyak diprediksi sebagai positif. Fenomena ini dapat disebabkan oleh kemiripan kata atau frasa yang digunakan dalam komentar yang secara konteks bisa bermakna netral maupun positif, tergantung pada interpretasi kalimat secara keseluruhan.

Berdasarkan hasil ini, memperoleh kesimpulan bahwa model *Naive Bayes* cukup efektif digunakan sebagai model dasar (*baseline*) bagi klasifikasi komentar berbasis teks, terutama pada kelas mayoritas. Namun, performanya masih terbatas untuk kelas minoritas yang cenderung lebih sulit dikenali. Supaya mengatasi keterbatasan tersebut, Penelitian ini turut menerapkan algoritma *SVM* sebagai salah satu metode klasifikasi yang digunakan. sebagai pendekatan alternatif untuk melihat potensi peningkatan performa klasifikasi secara menyeluruh.

3.4. *Support Vector Machine (SVM)*

Pemodelan *SVM* digunakan bagi memproses data yang sebelumnya telah melewati tahapan *preprocessing* serta transformasi menjadi vektor melalui metode *TF-IDF*. Berdasarkan hasil *tuning hyperparameter*, diperoleh kombinasi terbaik dengan parameter C = 1, *gamma = 'scale'*, dan *kernel linear*. *SVM* menghasilkan akurasi sebesar 70,51%, lmelampaui capaian dari *Naïve Bayes* yang hanya memperoleh 66,34%.

Tabel 4. Hasil Evaluasi *SVM*

kelas	*Precision*	*Recall*	*F1-Score*
0	0.63	0.72	0..67
1	0.78	0.88	0.83
2	0.61	0.30	0.40
Accuracy			0.70

Lihat Tabel 4 jika dibandingkan, performa algoritma *SVM* secara keseluruhan menunjukkan hasil yang lebih baik dibandingkan *Naïve Bayes*, dengan selisih akurasi sebesar 4,17% dan peningkatan *F1-score* pada kelas positif dari 0,81 menjadi 0,83. Hal ini menunjukkan bahwa *SVM* mampu membedakan opini dengan polaritas positif secara lebih konsisten.

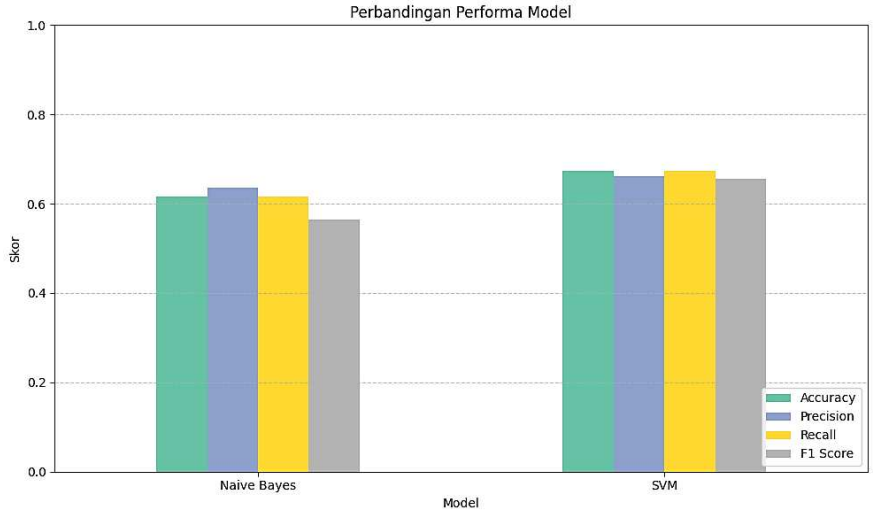
Selain itu, model *SVM* memiliki keunggulan dalam mengenali kelas negatif dengan *recall* 0,72 dibandingkan dengan *Naïve Bayes* yang hanya mencapai 0,60. Namun demikian, kedua model masih menghadapi tantangan dalam klasifikasi kelas netral, dengan *F1-score* masing-masing hanya 0,39 (*Naïve Bayes*) dan 0,40 (*SVM*), yang mengindikasikan bahwa komentar bersentimen netral lebih sulit diidentifikasi karena cenderung ambigu secara konteks dan struktur bahasa.

Keunggulan *SVM* dapat dijelaskan melalui pendekatannya yang tidak mengasumsikan independensi fitur seperti pada *Naïve Bayes*, serta kemampuannya dalam memisahkan kelas melalui *margin* optimal pada data berdimensi tinggi seperti representasi *TF-IDF*. Berdasarkan evaluasi ini, dapat disimpulkan bahwa *SVM* merupakan model yang lebih handal untuk klasifikasi sentimen pada isu penyitaan aset koruptor.

3.5. Hasil Analisis

Gambar 3 SVM menunjukkan keunggulan dalam akurasi dengan perolehan 70,51%, mengungguli Naïve Bayes yang mencatat 66,34%. Performa SVM ini menunjukkan konsistensinya yang lebih ideal guna mengklasifikasikan kelas positif dan negatif secara lebih akurat. Hal ini menunjukkan bahwa pendekatan SVM lebih mampu menangani variabilitas data teks di media sosial yang kompleks dan kontekstual. Sejalan

dengan temuan Rima et al. [5]. SVM juga terbukti unggul dalam mengklasifikasikan isu-isu sensitif seperti opini publik terhadap vaksinasi COVID-19, yang menunjukkan tingkat akurasi lebih tinggi dibandingkan Naïve Bayes.



Gambar 3. Hasil Analisa *Naïve Bayes* dan *SVM*

Berdasarkan distribusi sentiment, kelas 1 (positif) berhasil diklasifikasikan oleh kedua model dengan jumlah data tertinggi. Hal ini menunjukkan mayoritas opini masyarakat cenderung setuju bahwa anak koruptor turut menanggung dampak sosial dari perbuatan orang tuanya. Sebaliknya, kelas 0 (negatif) memiliki pandangan berbeda dan menolak bahwa anak tidak pantas dipersalahkan atas kesalahan orang tuanya. Serta pada kelas 2 (netral) memilih supaya bersikap objektif atau tidak berpihak dalam meanggapi isu ini. Analisis ini menunjukkan bahwa algoritma klasifikasi tidak hanya mampu mengukur performa teknis model, tetapi juga dapat menangkap kecendrungan opini *public* terhadap isu sosial dan politik. *SVM* terbukti memberikan hasil yang lebih stabil dalam konteks data media sosial, meskipun selisih performa dengan *Naïve Bayes* tidak terlalu besar.

4. KESIMPULAN

Penelitian ini membuktikan Algoritma *SVM* menunjukkan performa yang lebih optimal dibandingkan *Naïve Bayes* dalam melakukan klasifikasi sentimen terhadap komentar Twitter terkait isu penyitaan aset koruptor, dengan akurasi sebesar 70,51% serta *F1-score* 0,69. Keluaran ini mendukung hipotesis bahwa pemilihan algoritma berpengaruh signifikan terhadap akurasi klasifikasi teks. *Naïve Bayes* tetap efektif sebagai model dasar dengan keunggulan dalam efisiensi, namun kurang optimal pada kelas minoritas. Kekurangan dalam penelitian ini meliputi keterbatasan topik, periode pengambilan data yang relatif singkat, dan belum digunakannya pendekatan berbasis *deep learning*. Potensi pengembangan ke depan mencakup perluasan isu, memperpanjang periode pengambilan data, penggunaan model seperti *BERT* serta *LSTM*. Penelitian ini memperkuat efektivitas metode klasifikasi dalam analisis opini publik, serta secara praktis dapat menjadi dasar bagi pengembangan sistem monitoring sentimen kebijakan publik di media sosial sehingga berguna memperkuat dasar dalam menentukan keputusan yang lebih responsif serta berbasis data.

REFERENSI

[1] S. Lestari dan S. Berliani, "Analisis Sentimen Masyarakat Terhadap Isu Pecat Sri Mulyani Pada Twitter Menggunakan Metode Naive Bayes Dan Support Vector Machine," *Jurnal Sains dan Teknologi,* vol. 5, no. 3, pp. 951-960, Februari 2024. DOI: https://doi.org/10.55338/saintek.v5i3.2746

[2] A. Syakir dan F. N. Hasan, "Analisis Sentimen Masyarakat Terhadap Perilaku Korupsi Pejabat Pemerintah Berdasarkan Tweet Menggunakan Naive Bayes Classifier," *JURNAL MEDIA INFORMATIKA BUDIDARMA,* vol. 7, no. 4, pp. 1796-1805, Oktober 2023. DOI: https://doi.org/10.30865/mib.v7i4.6648

[3] J. Juliantono dan P. , "Persepsi Publik Terhadap Kepemimpinan Firli Bahuri di KPK: Pendekatan Sentimen Twitter dengan Naive Bayes," *JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika),* vol. 10, no. 2, pp. 1272-1285, Juni 2025. DOI: https://doi.org/10.29100/jipi.v10i2.6181

[4] R. R. Putri dan N. Cahyono, "Analisis Sentimen Komentar Masyarakat Terhadap Pelayanan Publik Pemerintah DKI Jakarta Dengan Algoritma Support Vector Machine dan Naive Bayes," *JATI (Jurnal Mahasiswa Teknik Informatika),* vol. 8, no. 2, pp. 2363-2371, April 2024. DOI: https://doi.org/10.36040/jati.v8i2.9472

[5] R. T. Aldisa dan P. Maulana, "Analisis Sentimen Opini Masyarakat Terhadap Vaksinasi Booster COVID 19 Dengan Perbandingan Metode Naive Bayes, Decision Tree dan SVM," *Building of Informatics, Technology and Science (BITS),* vol. 4, no. 1, pp. 106-109, Juni 2022. DOI: https://doi.org/10.47065/bits.v4i1.1581

[6] M. S. Başarslan dan F. Kayaalp, " Sentiment Analysis with Machine Learning Methods on Social Media," *ADCAIJ: Advances in Distributed Computing and Artificial Intelligence Journal ,* vol. 9 , no. 3, pp. 5-15, 2020. DOI: https://doi.org/10.14201/ADCAIJ202093515

[7] I. Kurniawan, A. L. Hananto, S. S. Hilabi, A. Hananto, B. Priyatna dan A. Y. Rahman, "Perbandingan Algoritma Naive Bayes Dan SVM Dalam Sentimen Analisis Marketplace Pada Twitter," *Jurnal Teknik Informatika dan Sistem Informasi,* vol. 10, no. 1, pp. 731-740, 2023. DOI: https://doi.org/10.35957/jatisi.v10i1.3582

[8] H. Ma'rifah, A. W. Wibawa dan M. I. Akbar, "Klasifikasi artikel ilmiah dengan berbagai skenario preprocessing," *Sains, Aplikasi, Komputasi dan Teknologi Informasi,* vol. 2, no. 2, pp. 70-78, April 2020. DOI: http://dx.doi.org/10.30872/jsakti.v2i2.2681

[9] Herimanto, K. Samosir dan F. Ginting, "A Comparative Analysis of Content-Based Filtering and TF-IDF Approaches for Enhancing Sports Recommendation Systems," *Innovation in Research of Informatics (INNOVATICS),* vol. 6, no. 2, pp. 90-97, 2024. DOI: https://doi.org/10.37058/innovatics.v6i2.12404

[10] W. Ningsih, B. Alfianda, R. dan D. Wulandari, "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,* vol. 4, no. 2, pp. 556-562, 2 April 2024. DOI: https://doi.org/10.57152/malcom.v4i2.1253

[11] A. Sabrani dan I. G. P. W. W. W, "Metode Multinomial Naïve Bayes untuk Klasifikasi Artikel Online Tentang Gempa di Indonesia," *JTIKA,* vol. 2, no. 1, pp. 89-100, Maret 2020. DOI: https://doi.org/10.29303/jtika.v2i1.87

[12] N. Sulistiyowati dan M. Jajuli, "Integrasi Naïve Bayes Dengan Teknik Sampling Smote Untuk Menangani Data Tidak Seimbang," *JURNAL NUANSA INFORMATIKA,* vol. 14, no. 1, pp. 34-37, Januari 2020. DOI: https://doi.org/10.25134/nuansa.v14i1.2411

[13] S. Rabbani, D. Safitri, N. Rahmadhani, A. A. F. Sani dan 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, October 2023. DOI: https://doi.org/10.57152/malcom.v3i2.897

[14] R. dan F. Thalib, "Implementasi Algoritma Naïve Bayes Classifier Dan Confusion Matrix Dalam Analisis Sentimen Terhadap Pelayanan Transportasi Umum Selama Pandemi Covid-19 Pada Media Sosial Twitter," *JURNAL TEKNOLOGI,* vol. 8, no. 1, pp. 65-75, 2020. DOI: 10.31479/jtek.v1i8.66

[15] A. P. Joshi dan B. V. Patel, "Data Preprocessing: the Techniques for Preparing Clean and Quality Data for Data Analytics Process," *Oriental Journal of Computer Science and Technology,* vol. 13, no. 2-3, pp. 78-81, 2020. DOI : http://dx.doi.org/10.13005/ojcst13.0203.03

[16] L. Hermawan dan M. B. Ismiati, "Pembelajaran Text Preprocessing berbasis Simulator Untuk Mata Kuliah Information Retrieval," *TRANSFORMATIKA,* vol. 17, no. 2, pp. 188-199, Januari 2020. DOI: https://doi.org/10.26623/transformatika.v17i2.1705

[17] R. Rinandyaswara, Y. A. Sari dan M. T. Furqon, "Pembentukan Daftar Stopword Menggunakan Term Based Random Sampling Pada Analisis Sentimen Dengan Metode Naïve Bayes" *Jurnal Teknologi Informasi dan Ilmu Komputer (JTIIK) ,* vol. 9, no. 4, pp. 717-724, Agustus 2022. DOI: https://doi.org/10.25126/jtiik.2022934707

[18] J. Petrus, E. S. dan E. , "A Novel Approach: Tokenization Framework based on Sentence Structure in Indonesian Language," *(IJACSA) International Journal of Advanced Computer Science and Applications,* vol. 14, no. 2, pp. 541-549, 2023. DOI: 10.14569/IJACSA.2023.0140264

[19] Z. Abidin, A. Junaidi dan W. , "Text Stemming and Lemmatization of Regional Languages in Indonesia: A Systematic Literature Review," *Journal of Information Systems Engineering ,* vol. 10, no. 2, pp. 217-231, Juni 2024. DOI: https://doi.org/10.20473/jisebi.10.2.217-231

[20] R. G. Whendasmoro dan Joseph, "Analisis Penerapan Normalisasi Data Dengan Menggunakan Z-Score Pada Kinerja Algoritma K-NN," *JURIKOM (Jurnal Riset Komputer),* vol. 9, no. 4, pp. 872-876, Agustus 2022. DOI: https://doi.org/10.30865/jurikom.v9i4.4526