IT-EXPLORE
Jurnal Penerapan Teknologi Informasi dan Komunikasi
ISSN 2828-7940, e-ISSN 2829-1727, DOI: 10.24246/itexplore. v4i3.2025.pp307-322
Volume 04 Nomor 03 Tahun 2025
Journal homepage: https://ejournal.uksw.edu/itexplore

Analisis sentimen kinerja bea cukai pada barang luar negeri dengan *Support Vector Machine*

Angelica Mailen Sipasulta [1], Teguh Indra Bayu [2*]
[1,2] Program Studi Teknik Informatika, Fakultas Teknologi Informasi,
Universitas Kristen Satya Wacana, Salatiga, Jawa Tengah

Informasi Artikel

Riwayat Artikel:

Dikirim:
11 April 2025
Direvisi:
23 Juni 2025
Diterima:
26 Juni 2025

Kata Kunci:
Analisis Sentimen
Bea Cukai
Support Vector Machine (SVM)
Twitter

ABSTRAK

Bea Cukai belakangan ini telah menjadi sorotan publik, terutama terkait pengawasan barang dari luar negeri. Berita dan tanggapan masyarakat mengenai pengawasan bea cukai menciptkan pro dan kontra, sehingga memicu beragam respon dari masyarakat. Twitter menjadi salah satu platform utama bagi masyarakat untuk berdiskusi dan menyampaikan pandangan mereka terkait kinerja bea cukai dalam pengawasan barang dari luar negeri. Opini yang disampaikan oleh masyarakat dapat berupa sentiment positif atau negatif, yang dapat dijadikan bahan pertimbangan untuk evaluasi pelayanan bea cukai ke depan. Penelitian ini bertujuan untuk menganalisis sentimen masyarakat Indonesia terhadap kinerja bea cukai dalam pengawasan barang dari luar negeri dengan memanfaatkan media sosial Twitter. Dalam penelitian ini, algoritma *Support Vector Machine* (SVM) diterapkan untuk mengklasifikasikan komentar masyarakat di twitter ke dalam sentimen positif atau negatif. Melalui proses *crawling* yang dilakukan dari tanggal 1 Juni 2023 hingga 12 Mei 2024, berhasil dikumpulkan data sebanyak 9.051 entri. Hasil penelitian menunjukkan bahwa sentimen masyarakat terhadap kinerja bea cukai cenderung negatif dengan presentase sebesar 60,49%, serta model SVM yang digunakan memiliki akurasi sebesar 93,87%, *precision* 94%, *recall* 93%, dan F1-*score* 94%. Ini menunjukkan bahwa metode SVM efektif dalam menganalisis sentimen publik, khususnya terkait dalam pengawasan bea cukai.

Keywords:
Sentiment Analysis
Bea Cukai
Support Vector Machine (SVM)
Twitter

ABSTRACT

Bea Cukai has recently been in the public spotlight, especially regarding the supervision of goods from abroad. News and public responses regarding Bea Cukai's supervision create pros and cons, thus triggering a variety of responses from the public. This study aims to analyze the sentiment of Indonesian people towards the performance of Bea Cukai in monitoring goods from abroad by utilizing Twitter social media. In this research, the Support Vector Machine (SVM) algorithm is applied to classify public comments on Twitter into positive or negative sentiments. Through the crawling process carried out from June 1, 2023, to May 12, 2024, 9,051 entries of data were collected. The results indicate that public sentiment toward customs performance tends to be negative, with 60.49% of comments expressing negative sentiment. The SVM model used in this study achieved an accuracy of 93.87%, precision of 94%, recall of 93%, and an F1-score of 94%. These findings demonstrate that the SVM method is effective in analyzing public sentiment, particularly regarding customs supervision.

This is an open access article under the CC BY-SA license

Corresponding Author:

Teguh Indra Bayu
Fakultas Teknologi Informasi, Universitas Kristen Satya Wacana
Jl. Diponegoro No.52-60, Sidorejo, Salatiga, Jawa Tengah, Indonesia
Email: teguh.bayu@uksw.edu

1. PENDAHULUAN

Tidak ada satu negara pun di dunia yang dapat mencukupi seluruh kebutuhan nasionalnya dari produk dan barang yang dibuat oleh negaranya sendiri [1]**.** Maka dari itu, negara juga membutuhkan *supply* barang dari luar negeri atau barang impor. Indonesia mempunyai regulasi tersendiri mengenai barang impor yang masuk ke Indonesia yang diatur dalam Undang-Undang negara. Dalam Undang-Undang Nomor 17 Tahun 2006 tentang Perubahaan atas Undang-Undang Nomor 10 Tahun 1995 tentang Kepabean dibuat untuk mencegah importir dan eksportir melakukan kecurangan. Dalam hal ini, Direktorat Jenderal Bea dan Cukai memiliki peran penting dalam upaya penegakan hukum atas segala bentuk pidana terkait dengan kegiatan keluar masuk barang dari luar negeri di daerah pabean. Untuk memenuhi tugas dan fungsi tersebut, Direktorat Jenderal Bea dan Cukai memiliki banyak kantor pelayanan kepabeanan di berbagai daerah [2]**.**

Tetapi beberapa waktu terakhir, pelayanan dan kinerja bea cukai menjadi pembicaraan hangat masyarakat. Salah satu kasus yang ramai dibicarakan mengenai alat bantu untuk anak berkebutuhan khusus yang didonasikan oleh Korea Selatan tertahan di bea cukai bandara Soekarno-Hatta sejak tahun 2022. Kejadian ini menjadi viral pada tahun 2024 dikarenakan alat bantu tersebut belum diterima oleh pihak SLB. Menurut Direktorat Jenderal Bea dan Cukai, alat bantu SLB tertahan sejak 2022 dikarenakan pihak penerima tidak melengkapi dokumen mekanisme barang hibah, sehingga baru diproses setelah viral di media sosial pada tahun 2024. Kasus ini memicu kritik dari masyarakat mengenai kinerja bea cukai yang dianggap memperlambat birokrasi penerimaan alat SLB kepada pihak yang membutuhkan [3].

Dari kasus tersebut, mulai muncul beberapa kisah serupa yang dibagikan oleh masyarakat di media sosial. Salah satu media sosial yang digunakan adalah Twitter. Di Twitter terdapat berbagai macam komentar mulai dari yang pro dan kontra, hingga penggiringan opini baik positif maupun negatif terhadap kinerja bea cukai. Banyak informasi yang dapat ditemukan melalui pengujian sentiment dari komentar di Twitter. Sangat penting untuk melakukan analisis sentimen dari komentar-komentar yang dibuat oleh masyarakat melalui media sosial Twitter mengenai kinerja bea cukai terutama dalam pengawasan barang dari luar negeri, apakah memiliki citra positif atau negatif di masyarakat. Dan berdasarkan latar belakang diatas, komentar tersebut perlu dianalisis dengan melakukan sentiment analisis.

Berdasarkan penjelasan sebelumnya, maka penelitian ini bertujuan untuk menganalisis sentimen masyarakat Indonesia terhadap kinerja bea cukai dalam pengawasan barang dari luar negeri berdasarkan komentar masyarakat melalui media sosial twitter. Algoritma SVM dipilih dalam penelitian ini karena efektif digunakan pada data yang besar [4]. Selain itu, algoritma SVM yang digunakan dalam penelitian ini bertujuan untuk melakukan klasifikasi komentar masyarakat di twitter ke dalam sentimen positif atau negatif. Penelitian ini diharapkan dapat meningkatkan kualitas kinerja bea cukai, khususnya dalam pengawasan barang dari luar negeri dengan menggunakan teknologi informasi.

2. KAJIAN PUSTAKA

2.1. State of The Art

Penelitian yang serupa pernah dilakukan mengenai analisis sentimen. Berdasarkan penelitian terdahulu dengan topik analisis terhadap relokasi Ibukota Nusantara menggunakan algoritma Naïve Bayes dan KNN, dengan data yang berhasil dikumpulkan sebanyak 2.000 data, Teknik yang dilakukan dalam penelitian adalah *case folding, cleaning, tokenize, stopword removal, weighting,* klasifikasi, validasi dan evaluasi. Hasil dari penelitian menggunakan metode Naïve Bayes mendapatkan hasil akurasi sebesar 82,27%, *precision* sebesar 86,36%, *recall* sebesar 76,93%. Sedangkan, kinerja dengan model KNN mendapat akurasi sebesar 88,12%, *precision* sebesar 93,98%, serta *recall* 81.53% [5]. Penelitian lain yang berjudul "Analisis Sentimen Calon Presiden 2024 Menggunakan Algoritma SVM Pada Media Sosial Twitter" melibatkan beberapa tahapan *cleansing, case folding, tokenize, remove stopword,* klasifikasi, serta hasil. Hasil penelitian ini mendapat akurasi tertinggi pada klasifikasi dataset Ganjar Pranowo dengan nilai rata-rata akurasi 98,61%, *precision* 98,81%, *recall* 99,79% [6].

Penelitian berjudul "Sentimen Analisis Aplikasi Belajar Online Menggunakan Klasifikasi SVM" mengumpulkan 30.000 yang kemudian diberi label dengan nilai 1,0,-1. Nilai 1 berarti positif, nilai 0 berarti netral, dan nilai -1 berarti negatif. Hasil yang didapatkan dengfan menggunakan metode SVM mendapat akurasi rata-rata Ruang Guru sebesar 99%, Zenius sebesar 96%, dan Quipper sebesar 82% [7]. Dan pada penelitian lain membahas terkait analisis opini mengenai feminisme dan dipatkan 600 data dengan rasio 80:20 berhasil mendapatkan akurasi sebesar 87% menggunakan metode Naïve Bayes dengan nilai *recall*

86%, *precision* 89%, dan *F1-Score* 86% [8].

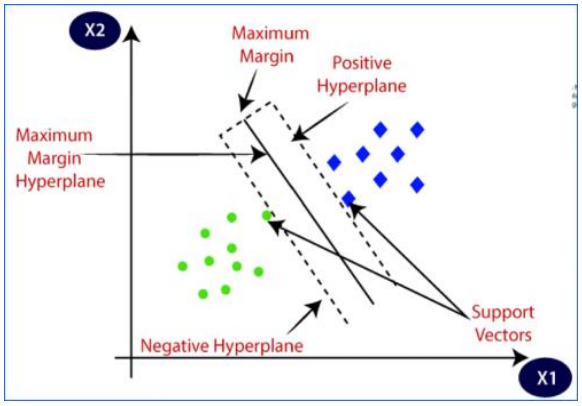
Berdasarkan penelitian mengenai analisis sentimen calon presiden 2024, SVM menghasilkan akurasi tinggi sebesar 98,61% dalam klasifikasi opini politik berbasis data twitter [6]**.** Dan, penelitian mengenai aplikasi belajar *online* menunjukkan akurasi sebesar 99% dalam domain edukasi [7] **.** Hasil-hasil tersebut mendukung bahwa SVM memiliki performa tinggi dan mampu menangani dataset besar dengan tingkat akurasi cukup tinggi.

2.2. Landasan Teori

Berdasarkan hukum kepabeanan, Direktorat Jenderal Bea dan Cukai adalah bagian dari Kementerian keuangan yang bertanggung jawab atas tugas dan peran utama dalam domain kepabeanan dan cukai, Direktorat ini bertanggung jawab atas perdagangan Internasional dan mengawasi barang yang masuk dan keluar wilayah bea cukai [9]. Namun, selama menjalankan tugasnya bea cukai sering menjadi topik diskusi di masyarakat, terutama tentang kebijakan pengawasan barang dari luar negeri, yang sering menimbulkan perbedaan pendapat di media sosial.

Maka, diperlukannya analisis sentimen. Analisis sentimen adalah teknik yang banyak digunakan dalam penelitian berbasis media sosial untuk memahami presepsi publik terhadap masalah tertentu. Ini adalah cara untuk menemukan dan mengklasifikasikan emosi atau sikap yang terkandung dalam teks [10].

Metode klasifikasi yang digunakan pada penelitian ini adalah *Support Vector Machine (SVM).* Klasifikasi *Support Vector Machine (SVM)* bekerja berdasarkan prinsip menemukan hyperlane dengan margin maksimal. *Hyperlane* berfungsi sebagai pemisah antar data dari berbagai kelas, dan margin adalah jarak antara data yang paling dekat dari setiap kelas dan *hyperlane*. Sedangkan, *Support Vector* adalah data yang berada paling dekat dari setiap kelas [11]. Algoritma SVM bertujuan menemukan hyperlane terbaik dengan memaksimalkan jarak antar kelas.



Gambar 1: Konsep *Hyperlane* pada SVM [12]

Sepasang bidang pembatas sejajar dapat digunakan untuk memisahkan dua kelas, seperti yang ditunjukan pada Gambar 1. Maka, dapat diperoleh kedua persamaan berikut [13]:

$$x_i w + b \geq 1, y_1 = +1 \quad (1)$$

$$x_i w + b \leq 1, y_1 = -1 \quad (2)$$

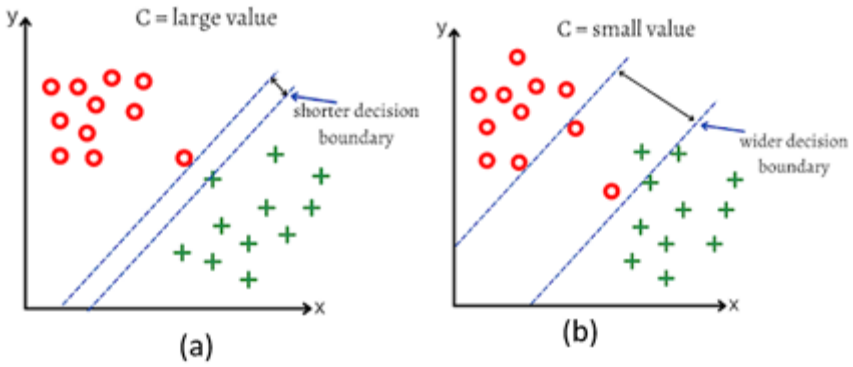
Pada persamaan 1 menunjukkan bahwa data yang diklasifikasikan sebagai positif harus berada di sisi positif dari *hyperlane*. Di sisi lain, pada persamaan 2 menunjukkan bahwa data yang diklasifikasikan sebagai negatif harus berada di sisi lain dari hyperlane. *w* berfungsi sebagai *vector* yang menentukan arah *hyperlane*, dan b menunjukkan jarak *hyperlane* dari titik asal.

Pembentukan *hyperlane* dilakukan dengan memanfaatkan fungsi kernel. Jika data dapat dipisahkan secara linear, maka digunakan kernel linear, yang memiliki persamaan sebagai berikut [14] :

$$K(x, x_i) = x, x_i. \quad (3)$$

Tetapi, performa *hyperplane* sangat bergantung pada parameter Cost (C). Parameter C merupakan pengatur keseimbangan antara lebar margin pemisah *(decision boundary)* dan jumlah kesalahan klasfikasi pada data pelatihan. Nilai C yang besar akan memaksa model untuk meminimalkan kesalahan klasifikasi sekecil mungkin, sehingga membentuk margin sempit. Akibatnya, *hyperplane* menjadi terlalu sensitif terhadap data

pelatihan, yang dapat mengurangi kemampuan generalisasi model dan meningkatkan risiko *overfitting*. Sebaliknya, nilai C yang kecil memperbolehkan lebih banyak kesalahan klasifikasi pada data pelatihan, menghasilkan margin yang lebih lebar, namun berisiko menyebabkan *underfitting* karena pemisahan kelas menjadi terlalu longgar. Pengaruh nilai C terhadap bentuk *decision boundary* divisualisasikan pada Gambar 2, di mana model dengan C besar (a) menghasilkan margin yang sempit dan garis pemisah yang ketat terhadap data pelatihan, sehingga model berisiko mengalami *overfitting*, sementara model dengan C kecil (b) menghasilkan margin yang lebih lebar dan garis pemisah yang lebih toleran terhadap kesalahan klasifikasi, namun berisiko menyebabkan *underfitting* [15].

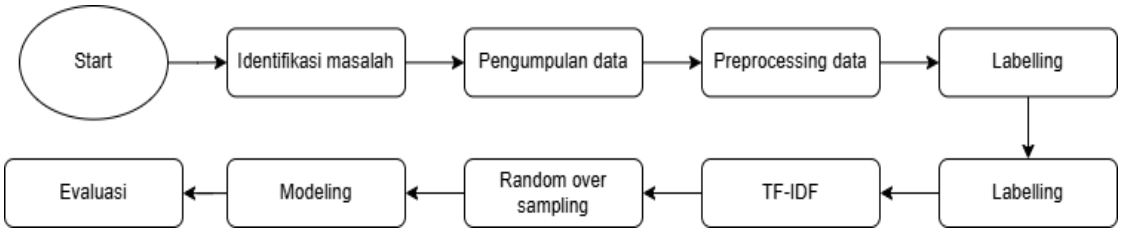


Gambar 2: (a) nilai C besar (b) nilai C kecil

Dalam penelitian ini implementasi dilakukan dengan library *Scikit-learn* dan parameter C disesuaikan melalui pengujian bertahap untuk memperoleh akurasi terbaik.

3. METODE PENELITIAN

Gambar 3 menunjukkan tahapan dalam penelitian ini yang dilakukan secara bertahap dan terstruktur. Pertama, mengidentifikasi masalah kemudian data dikumpulkan dalam jangka waktu tertentu. Data yang telah dikumpulkan akan melalui proses pembersihan data. Selanjutnya, setiap data yang telah dibersihkan diberi label sebelum melalui proses *Term Frequency-Inverse Document Frequency (TF-IDF*) dan random *oversampling* guna menyeimbangkan distribusi data. Setelah tahap ini, dilakukan proses membangun model klasifikasi. Terakhir, model yang telah dibuat dievaluasi guna mengukur performa dan akurasinya dalam menyelesaikan permasalahan yang telah diidentifikasi sebelumnya.



Gambar 3: Tahap Penelitian

3.1. Pengumpulan Data

Pada tahap pengumpulan data, *library Tweet Harvest* versi v2.6.0 digunakan untuk mengumpulkan data di media sosial Twitter menggunakan *Application Programming Interface (API).* Untuk mengakses tweets atau data, pengguna harus mendapatkan token autentikasi karena API memerlukan token autentikasi pengguna. Token autentikasi di dalamnya mencakup *consumer key, consumer secret, access token, dan access token secret* [16]. Proses *crawling* dilakukan untuk mengambil *tweet* yang mengandung kata kunci "Bea Cukai" dan "Barang dari Luar Negeri", karena kata kunci tersebut merepresentasikan isu yang relevan dengan pengawasan barang impor dan pelayanan kepabeanan, yang menjadi fokus pembahasan dalam konteks penelitian ini.

Pemilihan kata kunci tersebut dilakukan secara sistematis dengan mengacu pada diskusi publik di media sosial yang berkembang selama periode pengumpulan data, yakni dari 1 Juni 2023 hingga 12 Mei 2024. Seluruh tweet hasil *crawling* disimpan dalam format file CSV untuk diproses lebih lanjut pada tahap *preprocessing* dan klasifikasi.

3.2. *Data Preprocessing*

Data preprocessing adalah langkah penting sebelum masuk ke bagian klasifikasi model. Agar algoritma atau model *machine learning* dapat berkerja dengan lebih optimal, maka diperlukan proses *preprocessing* yang memiliki tujuan uuntuk membersihkan serta menyiapkan sebelum masuk ke proses selanjutnya [17]. Dataset ini terlebih dahulu dibersihkan dan di transformasi tanpa mengubah keasliannya, sehingga dapat memaksimalkan hasil klasifikasi. Data sebelum preprocessing tersebut dapat mengandung *missing value, outlier, duplicate value*, dsb [18]. Beberapa tahapan dalam *preprocessing* dalam penelitian ini mencakup [19]:

a. *Cleansing data*, tujuan dari bagian ini untuk membersihkan dataset dari karakter yang mengandung *noise*, yang dapat berdampak pada hasil klasifikasi. Pada bagian ini karakter seperti tagar, username, angka, URL, emotikon, dan tanda baca akan dihapus.

b. *Case Folding*, yaitu mengubah huruf kapital menjadi huruf kecil agar seragam.

c. *Tokenization* merupakan proses memilih dan memotong kata yang ada di dalam dokumen menjadi kata-kata yang disesuaikan dengan spasi. Tujuan dari proses ini untuk menjadikan kalimat yang berkaitan dengan atribut teks twitter menjadi fragmen kata yang lebih mudah digunakan.

d. Normalisasi, pada tahap ini kalimat tidak baku atau gaul akan diubah menjadi kalimat yang sesuai dengan ketentuan yang tercantum dalam Kamus Besar Bahasa Indonesia (KBBI).

e. *Stopword Removal (Filtering),* tahap ini adalah proses untuk menghapus kata-kata dari dokumen yang tidak relevan untuk analisis sentimen. Tahap ini menghasilkan pengurangan jumlah kata dalam kalimat.

f. *Stemming* adalah proses mengubah kata melalui modifikasi akhiran atau awalan, termasuk konjungsi, preposisi, serta kata Ganti, sehingga menghasilkan kata dasar yang sesuai dengan Kamus Besar Bahasa Indonesia (KBBI).

3.3. *Labeling*

Labeling atau pelabelan adalah langkah di mana kumpulan data kalimat atau dokumen diberi label atau diklasifikasikan dengan cara tertentu [20]. Dalam penelitian ini, setiap tweet dilabelkan dan diklasifikasikan sebagai positif atau negatif. *Lexicon InSet* (Indonesia Sentimen Lexicon), yang terdiri dari kumpulan kata dalam bahasa Indonesia dengan nilai -5 hingga +5 digunakan untuk menjalankan proses ini. Tahap ini menghasilkan data dengan label positif dan negatif. Kamus Lexicon ini sendiri mencakup 3.609 kata positif dan 6.609 kata negatif.

3.4. TD-IDF

Pembobotan kata, juga dikenal sebagai (*term weighting*), memiliki dua tujuan. Yang pertama adalah untuk mengubah teks yang tidak terstruktur menjadi terstruktur, dan kedua adalah untuk mengukur kontribusi setiap kata dalam dokumen terhadap kategori klasifikasi tertentu. Dalam penelitian ini menggunakan *Term Frequency-Inverse Document Frequency (TF-IDF)* sebagai salah satu metode pembobotan kata, yaitu *unsupervised learning*. *Term Frequency-Inverse Document Frequency (TF-IDF)* adalah perhitungan bobot kata pada dokumen berdasarkan seringnya kata tersebut muncul, yang menunjukan bahwa semakin banyak kata yang muncul dalan dokumen, semakin penting term tersebut [21].

Data teks dalam TF-IDF dikonversi menjadi bentuk vektor untuk mempertimbangkan bobot setiap kata dalam dokumen. Berdasarkan TF-IDF setiap kata dalam korpus diberi nilai numerik yang menunjukan seberapa penting kata tersebut dalam konteks korpus. Nilai TF-IDF dihitung dengan *TfidfVectorizer* dari *pustaka Python Sklearn* [22]. Berikut rumus pembobotan kata.

$$W_{ij} = tf_{ij} \times log\ log\ \frac{n}{df} \qquad (4)$$

Dimana W_{ij} adalah bobot yang menunjukkan tingkat kepentingan suatu kata atau term, tf_{ij} adalah frekuensi term atau jumlah kali term tersebut muncul, dan *ndf* adalah frekuensi total dokumen yang mengandung term dalam koleksi dokumen.

3.5. *Random Over-Sampling*

Random Oversampling mengatasi ketidakseimbangan kelas dalam dataset dengan menduplikasi sampel data dari kelas minoritas secara acak, memasukkannya ke dalam dataset pelatihan dan meningkatkan ukuran dataset dengan mengulang sampel asli hingga distribusi menjadi seimbang [23]. Metode ini memungkinkan dataset yang lebih seimbang dan meningkatkan kemampuan model untuk mengidentifikasi kelas minoritas [24].

3.6. Klasifikasi SVM

Algoritma ini membangun model yang dapat memprediksi kelas data baru, apakah termasuk dalam salah satu kategori yang telah ditentukan berdasarkan data yang telah dilatih dengan label kelas atau kategori tertentu. Yang dalam penelitian ini adalah kelas positif atau negatif.

3.7. Evaluasi

Tahapan selanjutnya adalah melakukan evaluasi terhadap performa model yang telah dijalankan. Untuk melakukan tahapan evaluasi, *confusion matrix* digunakan, yang menunjukkan hasil prediksi dibandingkan dengan kondisi aktual dari data yang diproses oleh algoritma. Selain itu, *classification report* memberikan gambaran lebih dalam tentang kualitas prediksi melalui metrik evaluasi seperti *precision, recall, accuracy,* dan *F1-Score* [25] . *Precision*, yaitu mengukur keakuratan model ketika memprediksi kelas positif, yang dapat dilihat pada persamaan 5. *Recall* atau *True Positive Rate* adalah ukuran kemampuan model untuk menemukan semua kasus positif, hal ini dapat dilihat pada persamaan 6. *Accuracy* adalah persentase prediksi yang benar terhadap total data, yang dapat dilihat pada persamaan 7. Dan yang terakhir, *F1-Score* adalah menemukan keseimbangan atau *harmonic mean* dai *precision* dan *recall*, yang dapat dilihat pada persamaan 8 [26].

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

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

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (7)$$

$$F1-Score = 2 \times \frac{Precision \times Recall}{Precision+Recall} \quad (8)$$

Digunakan *confusion matrix* untuk mengetahui jumlah TP, TN, FP, dan FN. *True Positive (TP)* adalah data positif yang diprediksi positif oleh model, sedangkan *True Negative (TN)* adalah data negatif yang diprediksi negatif oleh model. *False Positive (FP)* adalah jumlah data yang sebenarnya negatif, tetapi model memprediksi sebagai positif. Dan, *False Negative (FN)* menjelaskan mengenai jumlah data sebenanrya adalah positif, tetapi model memprediksi sebagai negatif.

4. HASIL DAN PEMBAHASAN

4.1. Pengumpulan Data

Dataset ini terdiri dari kumpulan data teks yang menggunakan kata kunci bahasa Indonesia seperti “Bea Cukai”, “Barang Luar Negeri”, dan ”Pengawasan Bea Cukai”. Hasil dari proses *crawling* tersebut disimpan dalam file csv. File tersebut mengandung data tweets sebanyak 9.051 data. Data dikumpulkan dari dari tanggal 1 Juni 2023 hingga 12 Mei 2024. Data yang didapatkan masih perlu diolah terlebih dahulu, sebelum dilakukan modeling sehingga bisa mendapatkan hasil yang maksimal. Pada Tabel 1 menunjukkan output dari crawling data.

Tabel 1. Contoh Hasil Pengumpulan Data

No	Tweet
1	@jaehoney14_ Parah bgt bea cukai... Giliran barang haram lolos. Giliran begini di curigai ada apa dgn negeri ini
2	@roundhatboy @Dry_Sheriff @r_elhakim @dragunovsen @ryanmaretama @beacukaiRI giliran ada yg lolos juga netijen protes kok bisa masuk pengawasan kurang ketat. giliran ketat nanti masyarakat protes kenapa banyak barang yang disita beacukai (pdhl isinya ada narkoba) heran deh.
3	kalian tuh gak sadar ya para fans indonesia? pak STY dkk tu sengaja mengalah biar gak juara 1 karena kalo juara 1 nanti pialanya malah bisa ditahan beacukai kan jadi repot ya?
4	Mau ke kbri rasanya malas betul tapi kalau nanti lihat bawaan koper diacak acak beacukai malah jauh lebih malas dan jengkel
5	@beacukaiRI @Kemendag @BPOM_RI Semoga Allah mempermudah urusan orang-orang yang membuat aturan ini orang-orang bea cukai beserta keluarga mereka. Supaya tidak mempersulit hidup orang lain.

4.2. *Data Preprocessing*

Pada tahap ini, proses dilakukan seperti yang dijelaskan dalam metode penelitian sebelumnya. Hasil preprocessing dilakukan mulai dari *cleansing data, case folding, tokenization,* normalisasi*, stopword*, dan *stemming*. Tahapan ini penting dilakukan, sehingga dapat memaksimalkan hasil dari modeling. Hasil preprocessing ditunjukkan pada Tabel 2-7.

a. *Cleansing data,* data dibersihkan dari karakter, tanda baca, simbol, dan elemen lainnya yang tidak mempengaruhi hasil proses modeling. Pada Tabel 2 menunjukkan contoh hasil cleansing data.

Tabel 2 Contoh Hasil *Cleansing Data*

Sebelum *Cleansing Data*	Sesudah *Cleansing Data*
@jaehoney14_ Parah bgt bea cukai... Giliran barang haram lolos. Giliran begini di curigai ada apa dgn negeri ini	Parah bgt bea cukai Giliran barang haram lolos Giliran begini di curigai ada apa dgn negeri ini
@roundhatboy @Dry_Sheriff @r_elhakim @dragunovsen @ryanmaretama @beacukaiRI giliran ada yg lolos juga netijen protes kok bisa masuk pengawasan kurang ketat. giliran ketat nanti masyarakat protes kenapa banyak barang yang disita beacukai (pdhl isinya ada narkoba) heran deh.	giliran ada yg lolos juga netijen protes kok bisa masuk pengawasan kurang ketat giliran ketat nanti masyarakat protes kenapa banyak barang yang disita beacukai pdhl isinya ada narkoba heran deh.
kalian tuh gak sadar ya para fans indonesia? pak STY dkk tu sengaja mengalah biar gak juara 1 karena kalo juara 1 nanti pialanya malah bisa ditahan beacukai kan jadi repot ya?	kalian tuh gak sadar ya para fans indonesia pak STY dkk tu sengaja mengalah biar gak juara karena kalo juara nanti pialanya malah bisa ditahan beacukai kan jadi repot ya
Mau ke kbri rasanya malas betul tapi kalau nanti lihat bawaan koper diacak acak beacukai malah jauh lebih malas dan jengkel	Mau ke kbri rasanya malas betul tapi kalau nanti lihat bawaan koper diacak acak beacukai malah jauh lebih malas dan jengkel
@beacukaiRI @Kemendag @BPOM_RI Semoga Allah mempermudah urusan orang-orang yang membuat aturan ini orang-orang bea cukai beserta keluarga mereka. Supaya tidak mempersulit hidup orang lain.	Semoga Allah mempermudah urusan orang orang yang membuat aturan ini orang orang bea cukai beserta keluarga mereka Supaya tidak mempersulit hidup orang lain

b. *Case folding*, tahapan ini dilakukan untuk mengurangi variasi yang tidak diperlukan, sehingga setiap kata memiliki arti yang sama. Fungsi *str.lower()* dari library pandas digunakan untuk menangani tahapan *case folding*. Pada Tabel 3 menunjukkan hasil dari proses *case folding.*

Tabel 3 Contoh Hasil *Case Folding*

Sebelum *Case Folding*	Sesudah C*ase Folding*
Parah bgt bea cukai Giliran barang haram lolos Giliran begini di curigai ada apa dgn negeri ini	parah bgt bea cukai giliran barang haram lolos giliran begini di curigai ada apa dgn negeri ini
giliran ada yg lolos juga netijen protes kok bisa masuk pengawasan kurang ketat. giliran ketat nanti masyarakat protes kenapa banyak barang yang disita beacukai pdhl isinya ada narkoba heran deh.	giliran ada yg lolos juga netijen protes kok bisa masuk pengawasan kurang ketat. giliran ketat nanti masyarakat protes kenapa banyak barang yang disita beacukai pdhl isinya ada narkoba heran deh.
kalian tuh gak sadar ya para fans indonesia pak STY dkk tu sengaja mengalah biar gak juara karena kalo juara nanti pialanya malah bisa ditahan beacukai kan jadi repot ya	kalian tuh gak sadar ya para fans indonesia pak sty dkk tu sengaja mengalah biar gak juara karena kalo juara nanti pialanya malah bisa ditahan beacukai kan jadi repot ya
Mau ke kbri rasanya malas betul tapi kalau nanti lihat bawaan koper diacak acak beacukai malah jauh lebih malas dan jengkel	mau ke kbri rasanya malas betul tapi kalau nanti lihat bawaan koper diacak acak beacukai malah jauh lebih malas dan jengkel
Semoga Allah mempermudah urusan orang orangu yang membuat aturan ini orang orang bea cukai beserta keluarga mereka Supaya tidak mempersulit hidup orang lain	semoga allah mempermudah urusan orang orang yang membuat aturan ini orang orang bea cukai beserta keluarga mereka supaya tidak mempersulit hidup orang lain

c. *Tokenization*, proses ini akan memecah teks menjadi bagian yang lebih kecil dapat mempermudah proses pemodelan. Contoh hasil *tokenization* yang ditunjukkan pada Tabel 4.

Tabel 4. Contoh Hasil *Tokenization*

Sebelum *Tokenization*	Sesudah *Tokenization*
parah bgt bea cukai giliran barang haram lolos giliran begini di curigai ada apa dgn negeri ini	[‘parah’, ‘bgt’, ‘bea’, ‘cukai’, ’giliran’, ‘barang’, ‘haram’ ‘lolos’, ’giliran’, ’begini’, ‘di’, ‘curigai’, ‘ada’, ‘apa’, ‘dgn’, ‘negeri’, ‘ini’]
giliran ada yg lolos juga netijen protes kok bisa masuk pengawasan kurang ketat giliran ketat nanti masyarakat protes kenapa banyak barang yang disita beacukai pdhl isinya ada narkoba heran deh	[‘giliran’, ‘ada’, ‘yg’, ‘lolos’, ‘juga’, ‘netijen’,‘protes’, ‘kok’, ‘bisa’, ‘masuk’, ‘pengawasan’, ‘kurang’, ‘ketat’, ‘giliran’, ‘ketat’, ‘nanti’, ‘masyarakat’, ‘protes’, ‘kenapa’, ‘banyak’, ‘barang’, ‘yang’ ‘disita’, ‘bea’ ‘cukai’, ‘pdhl’, ‘isinya’, ‘ada’, ‘narkoba’, ‘heran’, ‘deh’]
kalian tuh gak sadar ya para fans indonesia pak sty dkk tu sengaja mengalah biar gak juara karena kalo juara nanti pialanya malah bisa ditahan beacukai kan jadi repot ya	[‘kalian’, ‘tuh’, ‘gak’, ‘sadar’, ‘ya’, ‘para’, ‘fans’, ‘indonesia’, ‘pak’, ‘sty’, ‘dkk’, ‘tu’, ‘sengaja’, ‘mengalah ‘biar’, ‘gak’, ‘juara’, ‘karena’, ‘kalo’, ‘juara’, ‘nanti’, ‘pialanya’, ‘malah’, ‘bisa’, ‘ditahan’, ‘beacukai’, ‘kan’, ‘jadi’, ‘repot’, ‘ ya’]
mau ke kbri rasanya malas betul tapi kalau nanti lihat bawaan koper diacak acak beacukai malah jauh lebih malas dan jengkel	[‘mau’, ‘ke’, ‘kbri’, ‘rasanya’, ‘malas’, ‘betul’, ‘tapi’, ‘kalau’, ‘nanti’, ‘lihat’, ‘bawaan’, ‘koper’, ‘diacak’, ‘acak’, ‘beacukai’, ‘malah’, ‘jauh’, ‘lebih’, ‘malas’, ‘dan’, ‘jengkel’]
semoga allah mempermudah urusan orang orang yang membuat aturan ini orang orang bea cukai beserta keluarga mereka supaya tidak mempersulit hidup orang lain	[‘semoga’, ‘allah’, ‘mempermudah’, ‘urusan’, ‘orang’, ‘orang’, ‘yang’, ‘membuat’, ‘aturan’, ‘ini’, ‘orang’, ‘orang’, ‘bea’, ‘cukai’, ‘beserta’, ‘keluarga’, ‘mereka’, ‘supaya’, ‘tidak’, ‘mempersulit’, ‘hidup’, ‘orang’, ‘lain’]

d. Normalisasi, pada tahapan ini digunakan kamus normalisasi yang ditemukan di *repository jbkalit* di github bernama normalisasi.csv, dan menambahkan kata-kata normalisasi guna melengkapi kamus normalisasi sebelumnya. Hasil dari normalisasi ditunjukkan pada tabel pada Tabel 5.

Tabel 5 Contoh Hasil Normalisasi

Sebelum Normalisasi	Sesudah Normalisasi
[‘giliran’, ‘ada’, ‘yg’, ‘lolos’, ‘juga’, ‘netijen’,‘protes’, ‘kok’, ‘bisa’, ‘masuk’, ‘pengawasan’, ‘kurang’, ‘ketat’, ‘giliran’, ‘ketat’, ‘nanti’, ‘masyarakat’, ‘protes’, ‘kenapa’, ‘banyak’, ‘barang’, ‘yang’ ‘disita’, ‘bea’, ‘cukai’, ‘pdhl’, ‘isinya’, ‘ada’, ‘narkoba’, ‘heran’, ‘deh’]	[‘giliran’, ‘ada’, ‘yang’, ‘lolos’, ‘juga’, ‘warganet’, ‘protes’, ‘kok’, ‘bisa’, ‘masuk’, ‘pengawasan’, ‘kurang’, ‘ketat’, ‘giliran’, ‘ketat’, ‘nanti’, ‘masyarakat’, ‘protes’, ‘kenapa’, ‘banyak’, ‘barang’, ‘yang’ ‘disita’, ‘bea’, ‘cukai’, ‘padahal’, ‘isinya’, ‘ada’, ‘narkoba’, ‘heran’, ‘deh’]
[‘parah’, ‘bgt’, ‘bea’, ‘cukai’, ’giliran’, ‘barang’, ‘haram’, ‘lolos’, ’giliran’, ’begini’, ‘di’, ‘curigai’, ‘ada’, ‘apa’, ‘dgn’, ‘negeri’, ‘ini’]	[‘parah’, ‘sekali’, ‘bea’, ‘cukai’, ’giliran’, ‘barang’, ‘haram’, ‘lolos’, ’giliran’, ’begini’, ‘di’, ‘curigai’, ‘ada’, ‘apa’, ‘dengan’, ‘negeri’, ‘ini’]
[‘kalian’, ‘tuh’, ‘gak’, ‘sadar’, ‘ya’, ‘para’, ‘fans’, ‘indonesia’, ‘pak’, ‘sty’, ‘dkk’, ‘tu’, ‘sengaja’, ‘mengalah’, ‘biar’, ‘gak’, ‘juara’, ‘karena’, ‘kalo’, ‘juara’, ‘nanti’, ‘pialanya’, ‘malah’, ‘bisa’, ‘ditahan’, ‘beacukai’, ‘kan’, ‘jadi’, ‘repot’, ‘ ya’]	['kalian', 'itu', 'tidak', 'sadar', 'ya', 'para', 'fans', 'indonesia', 'bapak', 'sty', 'dan kawan kawan', 'itu', 'sengaja', 'mengalah', 'agar', 'tidak', 'juara', 'karena', 'kalau', 'juara', 'nanti', 'pialanya', 'malah', 'bisa', 'ditahan', 'bea cukai', 'kan', 'jadi', 'repot', 'ya']
[‘mau’, ‘ke’, ‘kbri’, ‘rasanya’, ‘malas’, ‘betul’, ‘tapi’, ‘kalau’, ‘nanti’, ‘lihat’, ‘bawaan’, ‘koper’, ‘diacak’, ‘acak’, ‘beacukai’, ‘malah’, ‘jauh’, ‘lebih’, ‘malas’, ‘dan’, ‘jengkel’]	['mau', 'ke', 'kbri', 'rasanya', 'malas', 'betul', 'tapi', 'kalau', 'nanti', 'lihat', 'bawaan', 'koper', 'diacak', 'acak', 'beacukai', 'malah', 'jauh', 'lebih', 'malas', 'dan', 'jengkel']
[‘semoga’, ‘allah’, ‘mempermudah’, ‘urusan’, ‘orang’, ‘orang’, ‘yang’, ‘membuat’, ‘aturan’, ‘ini’, ‘orang’, ‘orang’, ‘bea’, ‘cukai’, ‘beserta’, ‘keluarga’, ‘mereka’, ‘supaya’, ‘tidak’, ‘mempersulit’, ‘hidup’, ‘orang’, ‘lain’]	['semoga', 'allah', 'mempermudah', 'urusan', 'orang', 'orang', 'yang', 'membuat', 'aturan', 'ini', 'orang', 'orang', 'bea', 'cukai', 'beserta', 'keluarga', 'mereka', 'supaya', 'tidak', 'mempersulit', 'hidup', 'orang', 'lain']

e. *Stopword removal (filtering),* pada tahapan ini akan menggunakan *library sastrawi* serta menambahkan beberapa kata yang tidak tercantum di *library sastrawi*. Hasil dari proses ini akan ditunjukkan pada Tabel 6.

Tabel 6 Contoh *Stopword Removal*

Sebelum *Stopword Removal*	Sesudah *Stopword Removal*
[‘parah’, ‘bgt’, ‘bea’, ‘cukai’, ’giliran’, ‘barang’, ‘haram’, ‘lolos’, ’giliran’,’begini’, ‘di’, ‘curigai’, ‘ada’, ‘apa’, ‘dgn’, ‘negeri’, ‘ini’]	[‘parah’, ‘bea’, ‘cukai’, ’giliran’, ‘barang’, ‘haram’, ‘lolos’, ’giliran’, ‘curigai’, ‘negeri’]
[‘giliran’, ‘ada’, ‘yang’, ‘lolos’, ‘juga’, ‘warganet’, ‘protes’, ‘kok’, ‘bisa’, ‘masuk’, ‘pengawasan’, ‘kurang’, ‘ketat’, ‘giliran’, ‘ketat’, ‘nanti’, ‘masyarakat’, ‘protes’, ‘kenapa’, ‘banyak’, ‘barang’, ‘yang’ ‘disita’, ‘bea’, ‘cukai’, ‘padahal’, ‘isinya’, ‘ada’, ‘narkoba’, ‘heran’, ‘deh’]	['giliran', 'lolos', 'warganet’', 'protes', 'masuk', 'pengawasan', 'kurang', 'ketat', 'giliran', 'ketat', 'masyarakat', 'protes', 'banyak', 'barang', 'disita', 'bea’, ‘cukai', 'narkoba', 'heran']
['kalian', 'itu', 'tidak', 'sadar', 'ya', 'para', 'fans', 'indonesia', 'bapak', 'sty', 'dan kawan kawan', 'itu', 'sengaja', 'mengalah', 'agar', 'tidak', 'juara', 'karena', 'kalau', 'juara', 'nanti', 'pialanya', 'malah', 'bisa',	['sadar', 'fans', 'indonesia', 'bapak', 'sty', 'kawan kawan', 'sengaja', 'mengalah', 'juara', 'pialanya', 'ditahan', 'bea cukai', 'repot']

'ditahan', 'bea cukai', 'kan', 'jadi', 'repot', 'ya']	
['mau', 'ke', 'kbri', 'rasanya', 'malas', 'betul', 'tapi', 'kalau', 'nanti', 'lihat', 'bawaan', 'koper', 'diacak', 'acak', 'beacukai', 'malah', 'jauh', 'lebih', 'malas', 'dan', 'jengkel']	['kbri', 'malas', 'lihat', 'bawaan', 'koper', 'acak', 'beacukai", 'jauh', 'lebih', 'malas', 'jengkel']
['semoga', 'allah', 'mempermudah', 'urusan', 'orang', 'orang', 'yang', 'membuat', 'aturan', 'ini', 'orang', 'orang', 'bea', 'cukai', 'keluarga', 'mereka', 'supaya', 'tidak', 'mempersulit', 'hidup', 'orang', 'lain']	['semoga', 'allah', 'mempermudah', 'urusan', 'membuat', 'aturan', 'bea', 'cukai', 'keluarga', 'mereka', 'mempersulit', 'hidup', 'lain']

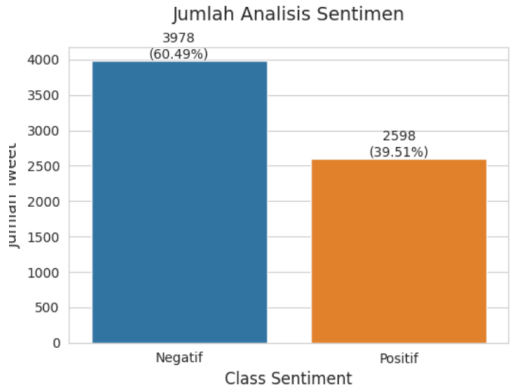
f. *Stemming*, dalam proses ini masih menggunakan library sastrawi. Proses ini akan mengubah setiap kata menjadi bentuk dasarnya. Pada Tabel 7 menunjukkan hasil dari tahapan ini. Setelah melalui proses preprocessing data termasuk proses drop nilai null, terdapat 6.576 data yang siap digunakan.

Tabel 7 Contoh Hasil *Stemming*

Sebelum *Stemming*	Sesudah *Stemming*
[‘parah’, ‘bea’, ‘cukai’, ’giliran’, ‘barang’, ‘haram’, ‘lolos’, ’giliran’, ‘curigai’, ‘negeri’]	[‘parah’, ‘bea’, ‘cukai’, ’gilir’, ‘barang’, ‘haram’, ‘lolos’, ’gilir’, ‘curiga’, ‘negeri’]
[‘giliran’, ‘lolos’, ‘warganet’, ‘protes’, ‘masuk’, ‘pengawasan’, ‘kurang’, ‘ketat’, ‘giliran’, ‘ketat’, ‘masyarakat’, ‘protes’, ‘banyak’, ‘barang’, ‘disita’, ‘bea’, ‘cukai’, ‘narkoba’, ‘heran’]	[‘gilir’, ‘lolos’, ‘warganet’, ‘protes’, ‘masuk’, ‘awas’, ‘kurang’, ‘ketat’, ‘gilir’, ‘ketat’, ‘masyarakat’, ‘protes’, ‘banyak’, ‘barang’, ‘sita’, ‘bea’ ‘cukai’, ‘narkoba’, ‘heran’,]
['sadar', 'fans', 'indonesia', 'bapak', 'sty', 'kawan kawan’, 'sengaja', 'mengalah', 'juara', 'pialanya', 'malah', 'bisa', 'ditahan', 'bea cukai', 'repot']	['sadar', 'fans', 'indonesia', 'bapak', 'sty', 'kawan kawan’', 'sengaja', 'kalah', 'juara', 'piala', 'tahan', 'beacukai', 'repot']
['kbri', 'malas', 'lihat', 'bawaan', 'koper', 'acak', 'beacukai', 'malah', 'jauh', 'lebih', 'malas', 'jengkel']	['kbri', 'malas', 'lihat', 'bawa', 'koper', 'acak', 'beacukai', 'jauh', 'lebih', 'malas', 'jengkel']
['semoga', 'allah', 'mempermudah', 'urusan', 'membuat', 'aturan', 'bea', 'cukai', 'keluarga', 'mereka', 'mempersulit', 'hidup', 'lain']	['semoga', 'allah', 'mudah', 'urus', 'buat', 'atur', 'bea', 'cukai', 'keluarga', 'mereka', 'sulit', 'hidup', 'lain']

4.3. Labeling

Labeling atau Pelabelan menggunakan Kamus Lexicon Bahasa Indonesia. Kamus Lexicon terdiri dari daftar kata positif dan kata negatif, apabila jumlah kata positif dalam teks lebih besar daripada kata negatif, teks tersebut diberi label positif. Sebaliknya, jika jumlah kata negatif lebih banyak daripada kata positif, teks akan diberi label negatif. Proses ini diterapkan untuk setiap teks, dan hasil akhirnya adalah kumpulan sentimen positif atau negatif berdasarkan jumlah kemunculan kata dalam kedua daftar kamus lexicon tersebut. Seperti yang ditunjukkan pada Gambar 4, data yang sudah diberi label dari 6.576 data dengan 2.598 label positif (60,49%) dan 3.978 label negatif (39,51%).



Gambar 4: Jumlah kelas positif dan negatif

Setelah melakukan labeling dengan kelas positif dan negatif, maka langkah selanjutnya adalah menerapkan metode *Encoding*. *Encoding* memastikan bahwa model memproses semua fitur secara konsisten dan menghindari bias model terhadap teks. Pada Tabel 8, metode *encoding* mengubah label sentimen dari teks menjadi format numerik. Dengan kelas negatif diubah menjadi 0 dan kelas positif diubah menjadi 1.

Tabel 8 Contoh Hasil *Encoding*

Teks	*Labeling* Sebelum *Encoding*	*Labeling* Setelah *Encoding*
[‘parah’, ‘bea’, ‘cukai’, ’gilir’, ‘barang’, ‘haram’, ‘lolos’, ’gilir’, ‘curiga’, ‘negeri’]	Negatif	0
[‘gilir’, ‘lolos’, ‘warganet’, ‘protes’, ‘masuk’, ‘awas’, ‘kurang’, ‘ketat’, ‘gilir’, ‘ketat’, ‘masyarakat’, ‘protes’, ‘banyak’, ‘barang’, ‘sita’, ‘bea’ ‘cukai’, ‘narkoba’, ‘heran’]	Positif	1
['sadar', 'fans', 'indonesia', 'bapak', 'sty', 'kawan kawan’', 'sengaja', 'kalah', 'juara', 'piala', 'tahan', 'beacukai', 'repot']	Negatif	0
['kbri', 'malas', 'lihat', 'bawa', 'koper', 'acak', 'beacukai', 'jauh', 'lebih', 'malas', 'jengkel']	Negatif	0
['semoga', 'allah', 'mudah', 'urus', 'buat', 'atur', 'bea', 'cukai', 'keluarga', 'mereka', 'sulit', 'hidup', 'lain']	Positif	1

4.4. TF-IDF

Dalam tahapan ini akan menggunakan *library TfidfVectorizer()* untuk mengubah teks menjadi bentuk numerik, sehingga bisa diproses pada model *machine learning*. Pada Tabel 9 terdapat contoh data sampel sebelum dilakukan perhitungan TF-IDF.

Tabel 9 Contoh data sampel TF-IDF

Doc	Sebelum	Sesudah
D1	Pengawasan bea cukai harus lebih diperketat agar barang ilegal tidak lolos masuk	[‘awas’, ‘bea’, ’cukai’, ’ketat’, ’barang’, ‘ilegal’, ‘lolos’, ‘masuk’]
D2	Giliran barang elektronik diperiksa ketat, barang ilegal malah dibiarkan.	[‘gilir’, ‘barang’, ‘elektronik’, ‘periksa’, ‘ketat’, ‘ilegal’, ‘biar’]
D3	kalo juara nanti pialanya malah bisa ditahan beacukai kan jadi repot ya	[juara', 'piala', 'tahan', 'beacukai', 'repot']
D4	Mau ke kbri rasanya malas betul tapi kalau nanti lihat bawaan koper diacak acak beacukai malah jauh lebih malas dan jengkel	['kbri', 'malas', 'lihat', 'bawa', 'koper', 'acak', 'beacukai', 'jauh', 'lebih', 'malas', 'jengkel']
D5	semoga allah mempermudah urusan orang orang yang membuat aturan ini orang orang bea cukai beserta keluarga mereka supaya tidak mempersulit hidup orang lain	['semoga', 'allah', 'mudah', 'urus', 'buat', 'atur', 'bea', 'cukai', 'keluarga', 'mereka', 'sulit', 'hidup', 'lain']

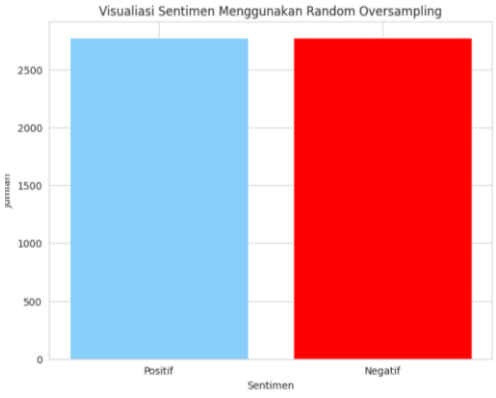
Hasil perhitungan TF-IDF untuk menghitung bobot setiap kata dari Doc D1 dan D2 ditunjukkan pada Tabel 10. Hasil perhitungan TF-IDF untuk kedua sampel data pada Tabel 10 menunjukkan bahwa salah satu nilai pembobotan terhadap kata dengan TF-IDF adalah 0.23299. Dengan kata yang muncul di kedua dokumen memiliki IDF = 0, sehingga TF-IDF = 0.

Tabel 10 Perhitungan TF-IDF

Kata	TF(D1)	TF(D2)	DF	IDF	TF-IDF(D1)	TF-IDF (D2)
Awas	1	0	1	0.301	0.23299	0
Bea	1	0	1	0.301	0.23299	0
Cukai	1	0	1	0.301	0.23299	0
Ketat	1	1	2	0	0	0
Barang	1	2	2	0	0	0
Ilegal	1	1	2	0	0	0
lolos	1	0	1	0.301	0.23299	0
Masuk	1	0	1	0.301	0.23299	0
Gilir	0	1	1	0.301	0	0.23299
Elektronik	0	1	1	0.301	0	0.23299
Periksa	0	1	1	0.301	0	0.23299
Biar	0	1	1	0.301	0	0.23299

4.5. Random Over-Sampling

Tahapan ini melakukan duplikasi pada sampel data kelas minoritas, yaitu kelas positif. Sehingga jumlahnya seimbang dengan kelas mayoritas, yaitu kelas negatif. Tujuan dari tahapan ini membantu model untuk tidak bias terhadap kelas mayoritas dan belajar dengan baik pada kedua kelas. Hasil dari *random over-sampling* terdapat pada Gambar 5.



Gambar 5: Hasil Sentimen menggunakan *random over-sampling*

4.6. Klasifikasi SVM

Tahap awal sebelum memulai proses klasifikasi, yaitu data dibagi terlebih dahulu. Pembagian ini dilakukan dengan membagi data ke dalam dua kategori, yaitu data *training* dan data *testing*, dengan rasio pembagian data 70:30. Dengan data *training* sudah dilakukan metode *random over-sampling*. Penelitian ini menggunakan klasifikasi SVM dengan kernel linear. Ini dilakukan dengan menguji sejumlah nilai parameter C yang akan ditunjukkan pada Tabel 11.

Tabel 11 Nilai parameter C

Parameter C	Akurasi
0.01	86.92%
0.05	90.31%
0.25	92.80%
0.5	93.25%
0.75	93.66%
1.0	93.87%

Hasil eksperimen menunjukkan bahwa parameter C sangat mempengaruhi akurasi model. Seiring dengan peningkatan nilai C, akurasi meningkat secara bertahap hingga mencapai nilai tertinggi C = 1.0, dengan akurasi 93.87%. Setelah C mencapai 0.5 peningkatan akurasi mulai melambat menunjukkan bahwa model telah mencapai titik optimalnya. Hasil eksperimen menunjukkan bahwa kombinasi proses sebelumnya dan SVM dengan pemilihan parameter terbaik, yaitu C = 1.0 adalah pilihan efektif untuk analisis sentimen. Selain itu, penggunaan *CalibratedClassifierCV* atau kalibrasi model menunjukkan bahwa estimasi probabilitas yang lebih akurat, yang berarti bahwa hasilnya dapat diandalkan untuk membuat pengambilan keputusan yang lebih baik.

4.7. Evaluasi

Hasil pengujian terakhir model *Support Vector Machine (SVM)* ditunjukkan pada Gambar 6. Hasil ini menunjukkan bahwa model memiliki kemampuan sangat baik dalam melakukan klasifikasi sentimen. Model menunjukkan akurasi 93.87% yang menunjukkan bahwa model memiliki tingkat keberhasilan yang tinggi dalam mengidentifikasi sentimen dalam teks yang dianalisis.

```
[ ]  # Final Accuracy Score Model

     print('Accuracy score model final: %s ' %accuracy_score(y_test, svm.predict(X_test_vec)))
```

```
Accuracy score model final: 0.9386720729853015
```

Gambar 6: *Final Accuracy Score Model*

```
Akurasi Model Support Vector Machine : 0.9387

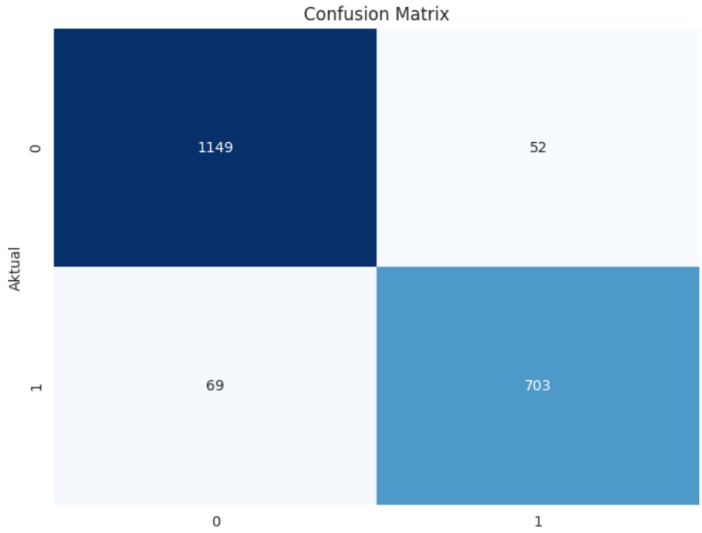
Laporan Klasifikasi :
              precision    recall  f1-score   support

           0       0.94      0.96      0.95      1201
           1       0.93      0.91      0.92       772

    accuracy                           0.94      1973
   macro avg       0.94      0.93      0.94      1973
weighted avg       0.94      0.94      0.94      1973
```

Gambar 7: *Classification Report*

Classification Report pada Gambar 7 menunjukkan bahwa model bekerja dengan sangat baik pada setiap kelas sentimen. Untuk kelas 0 (sentimen negatif) mengidentifikasi 96% dari seluruh sentimen negatif yang ada *(recall)* dan 94% dari prediksi sentimen negatif yang dibuat adalah benar *(precision)*. Serta, untuk kelas 1 (sentimen positif) menunjukkan bahwa model juga efektif dalam mengidentifikasi sentimen positif dengan berhasil mengenali 91% dari seluruh sentimen positif yang ada dan 93% dari prediksi sentimen positif yang dibuat adalah tepat. Secara keseluruhan, model menunjukkan kinerja yang seimbang dengan *precision, recall,* dan *f1-score* masing-masing sebesar 0.94 menunjukkan konsistensi model dalam klasifikasi kedua kelas sentimen.



Gambar 8: *Confusion Matrix*

Pada Gambar 8 menunjukkan performa model dalam bentuk matriks 2 × 2 dengan total 1973 total sampel yang diuji *(testing)*:

- *True Negative (TN)*: 1149 sampel berhasil diidentifikasi dengan benar sebagai sentimen negatif
- *False Positive (FP)*: 52 sentimen negatif salah diklasifikasikan sebagai positif
- *False Negative (FN)*: 69 sampel sentimen positif salah diklasifikasikan sebagai negatif
- *True Positive (TP)*: 703 sampel berhasil diidentifikasi sebagai sentimen positif.

Hasil menunjukkan bahwa model memiliki tingkat kesalahan klasifikasi yang sangat rendah; dari 1973 sampel, hanya 121 kesalahan klasifikasi.

```
# Hitung Classification Error
classification_error = 1 - svm.score(X_test_vec, y_test)

print('Classification Error: {:.4f}'.format(classification_error))
```

```
Classification Error: 0.0613
```

Gambar 8: *Classification Error*

Pada Gambar 8 perhitungan *classification Error* menghasilkan nilai 0.0613 yang merupakan komplemen dari nilai akurasi. Model menunjukkan keakuratan yang tinggi dalam klasifikasi sentimen dengan hanya melakukan kesalahan pada sekitar 6 kasus dari setiap 100 prediksi yang dibuat.

V. SIMPULAN DAN SARAN

Analisis sentimen digunakan untuk mengidentifikasi persepsi masyarakat atau publik terhadap isu tertentu. Hasil dari analisis sentimen bea cukai terhadap pengawasan barang dari luar negeri memberikan pandangan bagaimana masyarakat dalam menilai institusi pemerintah, khususnya bea cukai. Dengan data sebanyak 6.576 entri setelah dilakukan proses preprocessing, yang dibagi dengan rasio 70:30. Analisis sentimen bea cukai terhadap pengawasan barang dari luar negeri menggunakan metode SVM menunjukkan akurasi sebesar 93.87%, *precision* sebesar 94%, *recall* sebesar 93%, dan *f1-score* sebesar 94%. Hasil evaluasi ini menunjukkan bahwa metode SVM efektif dalam menganalisis sentimen, khususnya bea cukai.

Selain itu, hasil penelitian ini menunjukkan bahwa dari data yang dilabeli, terdapat 60,49% sentimen positif dan 39,51% sentimen negatif, yang menunjukkan bahwa sentimen negatif terhadap kinerja bea cukai dalam pengawasan barang dari luar negeri lebih dominan. Hal ini dapat menjadi perhatian pemerintah untuk mengevaluasi dan memperbaiki layanan bea cukai, guna meningkatkan kepercayaan dan kepuasaan

masyarakat terhadap kinerja bea cukai, serta dapat mendorong transparansi dan efisiensi dalam proses pengawasan barang dari luar negeri. Penelitian ini dapat memberikan dasar untuk penelitian lebih lanjut. Seperti, memperluas cakupan data dari berbagai media sosial untuk menghasilkan analisis yang lebih mendalam. Pada penelitian ini hanya menggunakan data dari Twitter. Saran, kepada penelitian kedepannya dapat menggunakan opii dari berbagai platform media sosial.

DAFTAR PUSTAKA

[1] N. S. Amanda and N. Aslami, "Analisis Kebijakan Perdagangan Internasional," *Journal Economy And Currency Study (JECS)*, vol. 4, no. 1, pp. 14–23, 2022.

[2] C. E. Stefhani, "Peran Direktorat Jenderal Bea Cukai Dalam Memberantas Penyelundupan Narkotika Melalui Bandar Udara Internasional Soekarno-Hatta.," *Jurnal Indonesia Sosial Teknologi*, vol. 3, no. 12, pp. 1335–1347, 2022, doi: 10.36418/jist.v3i12.540.

[3] A. C. Ramdani, H. R. Budiana, F. X. Ari, and A. Prastowo, "Manajemen Krisis Direktorat Jenderal Bea dan Cukai Kementerian Keuangan dalam Mengembalikan Kepercayaan Publik," vol. 2, no. September, pp. 67–83, 2024.

[4] N. Fajriyah, N. T. Lapatta, D. W. Nugraha, and R. Laila, "IMPLEMENTASI SVM DAN SMOTE PADA ANALISIS SENTIMEN MEDIA SOSIAL X TERHADAP PELANTIKAN AGUS HARIMURTI YUDHOYONO," *JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)*, vol. 10, no. 2, pp. 1359–1370, 2025.

[5] Syahril Dwi Prasetyo, Shofa Shofiah Hilabi, and Fitri Nurapriani, "Analisis Sentimen Relokasi Ibukota Nusantara Menggunakan Algoritma Naïve Bayes dan KNN," *Jurnal KomtekInfo*, vol. 10, pp. 1–7, 2023, doi: 10.35134/komtekinfo.v10i1.330.

[6] A. P. Nardilasari, A. L. Hananto, S. S. Hilabi, T. Tukino, and B. Priyatna, "Analisis Sentimen Calon Presiden 2024 Menggunakan Algoritma SVM Pada Media Sosial Twitter," *JOINTECS (Journal of Information Technology and Computer Science)*, vol. 8, no. 1, p. 11, Mar. 2023, doi: 10.31328/JOINTECS.V8I1.4265.

[7] A. A. Munandar, F. Farikhin, and C. E. Widodo, "Sentimen Analisis Aplikasi Belajar Online Menggunakan Klasifikasi SVM," *JOINTECS (Journal of Information Technology and Computer Science)*, vol. 8, no. 2, p. 77, Jul. 2023, doi: 10.31328/JOINTECS.V8I2.4747.

[8] W. Wahyuni, "Analisis Sentimen terhadap Opini Feminisme Menggunakan Metode Naive Bayes," *Jurnal Informatika Ekonomi Bisnis*, pp. 148–153, Sep. 2022, doi: 10.37034/INFEB.V4I4.162.

[9] D. A. P. Maharani, I Made Mulyawan Subawa, and I Gusti Ika Laksmi Mahadewi, "Pengawasan Pelayanan Bea dan Cukai Dalam Perdagangan Ekspor Di Bandara Ngurah Rai," *AL-DALIL: Jurnal Ilmu Sosial, Politik, dan Hukum*, vol. 1, no. 3, pp. 29–38, 2023, doi: 10.58707/aldalil.v1i3.567.

[10] F. F. Rachman and S. Pramana, "Analisis Sentimen Pro dan Kontra Masyarakat Indonesia tentang Vaksin COVID-19 pada Media Sosial Twitter," *Health Information Management Journal*, vol. 8, no. 2, pp. 100–109, 2020, [Online]. Available: https://inohim.esaunggul.ac.id/index.php/INO/article/view/223/175

[11] H. Irmanda and R. Astriratma, "Klasifikasi Jenis Pantun Dengan Metode Support Vector Machines (SVM)," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 4, pp. 915–922, Oct. 2020, doi: 10.29207/resti.v4i5.2313.

[12] P. Chowdhuri, P. Pal, and T. Si, "A novel steganographic technique for medical image using SVM and IWT," *Multimed Tools Appl*, vol. 82, no. 13, pp. 20497–20516, 2023, doi: 10.1007/s11042-022-14301-0.

[13] O. Rahman, G. Abdillah, and A. Komarudin, "Klasifikasi Ujaran Kebencian pada Media Sosial Twitter Menggunakan Support Vector Machine," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 5, pp. 17–23, Feb. 2021, doi: 10.29207/resti.v5i1.2700.

[14] S. Rahayu and Y. Yamasari, "Klasifikasi Penyakit Stroke dengan Metode Support Vector Machine (SVM)," *Journal of Informatics and Computer Science (JINACS)*, vol. 5, no. 03, pp. 440–446, 2024.

[15] P. Fremmuzar and A. Baita, "Uji Kernel SVM dalam Analisis Sentimen Terhadap Layanan Telkomsel

di Media Sosial Twitter," *Komputika: Jurnal Sistem Komputer*, vol. 12, no. 2, pp. 57–66, 2023.

[16] P. Arsi and R. Waluyo, "Analisis Sentimen Wacana Pemindahan Ibu Kota Indonesia Menggunakan Algoritma Support Vector Machine (SVM)," *Jurnal Teknologi Informasi dan Ilmu Komputer*, vol. 8, no. 1, p. 147, Feb. 2021, doi: 10.25126/JTIIK.0813944.

[17] F. Putra, H. F. Tahiyat, R. M. Ihsan, R. Rahmaddeni, and L. Efrizoni, "Penerapan Algoritma K-Nearest Neighbor Menggunakan Wrapper Sebagai Preprocessing untuk Penentuan Keterangan Berat Badan Manusia: Application of K-Nearest Neighbor Algorithm Using Wrapper as Preprocessing for Determination of Human Weight Information," *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, vol. 4, no. 1, pp. 273–281, 2024.

[18] M. D. Purbolaksono, M. Irvan Tantowi, A. Imam Hidayat, and A. Adiwijaya, "Perbandingan Support Vector Machine dan Modified Balanced Random Forest dalam Deteksi Pasien Penyakit Diabetes," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 5, no. 2, pp. 393–399, 2021, doi: 10.29207/resti.v5i2.3008.

[19] D. Darwis, E. S. Pratiwi, and A. F. O. Pasaribu, "PENERAPAN ALGORITMA SVM UNTUK ANALISIS SENTIMEN PADA DATA TWITTER KOMISI PEMBERANTASAN KORUPSI REPUBLIK INDONESIA," *Edutic - Scientific Journal of Informatics Education*, vol. 7, no. 1, Nov. 2020, doi: 10.21107/EDUTIC.V7I1.8779.

[20] M. R. Arjunanto and S. Waluyo, "Analisis Sentimen Pada Media Sosial Twitter Terhadap Infrastruktur Jalan Di Lampung Menggunakan Algoritma K-Nearest Neighbor," *Prosiding Seminar Nasional ...*, vol. 2, no. September, pp. 810–817, 2023, [Online]. Available: http://senafti.budiluhur.ac.id/index.php/senafti/article/view/914%0Ahttps://senafti.budiluhur.ac.id/index.php/senafti/article/download/914/557

[21] R. A. Supono and Muhammad Azis Suprayogi, "Perbandingan Metode TF-ABS dan TF-IDF Pada Klasifikasi Teks Helpdesk Menggunakan K-Nearest Neighbor," vol. 5, no. 10, pp. 911–918, 2021.

[22] S. Rabbani, D. Safitri, N. Rahmadhani, A. A. F. Sani, and M. K. Anam, "Perbandingan Evaluasi Kernel SVM untuk Klasifikasi Sentimen dalam Analisis Kenaikan Harga BBM," *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, vol. 3, no. 2, pp. 153–160, 2023, doi: 10.57152/malcom.v3i2.897.

[23] U. Hasanah, A. M. Soleh, and K. Sadik, "Effect of Random Under sampling , Oversampling , and SMOTE on the Performance of Cardiovascular Disease Prediction Models," vol. 21, no. 1, pp. 88–102, 2024, doi: 10.20956/j.v21i1.35552.

[24] R. Aryanti, T. Misriati, and R. Hidayat, "Klasifikasi Risiko Kesehatan Ibu Hamil Menggunakan Random Oversampling Untuk Mengatasi Ketidakseimbangan Data," *Media Online*, vol. 3, no. 5, pp. 409–416, 2023, [Online]. Available: https://djournals.com/klik

[25] N. L. P. C. Savitri, R. A. Rahman, R. Venyutzky, and N. A. Rakhmawati, "Analisis klasifikasi sentimen terhadap sekolah daring pada twitter menggunakan Supervised Machine Learning," *Jurnal Teknik Informatika dan Sistem Informasi*, vol. 7, no. 1, 2021.

[26] N. Azizah, M. Riyad Firdaus, R. Suyaningsih, and F. Indrayatna, "Penerapan Algoritma Klasifikasi K-Nearest Neighbor pada Penyakit Diabetes," *Prosiding Seminar Nasional Statistika Aktuaria*, vol. 2, no. 1 SE-Articles, pp. 119–126, Aug. 2023, doi: 10.1234/snsa.v2i1.344.