Building of Informatics, Technology and Science (BITS)
Volume 7, No 1, June 2025 Page: 814−822
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i1.7411

Analisis Sentimen Rencana Penerapan Cukai Pada Minuman Manis Kemasan Menggunakan Algoritma Naive Bayes dan Logistic Regression

Gozali Gozali [*], Kiki Ahmad Baihaqi, Cici Emilia Sukmawati, Deden Wahiddin

Fakultas Ilmu Komputer, Program Studi Teknik Informatika, Universitas Buana Perjuangan Karawang, Karawang, Indonesia
Email: [1,*]if21.gozali@mhs.ubpkarawang.ac.id, [2]kikiahmad@ubpkarawang.ac.id, [3]cici.emilia@ubpkarawang.ac.id, [4]deden.wahiddin@ubpkarawang.ac.id
Email Penulis Korespondensi: if21.gozali@mhs.ubpkarawang.ac.id
Submitted: **20/05/2025**; Accepted: **30/06/2025**; Published**: 30/06/2025**

Abstrak−Rencana pemberlakuan cukai terhadap minuman berpemanis dalam kemasan (MBDK) diusulkan sebagai langkah strategis untuk menekan tingkat konsumsi gula di kalangan masyarakat. Kebijakan ini memicu berbagai respons dari masyarakat, khususnya melalui platform media sosial seperti *TikTok.* Tujuan dari penelitian ini adalah untuk mengevaluasi sentimen yang muncul di masyarakat terhadap kebijakan cukai MBDK melalui komentar-komentar masyarakat yang disampaikan pada platform TikTok, dengan membandingkan performa algoritma *Naive Bayes* dan Logistic Regression. Data dikumpulkan dari komentar video pemberitaan penerapan cukai MBDK pada akun jurnalis resmi di TikTok menggunakan TikTok Comments Scraper yang tersedia di web apipy dan memperoleh 1.332. Data dianalisis melalui tahapan pra-pemrosesan seperti pembersihan teks, tokenisasi, stemming, serta pembobotan kata menggunakan *TF-IDF*. Setelah pelabelan sentimen oleh pakar, Data kemudian dipisahkan menjadi *training set* dan *testing set* dengan perbandingan 80 banding 20. Evaluasi dilakukan menggunakan *confusion matrix* untuk memperoleh Kinerja masing-masing model dievaluasi berdasarkan nilai *accuracy*, *precision*, *recall*, dan *F1-score*. Berdasarkan hasil analisis, diketahui bahwa komentar negatif mendominasi sebesar 65,2%, sedangkan komentar positif sebesar 34,8%. *Logistic Regression* memperoleh akurasi 81,37%, *precision* 86,22%, *recall* 75,14%, dan *F1-score* 77,06%. Sedangkan *Naive Bayes* memperoleh akurasi 79,85%, *precision* 82,19%, *recall* 74,17%, dan *F1-score* 75,76%. Dapat disimpulkan bahwa mayoritas pengguna TikTok masih memberikan tanggapan negatif terhadap kebijakan cukai MBDK dan algoritma *Logistic Regression* Menunjukkan kinerja yang lebih unggul dalam melakukan klasifikasi sentimen jika dibandingkan dengan algoritma *Naive Bayes*.

Kata Kunci: Komentar; Logistic Regression; MBDK; Naive Bayes; Sentimen; TikTok

Abstract−The plan to impose excise tax on packaged sweetened beverages (PSB) is proposed as a strategic measure to reduce sugar consumption among the public. This policy has elicited various responses from society, especially on social media platforms such as TikTok. The purpose of this study is to evaluate public sentiment towards the PSB excise tax policy by analyzing comments posted on the TikTok platform, comparing the performance of the Naive Bayes and Logistic Regression algorithms. Data were collected from comments on news videos about the implementation of the excise tax on PSB posted by official journalist accounts on TikTok, using the TikTok Comments Scraper available on the apipy website, resulting in 1,332 comments. The data were processed through preprocessing steps including text cleaning, tokenization, stemming, and word weighting using TF-IDF. After expert sentiment labeling, the data were then split into training and testing sets with an 80:20 ratio. Evaluation was conducted using a confusion matrix to obtain performance metrics such as accuracy, precision, recall, and F1-score for each model. The analysis revealed that negative comments dominated at 65.2%, while positive comments accounted for 34.8%. The Logistic Regression algorithm achieved an accuracy of 81.37%, precision of 86.22%, recall of 75.14%, and an F1-score of 77.06%. Meanwhile, the Naive Bayes algorithm obtained an accuracy of 79.85%, precision of 82.19%, recall of 74.17%, and an F1-score of 75.76%. It can be concluded that the majority of TikTok users still express negative responses to the PSB excise tax policy, and the Logistic Regression algorithm demonstrates superior performance in sentiment classification compared to the Naive Bayes algorithm.

Keywords: Comments; Logistic Regression; Naive Bayes; Sentiment; Sugar-Sweetened Beverages (SSB); TikTok

1. PENDAHULUAN

Media sosial biasa digunakan sebagai wadah untuk masyarakat dalam menyuarakan ide, gagasan, pendapat, dan kritikan termasuk Kebijakan penerapan cukai terhadap minuman berpemanis dalam kemasan (MBDK)[1], TikTok Merupakan salah satu *platform* media sosial yang saat ini sangat populer dan banyak digunakan oleh pengguna di seluruh dunia, dan banyak pengguna berbagai macam kalangan[2].

Berdasarkan *IDF Diabetes Atlas 2025* dari International Diabetes Federation (IDF), jumlah orang dewasa berusia 20 hingga 79 tahun yang hidup dengan diabetes di Indonesia diperkirakan mencapai 20,4 juta jiwa pada tahun 2024. Pencapaian ini menempatkan Indonesia pada posisi kelima sebagai negara dengan jumlah penderita diabetes tertinggi di dunia. Fakta tersebut mencerminkan meningkatnya beban masalah kesehatan masyarakat yang memerlukan perhatian khusus dan penanganan intensif dari berbagai pihak terkait[3].

Pengenaan cukai pada MBDK dinilai sebagai langkah strategis dalam upaya membatasi pola penggunaan masyarakat terhadap gula, serta menekan biaya penanganan penyakit akibat konsumsi gula berlebih. [4]. Untuk mengendalikan tingkat obesitas, penerapan aturan terhadap tarif cukai minuman berpemanis dalam kemasan (MBDK) perlu dilaksanakan. Pengenaan cukai tersebut bisa menjadi solusi untuk mengurangi konsumsi gula berlebih. Untuk mewujudkannya, Kementerian Kesehatan (Kemenkes) sudah mengirim surat kepada Kementerian Keuangan mengenai usulan penerapan cukai terhadap minuman berpemanis dalam kemasan (MBDK) [5]. Menurut Undang-Undang Nomor 39 Tahun 2007 tentang Cukai, barang yang dapat dikenakan cukai adalah barang yang konsumsinya perlu diawasi atau menimbulkan pengaruh negatif terhadap kesehatan dan lingkungan. Berdasarkan hal tersebut, pemerintah mengkategorikan MBDK sebagai barang yang layak dikenai cukai[6].

Building of Informatics, Technology and Science (BITS)
Volume 7, No 1, June 2025 Page: 814−822
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i1.7411

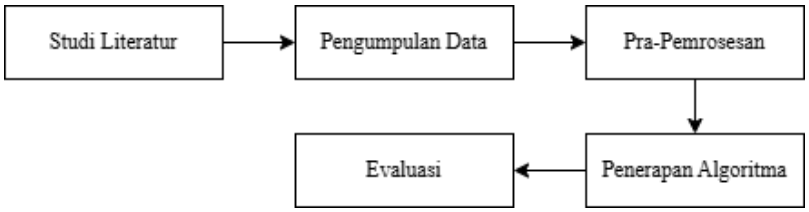
Kebijakan pengenaan cukai pada MBDK menimbulkan respons beragam dari masyarakat terutama pengguna TikTok, Beragam respons muncul terhadap kebijakan tersebut, mencakup pendapat yang mendukung maupun yang menolak. Situasi ini membuka peluang untuk dilakukan analisis sentimen, dengan memanfaatkan data yang bersumber dari Tiktok[7]. Analisis sentimen adalah teknik pengolahan teks yang digunakan untuk mengidentifikasi pandangan masyarakat yang mencakup sentimen positif, negatif maupun netral. Proses ini berfokus pada penggalian emosi atau pandangan publik yang terekspresi melalui teks yang diunggah di media sosial [8]. Di bidang pemerintahan, pemanfaatan analisis sentimen dapat membantu mengidentifikasi sudut pandang masyarakat mengenai kebijakan yang dijalankan pemerintah [9].

Berdasarkan penelitian terdahulu yang telah dilakukan oleh Setiawan, dkk, menunjukan bahwa analisis sentimen pada media sosial seperti TikTok telah banyak dimanfaatkan untuk mengukur persepsi publik terhadap suatu produk maupun kebijakan[10]. Berdasarkan hasil-hasil penelitian terdahulu, algoritma *Logistic Regression* dan *Naïve Bayes* menunjukkan performa yang cukup kuat dan konsisten dalam tugas klasifikasi sentimen berbasis teks, seperti yang ditunjukkan oleh Pratama, dkk., yang menggunakan algoritma *Naïve Bayes* dalam mengkaji komentar di masyarakat terhadap regulasi pengadaan gas LPG 3 Kg melalui identitas KTP. Penelitian tersebut mampu mencapai akurasi sebesar 84% setelah melalui tahapan penting dalam pemrosesan data teks, termasuk *cleansing*, *normalization*, *tokenization*, *stemming, stopword removal*, serta pembobotan menggunakan teknik TF-IDF [11]. Insan, dkk, juga menunjukkan performa *Naïve Bayes* dapat mengklasifikasikan sentimen ulasan aplikasi Brimo dengan akurasi sebesar 84,52% dan nilai *precision* serta *recall* yang tinggi[12]. Di sisi lain, Rizal, dkk. telah dimanfaatkan *Logistic Regression* secara optimal sebagai upaya mengevaluasi sentimen pengguna TikTok, dengan akurasi mencapai 83%.menjadikannya model yang dapat diandalkan untuk klasifikasi sentimen[13]. Temuan serupa juga dikemukakan oleh Savitri, dkk, yang dalam perbandingan beberapa algoritma supervised learning menunjukkan bahwa *Logistic Regression* memperoleh akurasi tertinggi sebesar 87% dalam menganalisis sentimen terhadap sekolah daring[14]. Meskipun berbagai penelitian sebelumnya telah membuktikan efektivitas penggunaan *Naïve Bayes* dan *Logistic Regression* dalam pengklasifikasian sentimen, terutama pada isu kebijakan publik dan ulasan aplikasi digital, namun belum ada yang secara spesifik menganalisis sentimen publik terhadap kebijakan pengenaan cukai pada minuman manis dalam kemasan menggunakan data komentar TikTok. sekaligus memperkaya kajian akademik dalam pemanfaatan media sosial sebagai alat evaluasi kebijakan berbasis data. Berdasarkan bukti empiris tersebut, pemilihan *Naïve Bayes* dan *Logistic Regression* dalam penelitian ini dinilai relevan untuk digunakan dalam konteks klasifikasi sentimen terhadap komentar pengguna TikTok terkait isu pengenaan cukai pada minuman manis dalam kemasan.

2. METODOLOGI PENELITIAN

2.1 Tahapan Penelitian

Semua proses yang dilakukan dalam penelitian ini dirancang guna memastikan kelancaran proses pelaksanaan, sehingga diperoleh hasil yang tersusun secara terstruktur. Urutan tahapan tersebut ditunjukkan pada Gambar 1.



Gambar 1. Tahapan Penelitian

Sebagaimana terlihat pada Gambar 1, penelitian dimulai dengan tahapan studi literatur. Selanjutnya, dilakukan tahap pengumpulan data komentar dari video pemberitaan penerapan cukai pada minuman manis kemasan pada tiktok. Setelah data terkumpul, dataset tersebut diproses melalui tahapan pra-pemrosesan sebelum penerapan algoritma. Tahap berikutnya adalah penerapan algoritma menggunakan *Logistic Regression* dan *Naive Bayes*. Terakhir, penelitian dievaluasi berdasarkan analisis *confusion matrix*.

2.1.1 Studi Literatur

Penelitian ini diawali dengan studi literatur guna memperoleh pemahaman yang mendalam mengenai algoritma *Logistic Regression*, *Naive Bayes*, serta penerapan analisis sentimen menggunakan data komentar dari platform TikTok. Literatur yang digunakan bersumber dari jurnal ilmiah, situs daring, dan buku referensi.

2.1.2 Pengumpulan Data

Tahap pengumpulan data menggunakan TikTok *Comments Scraper* yang tersedia di situs web *apify* untuk mengabil kometar-kometar di media sosial Tiktok,kometar-kometar yang diambil adalah komentar dari video pemberitaan

Building of Informatics, Technology and Science (BITS)
Volume 7, No 1, June 2025 Page: 814−822
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i1.7411

penerapan cukai yang di ungguh di platform media sosial TikTok, oleh akun resmi jurnalis seperti CNN Indonesia, Kompas.com dan kumparan.

2.1.3 Pra-Pemrosesan

Pada tahapan ini, data yang digunakan dalam penelitian akan dinormalisasi agar dapat diproses ke tahap analisis berikutnya. Proses pra-pemrosesan ini terdiri dari beberapa langkah, termasuk pembersihan data (*cleaning*), penyeragaman huruf (*case folding*), pemisahan kata (*tokenisasi*), Transformasi kata menjadi bentuk dasarnya (*stemming*), serta penghapusan kata-kata yang bersifat umum (*stopword removal*). Seluruh tahapan tersebut dilakukan dalam penelitian ini untuk mengubah data teks menjadi bentuk yang lebih terstruktur dan siap untuk dianalisis:

a. Data *Cleansing* : Tujuan dari data cleansing adalah menghilangkan atribut-atribut yang tidak berfungsi untuk analisis, seperti hashtag, *URL*, tag *HTML*, dan lainnya.
b. *Case Folding* : Proses case folding dilakukan untuk menyamakan format penulisan huruf dengan cara mengubah seluruh karakter menjadi huruf kecil (lowercase).
c. Normalisasi Kata : Normalisasi terhadap kata-kata dalam teks dengan cara mengubahnya ke dalam bentuk baku atau standar yang sesuai kaidah bahasa.
d. *Tokenizing* : bertujuan untuk memisahkan teks menjadi token, seperti kata atau frasa.
e. *Stemming* : Proses ini dilakukan untuk mengembalikan kata dalam teks ke bentuk dasarnya dengan menghilangkan berbagai imbuhan, baik yang berada di awal, akhir, maupun tengah kata.
f. *Remove Stop Word* : Digunakan untuk menghapus kata-kata umum yang tidak memiliki nilai informasi signifikan dalam proses analisis, seperti 'dan', 'atau', 'yang', 'di', 'ke', serta kata hubung lainnya.
g. Pemboboran Kata : Pembobotan kata menggunakan TF-IDF (*Term Frequency-Inverse Document Frequency*) digunakan untuk mentransformasikan data berbentuk teks ke dalam representasi numerik. Metode ini merupakan teknik statistik bertujuan untuk menghitung frekuensi kemunculan sebuah kata dalam dataset secara keseluruhan.
h. Pelabelan Data : bertujuan untuk membagi data menjadi 2 bagian dengan diberikan label positive dan negative

2.1.4 World Cloud

Word cloud adalah representasi visual yang menampilkan kumpulan kata-kata dari sebuah teks yang menampilkan kata-kata dominan dari sebuah kumpulan data secara grafis [15]. *Word cloud* digunakan untuk menyajikan visualisasi kata-kata yang paling sering muncul dalam suatu kumpulan data teks secara grafis. Frekuensi kemunculan kata tercermin melalui variasi ukuran huruf; Kata-kata dengan frekuensi kemunculan yang lebih tinggi akan ditampilkan dengan ukuran yang lebih besar dibandingkan dengan kata yang lebih jarang muncul.[16].

2.1.5 Implementasi Algoritma

Implementasi algoritma menggunakan dua algoritma, yaitu *Naive Bayes* dan *Logistic Regression*, dirancang untuk menangani klasifikasi dua kelas, untuk mengelompokkan data yang terstruktur secara sistematis [17]. dengan proses persiapan data yang sudah melalui tahapan pembobotan TF-IDF lalu pembagian dataset berlabel Dibagi menjadi data pelatihan dan data pengujian dengan perbandingan 80 banding 20.

a. Algoritma *Logistik Regression*
Regresi logistik termasuk salah satu metode *regresi* yang memiliki tingkat efektivitas tinggi dalam menyelesaikan masalah klasifikasi. Teknik ini memungkinkan analisis dan pemodelan hubungan antara variabel input dan output dengan cara yang sederhana dan mudah dimengerti[18]. Persamaan model regresi logistik tertera sebagai berikut:

$$\pi(x_i) = \frac{\exp(\beta_0+\beta_1 x_1+ \beta_2 x_2+\cdots+\beta_p x_p)}{1+\exp(\beta_0+\beta_1 x_1+ \beta_2 x_2+\cdots+ \beta_p x_p)} \quad (1)$$

$\pi\ (x_i)$ adalah probabilitas tertentu, β_0 merupakan intercept, $\beta_1\beta_2, \ldots, \beta_p$ adalah koefisien variabel independen, $x_1x_2, \ldots, x_p$ adalah nilai variabel independen.

b. Algoritma *Naive Bayes*
Naive Bayes adalah metode klasifikasi yang didasarkan pada teorema Bayes. Metode ini sangat cocok diterapkan ketika terdapat jumlah data masukan yang besar. Teknik klasifikasi ini sering dipilih karena keunggulannya dalam hal kecepatan dan kesederhanaan [19], [20], Selain itu, *Naive Bayes* memiliki keunggulan dengan waktu klasifikasi yang relatif singkat, metode ini mampu mempercepat proses analisis sentimen secara keseluruhan [21], Teorema Bayes dapat ditulis dalam bentuk berikut:

$$P(H|X) = \frac{P(X|H)\cdot P(H)}{P(X)} \quad (2)$$

X merupakan data dengan kelas yang belum diketahui, sedangkan H adalah hipotesis yang menyatakan bahwa data X termasuk ke dalam kelas tertentu. P(H|X) melambangkan probabilitas posterior hipotesis H berdasarkan data X, sementara P(H) adalah probabilitas awal (prior) dari hipotesis tersebut. P(X|H) menggambarkan kemungkinan munculnya data X jika hipotesis H benar, dan P(X) merupakan probabilitas keseluruhan terjadinya data X.

Building of Informatics, Technology and Science (BITS)
Volume 7, No 1, June 2025 Page: 814−822
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i1.7411

2.1.6 Evaluasi

Setelah proses algoritma selesai, data dievaluasi dengan *confusion Matrix* guna mengukur efektivitas klasifikasi melalui nilai akurasi, *precision*, dan *recall,* metode ini bisa penghitungan tingkat keakuratan atau kebenaran dalam proses klasifikasi. Dengan menggunakan *confusion matrix*, dapat dianalisis sejauh mana klasifikasi mampu mengenali data dari berbagai kelas[22]. Representasi *confusion Matrix* dapat disajikan seperti yang ditampilkan pada Tabel 1.

Tabel 1. *Confusion Matrix*

Kelas	Positif	Negatif
Positif	*True Positive* (TP)	*False Positive* (FP)
Negatif	*False Negative* (FN)	*True Negative* (TN)

TP (*True Positive*) adalah data positif yang diklasifikasikan tepat oleh model, sedangkan FP (*False Positive*) adalah data negatif yang diklasifikasikan salah menjadi positif. FN (*False Negative*) merupakan data positif yang diklasifikasikan salah menjadi negatif, dan TN (*True Negative*) adalah data negatif yang diklasifikasikan tepat oleh model.

3. HASIL DAN PEMBAHASAN

3.1 Pengumpulan Data

Tahapan pengumpulan data dilakukan menggunakan TikTok *Comments Scraper* yang tersedia di situs web *apify*. Dari proses tersebut, berhasil dikumpulkan sebanyak 1.332 komentar TikTok yang diambil pada tanggal 22 April 2025. Sample dataset ditampilkan pada Gambar 2.

	text	uid	uniqueId
2	Air putih dingin ga ada tandingannya	6881458569117172737	sugamistress
3	mayan ladang cuan baru	6949781291798332418	mbahnaup
	eaakkk..		
4	gula di pajak.	6699056639109415937	kenzieal24
5	setuju, terus apakah gula akan naik juga?	95172554422	njay_7
6	sebentar lagi di indo ada pajak nafas	6910824860521268226	belomakan1
	trus gimana nasib orang² yang jualan es² di jalan, seperti pop ice,		
7	es campur, masa iya harus pakai cukai jga	7024929264240935962	cowo_amatir
8	Sri Mulyani udah pusing cari sumber APBN	6958333501235561477	temannmu__
9	bagus lha ngurangin diabetes	6811257599834096642	errizkyp
10	teh manis di warung2 gmn nasibnya....	7228921901576537090	arifj548
11	tutuo pabrik gula ganti gula yg baik ..	6943829487721563141	riadjuariah355
12	sumpah enakan minum air putih seger kali coyyy	7115845800786641946	yandimaulan7
	daripada beli minuman manis, mending liat senyum dia aja udh		
13	manis	6990310381135168538	vellkfd

Gambar 2 Sempel Dataset

Berdasarkan Gambar 2, dataset yang berhasil dikumpulkan terdiri dari tiga atribut. Namun, dalam penelitian ini hanya kolom *text* yang digunakan karena memuat isi komentar masyarakat terkait sentimen terhadap penerapan cukai pada minuman berpemanis dalam kemasan, kolom lainnya tidak dilibatkan dan data yang mengandung nilai kosong (*NaN*) akan dihapus pada tahap praproses guna menjaga kualitas data dan meningkatkan performa model.

3.2 Pra-Pemrosesan

Di tahap ini, data yang menjadi fokus dalam penelitian diproses terlebih dahulu melalui pra-pemrosesan *text* dengan beberapa tahapan seperti data *cleansing*, normalisasi kata, *tokenisasi*, *case folding*, *stemming*, dan *stopword*. yang bertujuan untuk mengubah teks menjadi bagian-bagian yang lebih terstruktur dan siap dianalisis.

a. Data *Cleansing*
Pada tahapan ini, data dari kolom *text* dihilangkan unsur-unsur yang tidak memiliki keterkaitan langsung untuk analisis, seperti hashtag, *URL*, tag *HTML*, emoticon dan karakter selain huruf. Hasil data *cleansing* ditunjukkan pada tabel 2.

Tabel 2. Hasil Data *Cleansing*

Sebelum Data *Cleansing*	Setelah Data *Cleansing*
Indonesia bnyk minum manis trus angka pengidap diabetes jg bnyak bahkan anak2 jdi perusahaan minuman untung, rumah sakit jg untung	Indonesia bnyk minum manis trus angka pengidap diabetes jg bnyak bahkan anak jdi perusahaan minuman untung rumah sakit jg untung

b. *Case folding*
Dilanjut dengan tahapan *case folding* untuk merubah semua huruf kedalam huruf kecil. Hasil *case folding* Dapat diamati pada Tabel 3.

Building of Informatics, Technology and Science (BITS)
Volume 7, No 1, June 2025 Page: 814−822
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i1.7411

Tabel 3. Hasil *Case Folding*

Sebelum *Case Folding*	Setelah *Case Folding*
Indonesia bnyk minum manis trus angka pengidap diabetes jg bnyak bahkan anak jdi perusahaan minuman untung rumah sakit jg untung	indonesia bnyk minum manis trus angka pengidap diabetes jg bnyak bahkan anak jdi perusahaan minuman untung rumah sakit jg untung

c. Normalisasi Kata
Setalah data *text* diseragamkan menjadi *lower text* dilanjut dengan normalisasi kata yang bertujuan untuk merubah kata-kata tidak baku, typo, atau singkatan ke bentuk yang baku, Dapat diamati pada tabel 4.

Tabel 4. Hasil Normalisasi Kata

Sebelum Normalisasi Kata	Setelah Normalisasi Kata
indonesia bnyk minum manis trus angka pengidap diabetes jg bnyak bahkan anak jdi perusahaan minuman untung rumah sakit jg untung	indonesia banyak minum manis terus angka pengidap diabetes juga banyak bahkan anak jadi perusahaan minuman untung rumah sakit juga untung

d. *Tokenizing*
Langkah berikutnya dalam pemrosesan adalah membagi teks ke dalam bagian-bagian kecil yang disebut *token*, biasanya berupa kata atau frasa. Hasil dari proses *tokenisasi* ini digunakan sebagai dasar untuk analisis lanjutan. Dapat diamati pada tabel 5.

Tabel 5. Hasil *Tokenizing*

Sebelum *Tokenizing*	Setelah *Tokenizing*
indonesia banyak minum manis terus angka pengidap diabetes juga banyak bahkan anak jadi perusahaan minuman untung rumah sakit juga untung	['indonesia', 'banyak', 'minum', 'manis', 'terus', 'angka', 'pengidap', 'diabetes', 'juga', 'banyak', 'bahkan', 'anak', 'jadi', 'perusahaan', 'minuman', 'untung', 'rumah', 'sakit', 'juga', 'untung']

e. *Stemming*
Tahapan ini bertujuan untuk mentransformasikan kata dalam teks ke bentuk dasar (*root word*) melalui proses penghilangan imbuhan, baik berupa awalan, akhiran, maupun sisipan. Hasil dapat diamati pada tabel 6.

Tabel 6. Hasil *stemming*

Sebelum *Stemming*	Setelah *Stemming*
['indonesia', 'banyak', 'minum', 'manis', 'terus', 'angka', 'pengidap', 'diabetes', 'juga', 'banyak', 'bahkan', 'anak', 'jadi', 'perusahaan', 'minuman', 'untung', 'rumah', 'sakit', 'juga', 'untung']	['indonesia', 'banyak', 'minum', 'manis', 'terus', 'angka', 'idap', 'diabetes', 'juga', 'banyak', 'bahkan', 'anak', 'jadi', 'usaha', 'minum', 'untung', 'rumah', 'sakit', 'juga', 'untung']

f. *Remove stop word*
Tahapan ini bertujuan untuk menghilangkan Kata-kata yang bersifat umum dan tidak membawa informasi bermakna dalam konteks analisis teks, contohnya "dan", "atau", "yang", dan lain-lain. Hasil *remove stop word* dilihat pada tabel 7.

Tabel 7. Hasil *remove stop word*

Sebelum *Remove Stop Word*	Setelah *Remove Stop Word*
['indonesia', 'banyak', 'minum', 'manis', 'terus', 'angka', 'idap', 'diabetes', 'juga', 'banyak', 'bahkan', 'anak', 'jadi', 'usaha', 'minum', 'untung', 'rumah', 'sakit', 'juga', 'untung']	['indonesia', 'minum', 'manis', 'angka', 'idap', 'diabetes', 'anak', 'usaha', 'minum', 'untung', 'rumah', 'sakit', 'untung']

g. Pembobotan kata
Proses ini bertujuan untuk mentransformasikan teks menjadi format numerik melalui proses vektorisasi, sehingga dapat digunakan sebagai fitur dalam proses klasifikasi. Metode vektorisasi yang digunakan adalah *Term Frequency-Inverse Document Frequency* (*TF-IDF*), yaitu teknik statistik yang tidak hanya menghitung seberapa sering suatu kata muncul dalam sebuah dokumen, tetapi juga mempertimbangkan pentingnya kata tersebut relatif terhadap seluruh dokumen dalam dataset. Dengan demikian, TF-IDF mampu menghasilkan bobot kata yang lebih relevan dan informatif bagi model klasifikasi.

Building of Informatics, Technology and Science (BITS)
Volume 7, No 1, June 2025 Page: 814−822
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i1.7411

Tabel 8. Hasil Pembobotan Kata

Term	TF-IDF
Konsumsi	0.6318231978737082
Cukai	0.25395852229612276
Pajak	0.14033513223373723
Diabetes	0.2518185620683344

Tabel 8 menyajikan hasil vektorisasi TF-IDF pada komentar dengan indeks 0 yang sudah melalui tahap pra-pemrosesan dengan isi komentar "beda cukai pajak abang cukai pakai barang konsumsi efek buruk badan konsumsi harga tekan konsumsi". Kata konsumsi memiliki nilai TF-IDF tertinggi sebesar 0.6318, yang menandakan kata tersebut paling sering muncul dalam komentar tersebut. Disusul oleh kata "cukai", "diabetes", dan "pajak" yang juga memiliki bobot cukup tinggi.

h. *Labelling*

Setelah dataset diperoleh, data tersebut kemudian diproses untuk pelabelan sentimen kedalam dua kategori, yakni positif dan negatif, dikerjakan oleh pakar bahasa Indonesia. Hasil dapat diamati pada tabel 9.

Tabel 9. Hasil Label Pakar

No	Text	Label
1	ini bagus bg buat menekan angka diabetes	Positive
2	semua di pajakin aja nanti sampai kita buang angin aja di pajak	Negative
3	1 tahun kemudian korupsi pajak minuman manis	Negative
4	Kalo ini gua setuju sih buat ngurangin diabetes diindo	Positive
5	naikan terus biar rakyatmu semakin sengsara	Negative
6	Padahal ini udah diterapkan juga di negara tetangga yaitu Singapura tapi kok pada kontra ya? Tujuan dari penerapannya juga bagus supaya bisa menekan angka diabetes di kalangan anak-anak dan juga anak muda yang disebabkan konsumsi makanan dan minuman yang mengandung gula yang cukup tinggi. Kalau ada kebijakan atau wacana dari pemerintah dan tujuannya bagus, tolong dipahami terlebih dahulu. Jangan belum ditonton atau dibaca sampai akhir dan modal baca judul aja udah komen negatif dan menuduh pemerintah dengan pernyataan negatif.	Positeve
7	pajak terus pikirannya dah pemerintah ini, tapi ngurus pendidikan, ekonomi negara, pembangunan negara, aja kagk becus, memang benar dah bea cukai otaknya mikirin duit,maklum bnyk koruptor	Negative

3.3 *WordCloud*

Representasi dari hasil pengolahan data ditunjukkan dalam visualisasi *Word cloud* . Label sentimen positif dan negatif, divisualisasikan secara terpisah. Gambar 3 dan Gambar 4 menampilkan hasil *Word cloud* untuk sentimen positif dan sentimen negatif.

Gambar 3. Visualisasi Worldcloud Positif

Berdasarkan visualisasi *Word cloud* pada Gambar 3, bisa diketahui bahwa unsur kata yang paling banyak ditemukan pada komentar dalam sentimen positif antara lain minum, gula, tuju, ya, sehat, dan sehat. Semakin sering suatu kata muncul, maka semakin besar pula ukurannya dalam visualisasi dalam komentar yang mengandung sentimen positif.

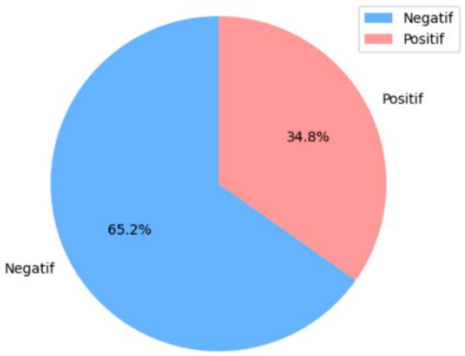
Gambar 4. Visualisasi Worldcloude Negatif

Building of Informatics, Technology and Science (BITS)
Volume 7, No 1, June 2025 Page: 814−822
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i1.7411

Berdasarkan visualisasi *Word cloud* pada Gambar 4, bisa bisa diketahui bahwa unsur kata yang paling banyak ditemukan pada komentar dalam sentimen negatif antara lain cukai, pajak, minum, gula, ya, dan harga. Semakin sering suatu kata muncul, maka semakin besar pula ukurannya dalam visualisasi dalam komentar yang mengandung sentimen negatif.

3.4 Penerapan Algoritma

Setelah selesai melalui tahap *pre-processing*, selanjutnya adalah implementasi algoritma untuk klasifikasi. Tahap ini dilakukan menggunakan visual studio code dengan memanfaatkan bahasa pemrograman *Python*, dengan menerapkan *Logistic Regression* dan *Naibe Bayes*. Data yang digunakan berjumlah 1.332 komentar yang telah melalui proses *pre-processing* dan diberi label sentimen positif atau negatif oleh ahli bahasa Indonesia. Visualisasi perbandingan antara jumlah komentar positif dan negatif dapat dilihat pada Gambar 5.



Gambar 5. Persentase Sentimen

Gambar 5 menunjukkan jumlah data berdasarkan label sentimen, dimana sentimen negatif menjadi mayoritas data dengan persentase sebesar 65,2%, sementara sentimen positif sebesar 34,8%. Setelah itu, proses persiapan data berupa pembagian dataset berlabel menjadi data pelatihan dan data pengujian dengan perbandingan 80 banding 20., atau 80% sebanyak 1066 untuk data latih dan 20% sebanyak 266 untuk testing, yang dimanfaatkan dalam proses evaluasi akurasi model klasifikasi. Tabel 10 menyajikan hasil performa berdasarkan kedua algoritma yang diterapkan pada penelitian ini.

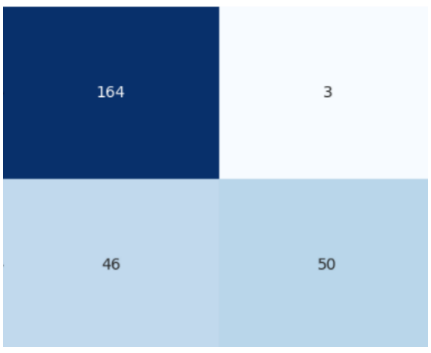
Tabel 10. Performa Algoritma

No	Algoritma	*F1-score*(%)	*Recall*(%)	*Precision*(%)	Akurasi(%)
1	*Logistic regression*	77.06%	75.14%	86.22%	81.37%
2	*Naibe Bayes*	75.76%	74.17%	82.19%	79.85%

Tabel 10 memperlihatkan performa algoritma Logistic Regression yang memperoleh nilai akurasi sebesar 81.37%, *precision* 86.22%, *recall* 75.14%, dan *F1-score* 77.06%. Sementara itu, algoritma *Naive Bayes* mendapatkan akurasi 79.85%, *precision* 82.19%, *recall* 74.17%, dan *F1-score* 75.76%. Dari hasil tersebut, *logistic Regression* terbukti memberikan kinerja yang lebih optimal dibandingkan *Naive Bayes*, terutama ditinjau dari aspek akurasi dan precision.

3.5 Evaluasi

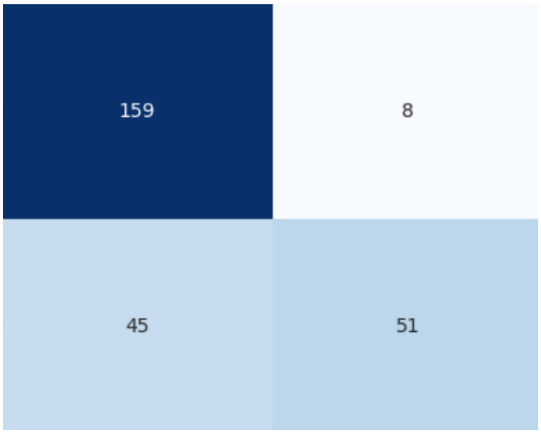
Pada proses evaluasi, *confusio n matrix* diterapkan guna melihat kinerja Setiap algoritma yang digunakan dalam mengklasifikasikan label data. Selain itu, juga dilakukan perhitungan ulang terhadap nilai akurasi yang diperoleh dari setiap algoritma



Gambar 6. *Confusion matrix Logistic Regression*

Building of Informatics, Technology and Science (BITS)
Volume 7, No 1, June 2025 Page: 814−822
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i1.7411

Berdasarkan Gambar 6, *confusion matrix* untuk algoritma *Logistic Regression* menunjukkan bahwa terdapat 164 *True Negative*, 46 *False Negative* ,50 *True Positive*, dan 3 *False Positive*. Dengan total jumlah data yang diuji adalah 263. Jika dihitung dengan rumus akurasi yaitu (TP + TN) / (TP + TN + FP + FN), maka diperoleh nilai (164 + 50) / 263 = 0.8137 atau 81.37%. Nilai ini sesuai dengan hasil akurasi dari algoritma *Logistic Regression*.



Gambar 7. *Confusion matrix Naive Bayes*

Berdasarkan Gambar 7, *confusion matrix* untuk algoritma *Naive Bayes* menunjukkan bahwa terdapat 159 *True Negative*, 45 *False Negative*, 51 *True Positive* dan 8 *False Positive*. Total data yang diuji adalah 263. Perhitungan akurasi dilakukan dengan menggunakan rumus (TP + TN) dibagi dengan jumlah total (TP + TN + FP + FN), diperoleh nilai (159 + 51) / 263 = 0.7985 atau 79.85%. Nilai ini sesuai dengan tingkat akurasi yang dihasilkan oleh algoritma *Naive Bayes*.

4. KESIMPULAN

Kesimpulan dari proses analisis sentimen terhadap kebijakan pengenaan cukai pada minuman berpemanis dalam kemasan (MBDK) berdasarkan komentar pengguna di platform TikTok menunjukkan bahwa mayoritas masyarakat memberikan respons negatif. Hal ini terlihat dari distribusi sentimen, di mana komentar negatif mencapai 65,2% atau sebanyak 869 komentar, sedangkan komentar positif hanya sebesar 34,8% atau 463 komentar dari total 1.332 data yang dianalisis. Temuan ini mengindikasikan bahwa kebijakan tersebut masih memicu pro dan kontra di kalangan masyarakat, sehingga pemerintah perlu mempertimbangkan pendekatan komunikasi publik yang terbukti lebih efisien. Selain itu, penelitian ini juga menghasilkan performa dua algoritma klasifikasi, yaitu *Logistic Regression* dan *Naive Bayes*, dalam mengolah data sentimen. Temuan dari evaluasi mengindikasikan bahwa algoritma *Logistic Regression* memberikan performa terbaik dengan tingkat akurasi sebesar 81.37%, *precision* sebesar 86.22%, *recall* sebesar 75.14%, dan *F1-score* sebesar 77.06%. Sementara itu, *Naive Bayes* memperoleh akurasi 79.85%, *precision* 82.19%, *recall* 74.17%, dan *F1-score* 75.76%. Berdasarkan temuan tersebut, dapat diambil kesimpulan bahwa *Logistic Regression* lebih optimal dalam menganalisis sentimen komentar pengguna terhadap isu cukai MBDK. Data yang telah diperoleh dapat digunakan sebagai dasar untuk pengembangan penelitian selanjutnya, dengan menambahkan jumlah data yang lebih banyak serta mempertimbangkan penggunaan metode pengolahan data yang lebih optimal guna meningkatkan kinerja dan tingkat akurasi algoritma yang diterapkan.

REFERENCES

[1] N. Rahmawati and M. Marizal, “Kebebasan Berpendapat Terhadap Pemerintah Melalui Media Sosial Dalam Perspektif Uu Ite,” Jurnal Kajian dan Penelitian Hukum, vol. 3, no. 1, 2021, doi 10.37631/widyapranata.v3i1.270

[2] D. P. R. Adawiyah, “Pengaruh Penggunaan Aplikasi TikTok Terhadap Kepercayaan Diri Remaja di Kabupaten Sampang,” *Jurnal Komunikasi*, vol. 14, no. 2, pp. 135–148, Oct. 2020, doi: 10.21107/ilkom.v14i2.7504.

[3] International Diabetes Federation, “IDF Diabetes Atlas, 11th Edition - Indonesia Data,” International Diabetes Federation. Accessed: Apr. 27, 2025. [Online]. Available: https://diabetesatlas.org/data-by-location/country/indonesia/

[4] A. N. Rahma, “Urgensi Pengenaan Cukai Pada Minuman Berpemanis Dalam Kemasan,” djpb.kemenkeu.go.id. Accessed: Nov. 28, 2024. [Online]. Available: https://djpb.kemenkeu.go.id/kanwil/sultra/id/data-publikasi/artikel/3134-urgensi-pengenaan-cukai-pada-minuman-berpemanis-dalam-kemasan.html

[5] BPKN-RI, “Penerapan Cukai Minuman Berpemanis, Kemenkeu Mundur Teratur?,” https://bpkn.go.id/. Accessed: Dec. 06, 2024. [Online]. Available: https://bpkn.go.id/beritaterkini/detail/penerapan-cukai-minuman-berpemanis-kemenkeu-mundur-teratur

[6] R. A. Pratama, “Harapan Manis, Cukai Minuman Manis,” https://mediakeuangan.kemenkeu.go.id/. Accessed: Jan. 27, 2025. [Online]. Available: https://mediakeuangan.kemenkeu.go.id/article/show/harapan-manis-cukai-minuman-manis

[7] D. Manuel, Y. Sinurat, D. E. Ratnawati, and D. W. Brata, “Analisis Sentimen Terhadap Kenaikan Cukai Rokok pada Media Sosial Twitter menggunakan Algoritma Naïve Bayes Classifier,” JPTIIK, vol. 7, no. 1, 2023

Building of Informatics, Technology and Science (BITS)
Volume 7, No 1, June 2025 Page: 814−822
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i1.7411

[8] J. S. Gea and H. Budiati, "Analisis Sentimen Masyarakat Terhadap Direktorat Jenderal Pajak," *Jurnal Sains Dan Komputer*, vol. 8, no. 01, pp. 30–36, Jan. 2024, doi: 10.61179/jurnalinfact.v8i01.466.

[9] E. M. Thoriq, D. E. Ratnawati, and B. Rahayudi, "Analisis Sentimen Opini Publik pada Media Sosial Twitter terhadap Vaksin Covid-19 menggunakan Algoritma Support Vector Machine dan Term Frequency-Inverse Document Frequency," JPTIIK, vol. 5, no. 12, 2021

[10] T. Setiawan, S. Liem, and D. M. R. Pribadi, "Perbandingan Algoritma SVM dan Naïve Bayes dalam Analisis Sentimen Komentar Tiktok pada Produk Skincare," Applied Information Technology and Computer Science, vol. 3, no. 2, 2024

[11] M. Ridwan Pratama, A. Fauzi, D. Wahiddin, and A. R. Pratama, "Analisis Sentimen Kebijakan Pembelian Gas 3 Kg dengan KTP Menggunakan Naïve Bayes" Jutisi, vol. 13, no. 2, 2024

[12] M. Khoirul, U. Hayati, and O. Nurdiawan, "Analisis Sentimen Aplikasi Brimo Pada Ulasan Pengguna Di Google Play Menggunakan Algoritma Naive Bayes," Jurnal Mahasiswa Teknik Informatika, Vol. 7, No. 1, 2023, doi 10.36040/jati.v7i1.6373

[13] F. Rizal, A. Wijaya, and F. Hasyim, "Analisis Sentimen Masyarakat Indonesia Terhadap Aplikasi TikTok Menggunakan Algoritma Logistic Regression," *AKIRATECH : Journal of Computer and Electrical Engineering*, vol. 1, no. 2, 2024, [Online]. Available: https://journal.ajbnews.com/index.php/akiratech

[14] 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, Apr. 2021, doi: 10.28932/jutisi.v7i1.3216.

[15] T. M. Fahrudin *et al.*, "Analisis Speech-to-Text pada Video Mengandung Kata Kasar dan Ujaran Kebencian dalam Ceramah Agama Islam Menggunakan Interpretasi Audiens dan Visualisasi Word Cloud," *SKANIKA: Sistem Komputer dan Teknik Informatika*, vol. 5, no. 2, pp. 190–202, 2022.

[16] K. Septiani, "Perbandingan Analisis Sentimen Terhadap Pembayaran Digital 'Go-Pay' Dan 'Ovo' Di Media Sosial Twitter Menggunakan Metode Naive Bayes Dan Word Cloud," Repository Telkom University, 2022.

[17] A. Ermillian and K. Nugroho, "Perancangan Model Deteksi Potensi Siswa Putus Sekolah Menggunakan Metode Logistic Regression Dan Decision Tree," *Jurnal Informatika: Jurnal Pengembangan IT*, vol. 9, no. 3, pp. 281–295, Dec. 2024, doi: 10.30591/jpit.v9i3.8007.

[18] R. Alwi and A. Arif Budiman, "Technology Information and Data Analytic Analisis Sentimen Kepuasan Pelanggan Parfum Scentplus dan Moris di Platform Tik Tok menggunakan Metode Regresi Logistik," *Jurnal Tifda*, vol. 1, no. 2, 2024, doi: 10.70491/tifda.v1i2.45.

[19] C. Sa, T. Widiharih, and A. Rachman Hakim, "Klasifikasi Pemberian Kredit Sepeda Motor Menggunakan Metode Regresi Logistik Biner Dan Chi-Squared Automatic Interaction Detection (Chaid) Dengan Gui R (Studi Kasus: Kredit Sepeda Motor di PT X)," *Jurnal Gaussian*, vol. 10, no. 2, pp. 159–169, 2021, https://ejournal3.undip.ac.id/index.php/gaussian/

[20] D. Putra Marbun *et al.*, "Klasifikasi Kelayakan Pinjaman Nasabah Koperasi Simpan Pinjam Menggunakan Metode Regresi Logistik Biner," *SNESTIK Seminar Nasional Teknik Elektro, Sistem Informasi, dan Teknik Informatika*, pp. 415–420, 2022, doi: 10.31284/p.snestik.2022.2810.

[21] Tania Puspa Rahayu Sanjaya, Ahmad Fauzi, and Anis Fitri Nur Masruriyah, "Analisis sentimen ulasan pada e-commerce shopee menggunakan algoritma naive bayes dan support vector machine," *INFOTECH : Jurnal Informatika & Teknologi*, vol. 4, no. 1, pp. 16–26, Jun. 2023, doi: 10.37373/infotech.v4i1.422.

[22] S. Proboningrum and A. Sidauruk, "Sistem Pendukung Keputusan Pemilihan Supplier Kain Dengan Metode Moora," *(JSiI) Jurnal Sistem Informasi*, vol. 8, no. 1, pp. 43–48, 2021.