Jurnal Ilmiah Teknologi Informasi Asia
Vol.18, No.02, Tahun 2024
ISSN: 2580-8397 (O); 0852-730X (P)

Analisis Sentimen Penggunaan Galon BPA Menggunakan Seleksi Fitur Chi-Square Dan Algoritma Support Vector Machine

Cenditya Ayu Aurelia[1], Trimono[2], I Gede Susrama Mas Diyasa[3]
[1]Universitas Pembangunan Nasional "Veteran" Jawa Timur
[2]Universitas Pembangunan Nasional "Veteran" Jawa Timur
[3]Universitas Pembangunan Nasional "Veteran" Jawa Timur
[1]20083010023@student.upnjatim.ac.id, [2]trimono.stat@upnjatim.ac.id,
[3]igsusrama.if@upnjatim.ac.id

ABSTRAK. Air Minum Dalam Kemasan (AMDK) menjadi elemen utama bagi keseimbangan tubuh. Adanya berita tentang bahaya galon yang mengandung BPA menimbulkan kekhawatiran di masyarakat terutama di platform media sosial Twitter sehingga menimbulkan keresahan masyarakat terhadap dampak negatif yang disebabkan dari penggunaan galon BPA. Hal tersebut menciptakan perdebatan antara dua pihak yang terdiri dari masyarakat yang mendukung penggunaan galon BPA dan masyarakat yang mendukung penggunaan galon non-BPA dari produk air minum tertentu. Penelitian ini melakukan analisis sentimen untuk mengelompokkan pendapat masyarakat terkait penggunaan galon menggunakan algoritma *Support Vector Machine* dan seleksi fitur *Chi-Square*. Hasil dari penelitian menunjukkan bahwa penerapan seleksi fitur *Chi-Square* meningkatkan akurasi hingga 0.95 pada *kernel Linear* dan RBF dengan 239 prediksi yang tepat dan 13 prediksi yang tidak tepat.
Kata Kunci: *Analisis Sentimen; Air Minum Dalam Kemasan; BPA; non-BPA; Support Vector Machine; Chi-Square*

ABSTRACT. Bottled Drinking Water (AMDK) is the main element for the balance of the body. The news about the dangers of gallons containing BPA raises concerns in the community, especially on the social media platform Twitter, causing public unrest about the adverse impacts resulting from the utilization of BPA-containing gallons. This creates a debate between two parties consisting of people who support the use of BPA gallons and people who support the use of non-BPA gallons from certain drinking water products. This research conducts sentiment analysis to categorize public opinions concerning the utilization of gallons using Support Vector Machine algorithm and Chi-Square feature selection. The results of the study show that the application of Chi-Square feature selection increases accuracy to 0.95 on Linear and RBF kernels with 239 correct predictions and 13 incorrect predictions.
Keywords: *Sentiment Analysis; Bottled Drinking Water, BPA; non-BPA; Support Vector Machine; Chi-Square*

1. PENDAHULUAN

Air Minum Dalam Kemasan (AMDK) menjadi bagian elemen utama yang penting dalam pemenuhan kebutuhan sumber air untuk menjaga keseimbangan tubuh (Kementerian Kesehatan RI, 2022). Pada kemasan air minum memiliki peranan yang penting untuk menjaga kualitas dan keamanan air. Jenis bahan kemasan yang umum digunakan dalam memproduksi air minum galon adalah *Polycarbonate* dan *Polietilen Tereftalat* (Santyningtya et al., 2023). Proses pembuatan kemasan berbahan *Polycarbonate* melibatkan bahan campuran yang mengandung senyawa kimia utama yang dikenal sebagai BPA atau Bisphenol A (Zulfa & Mulyawati, 2023). Dampak yang disebabkan oleh penggunaan galon mengandung BPA dapat menyebabkan penurunan fungsi sistem hormon tubuh dan berbagai masalah kesehatan lainnya seperti hipertensi, diabetes, obesitas, kanker, perkembangan kesehatan mental, serta penyakit ginjal (Aulia & Mita, 2023).

Sementara itu, kemasan galon lain yang berbahan *Polietilen Tereftalat* memiliki karakteristik jenis plastik transparan yang ditandai dengan kode daur ulang nomor 1 yang menunjukkan bahwa proses produksi tidak menggunakan bahan kimia berbahaya seperti BPA atau BPA *free* sehingga aman untuk dikonsumsi (Kompas.com, 2023). Namun, bahan tersebut dalam galon berdampak negatif terhadap lingkungan karena membutuhkan waktu untuk dapat terurai secara alami sehingga bertentangan dengan kebijakan yang telah diatur dalam Peraturan Pemerintah Republik Indonesia Nomor 27 Tahun 2020 Tentang Pengelolaan Sampah Plastik (Negara, 2020).

Hal tersebut menimbulkan keresahan masyarakat terhadap dampak negatif yang disebabkan dari penggunaan galon BPA. Masyarakat menanggapi kondisi tersebut dengan menyuarakan pendapat mereka melalui media sosial Twitter yang terdiri dari masyarakat yang mendukung penggunaan galon BPA dan masyarakat yang mendukung penggunaan galon non-BPA dari produk air minum tertentu (Musfiroh et al.,

2021). Maka dari itu, diperlukan analisis sentimen yang merupakan salah satu langkah dalam proses pengumpulan informasi text mining dalam klasifikasi teks (Muttaqin & Kharisudin, 2021). Tujuan dari analisis sentimen untuk membantu dalam mengelompokkan kalimat berdasarkan sentimen positif atau sentimen negatif sehingga dapat memudahkan dalam pengambilan keputusan (Rahmawati & Habibi, 2020). Salah satu algoritma klasifikasi yang dapat dimanfaatkan untuk menganalisis sentimen adalah *Support Vector Machine* (SVM).

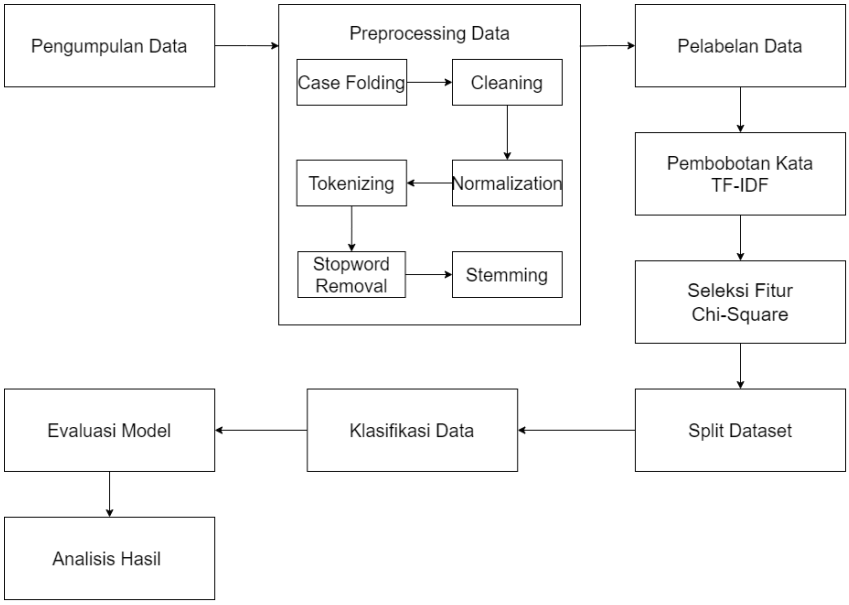
Algoritma *Support Vector Machine* dikembangkan dari prinsip teori *Structural Risk Minimization* yang memiliki tujuan untuk mendapatkan batas garis pemisah atau hyperplane yang terbaik diantara dua kelas data yang berbeda positif dan negatif sehingga dapat memaksimalkan jarak antar kelas atau margin (Agustina et al., 2020). Dalam studi kasus penelitian yang dilakukan oleh (Aziz & Purbolaksono, 2023), algoritma *Support Vector Machine* digunakan untuk menganalisis sentimen dari data ulasan film dari situs Rotten Tomatoes dengan membandingkan algoritma *Logistic Regression* dan *Naïve Bayes* menghasilkan akurasi algoritma *Support Vector Machine* tertinggi sebesar 96,4%. Penelitian tersebut menunjukkan proses klasifikasi dengan parameter *kernel Linear* dapat meningkatkan kinerja metode pengklasifikasi dengan baik dari data yang kompleks. Selain itu, penelitian terdahulu oleh (Iqbal et al., 2023) melakukan analisis sentimen pada galon BPA dengan membandingkan algoritma *Support Vector Machine* secara umum dan *Naïve Bayes* diperoleh algoritma *Support Vector Machine* akurasi tertinggi sebesar 96,15. Perbedaan penelitian sebelumnya terdapat pada algoritma *Support Vector Machine* menggunakan algoritma secara umum, sedangkan penelitian dilakukan ini memanfaatkan berbagai *kernel* dalam algoritma *Support Vector Machine* yang disesuaikan pada parameter tertentu.

Selain itu, penambahan seleksi fitur sebelum melakukan klasifikasi dapat mengoptimalkan kinerja model klasifikasi dengan meningkatkan akurasi (Abror, 2023). Seleksi fitur *Chi-Square* dapat mengurangi fitur kata yang tidak memiliki arti penting dalam klasifikasi dengan memilih subset yang optimal dari fitur asli. Salah satu penelitian oleh (Septiana et al., 2021), melakukan perbandingan seleksi fitur dari *Chi-Squared* dan *Particle Swarm Optimization* yang terbukti bahwa seleksi fitur *Chi-Square* mampu menghasilkan akurasi tertinggi sebesar 69,13%. Penelitian tersebut terbukti bahwa seleksi fitur *Chi-Square* dapat mengurangi jumlah term dalam proses klasifikasi pada teks.

Berdasarkan permasalahan diatas, maka dalam penelitian ini dilakukan untuk menganalisis sentimen penggunaan galon BPA menjadi sentimen positif dan sentimen negatif dengan penambahan seleksi fitur *Chi-Square* dan implementasi dari algoritma *Support Vector Machine*. Dalam penelitian ini bertujuan untuk meningkatkan kesadaran masyarakat Indonesia dalam memilih air minum yang berkualitas secara bijaksana sehingga dapat membuat keputusan yang tepat.

2. METODE PENELITIAN

Penelitian terkait permasalahan sentimen penggunaan galon BPA akan diselesaikan menggunakan pendekatan *Support Vector Machine* dengan penambahan seleksi fitur dari *Chi-Square*. Analisis tersebut terdiri dari beberapa tahapan untuk mempermudah penyelesaian penelitian diberikan melalui diagram alir penelitian berikut ini.



Gambar 1. Diagram Alir Penelitian

2.1. Pengumpulan Data

Data penelitian diambil dari platform media sosial Twitter untuk menganalisis sentimen terhadap penggunaan galon berbahan BPA dan non-BPA dalam kemasan air minum. Data *tweet* dikumpulkan dari tanggal 3 Februari hingga 11 November 2023 menggunakan teknik *crawling* data dengan beberapa kata kunci, antara lain "galon BPA", "galon non-BPA", "galon isi ulang", dan "galon sekali pakai", maupun komentar dalam *thread* Twitter. Proses pengumpulan data menggunakan *library* Python "*tweet*-harvest" dengan tujuan untuk mendapatkan opini masyarakat Indonesia terhadap penggunaan galon berbahan BPA dan non-BPA dalam kemasan air minum.

2.2. Preprocessing Data

Preprocessing data merupakan proses pertama dalam memproses kata tidak teratur menjadi bentuk data yang dapat digunakan pada tahap selanjutnya (Mala Olhang et al., 2020). Proses *preprocessing* data bertujuan untuk menghilangkan gangguan dan menjaga konsistensi data menjadi terstruktur (I Gede Susrama Mas Diyasa et al., 2023). Dalam penelitian ini terdapat enam tahapan *preprocessing* data meliputi:

a. *Case Folding*: mengubah huruf kapital pada huruf menjadi *lowercase* agar dapat dianalisis lebih mudah.
b. *Cleaning*: membersihkan karakter yang tidak relevan atau *noise* dari teks, seperti URL, nama akun media sosial, tagar, angka, tanda baca, emotikon, huruf tunggal, spasi berlebih, dan karakter lainnya untuk meningkatkan kualitas data.
c. *Normalization*: memperbaiki kesalahan kata menjadi bentuk kata baku berdasarkan Kamus Besar Bahasa Indonesia (KBBI) dengan daftar kamus normalisasi yang telah dibuat.
d. *Tokenizing*: proses memisahkan kalimat menjadi beberapa kata-kata yang lebih kecil agar lebih mudah dianalisis.
e. *Stopword Removal*: penghapusan kata teks yang tidak termasuk dalam kamus *stopwords* bahasa Indonesia dan daftar kata-kata tambahan yang dibuat untuk *stopword*.
f. *Stemming*: mengubah token yang terdiri dari kata imbuhan menjadi bentuk dasarnya kecuali kata yang memiliki istilah khusus yang terdapat dalam daftar kata stemming tambahan.

2.3. Pelabelan Data

Pelabelan data atau *labelling* dalam penelitian analisis sentimen merupakan proses memberikan sebuah label dengan menandai teks atau dokumen sebagai sentimen positif ataupun sentimen negatif. Dalam penelitian ini, penulis melakukan pelabelan data secara manual dari setiap tanggapan masyarakat di Twitter. Pelabelan data dilakukan setelah tahap *preprocessing data* pada kolom *normalization* agar data yang diberikan label sudah bersih dari *noise* atau gangguan dalam teks, sehingga dapat memberikan identifikasi atau klasifikasi sesuai label. Proses pelabelan data secara manual diterapkan dengan cara memberikan label kelas dari dua pihak yang berbeda untuk sentimen positif yang bernilai positif (+) dan label kelas sentimen negatif yang bernilai negatif (-). Hasil *output* dari proses pelabelan data adalah setiap data *tweets* yang sudah memiliki label atau kelas, sehingga dapat digunakan untuk proses selanjutnya.

2.4. Pembobotan Kata TF-IDF

Dalam proses pembobotan kata dengan TF-IDF terjadi transformasi data dari bentuk teks menjadi representasi numerik dengan tujuan memberikan nilai bobot pada tiap-tiap kata atau fitur (Septian et al., 2019). Rumus yang digunakan dalam perhitungan TF-IDF sebagai berikut (Alrajak et al., 2020):

$$TF = \begin{cases} 1 + log_{10}(tf_{t,d}), \text{if } tf_{t,d} > 0 \\ 0, \text{if } tf_{t,d} = 0 \end{cases} \quad (1)$$

$$idf_t = \log\left(\frac{N}{df_t}\right) \quad (2)$$

Keterangan:
TF : *Term frequency*
idf_t : Nilai IDF dari kata t
N : Total data atau dokumen
df_t : Jumlah kali *term* t muncul dalam suatu dokumen d

Pembobotan kata TF-IDF diterapkan menggunakan rumus persamaan sebagai berikut:

$$tf - idf_{td} = tf_{td} \times idf_t \quad (3)$$

Keterangan:
$tf - idf_{td}$: Bobot kata TF-IDF dalam data atau dokumen
tf_{td} : Frekuensi munculnya kata dalam dokumen
idf_t : Nilai IDF dari kata

2.5. Seleksi Fitur *Chi-Square*

Chi-Square merupakan metode pemilihan fitur kata dengan cara menghitung tingkat ketergantungan antar fitur. Dengan menggunakan *Chi-Square* dapat membantu mengurangi jumlah *term* dalam proses klasifikasi pada teks dengan menggunakan rumus persamaan sebagai berikut (Kurniawan et al., 2022):

$$x^2 = \sum \frac{(O_i - E_i)^2}{E_i} \quad (4)$$

Keterangan:
O_i : Frekuensi dari hasil observasi
E_i : Frekuensi dari hasil yang diharapkan
x^2 : Nilai *Chi-Square*

2.6. Split Dataset

Pembagian dataset dalam penelitian ini menggunakan data yang dilatih dan data yang akan diujikan. Pembagian dataset yang dilakukan adalah 80% untuk data yang dilatih dan 20% untuk data yang diujikan. Dengan demikian, pembagian dataset ini diharapkan dapat memberikan hasil optimal dalam analisis sentimen penggunaan galon BPA dengan menggunakan algoritma *Support Vector Machine*.

2.7. Klasifikasi Data

Klasifikasi data bertujuan untuk memprediksikan data *tweet* yang diberikan dengan mengelompokkan menjadi sentimen positif maupun sentimen negatif. Dalam klasifikasi data menggunakan algoritma *Support Vector Machine* dengan melakukan perbandingan tiga skema *kernel* algoritma *Support Vector Machine*, yaitu *kernel Linear*, *kernel Polynomial*, dan *kernel Radial Basis Function* (RBF). Tujuan dari algoritma *Support Vector Machine* adalah mendapatkan batas garis pemisah atau hyperplane yang terbaik dalam mengklasifikasikan data termasuk dalam studi kasus analisis sentimen dengan memaksimalkan jarak antar kelas atau margin (Agustina dkk., 2020). Berikut adalah rumus yang dapat digunakan untuk memperoleh *hyperplane* optimal algoritma *Support Vector Machine* (Nuraliza dkk., 2022):

$$(w.x_i) + b = 0 \quad (5)$$

Dalam melakukan proses klasifikasi dengan dua kelas yang berbeda dapat menggunakan rumus persamaan berikut:

$$h(x) = \sum_{i=0}^{N} a_i y_i K(x, x_i) + b \quad (6)$$

Terdapat empat jenis *kernel trick* pada algoritma *Support Vector Machine* yang terdiri dari *Linear*, *Polynomial*, dan *Radial Basis Function* (RBF) dalam proses pengujian (Anggoro dan Permatasari, 2023) sebagai berikut:

a. *Kernel Linear*
Kernel liner digunakan untuk menemukan garis lurus atau *hyperplane* dengan *margin* terbesar diantara kelas dan mengurangi kesalahan klasifikasi data melalui parameter C.

$$K(x, x_k) = x_k^T.x \quad (7)$$

b. *Kernel Polynomial*
Kernel polynomial digunakan ketika data tidak dapat dipisahkan dengan garis lurus yang dapat menghasilkan batas keputusan yang tidak *Linear*. Pada *kernel* ini menggunakan parameter C dan Degree (d).

$$K(x, x_k) = ((x_k^T.x) + C)^d \quad (8)$$

c. *Kernel Radial Basis Function* (RBF)
Kernel RBF digunakan ketika data tidak berdistribusi atau terpisah secara *Linear*. Pada *kernel* ini menggunakan parameter C dan Gamma (γ).

$$K(x, x_k) = \exp(-\gamma |x - x_k|^2) \quad (9)$$

Keterangan:

w	: Parameter garis *hyperplane* yang dicari	$K(x, x_k)$	: Fungsi *kernel*
x	: Data yang akan diklasifikasikan	a_i	: nilai alfa ke-i
x_i	: Titik data	C	: Parameter Cost
b	: Nilai bias atau konstanta	d	: Parameter Degree
y_i	: Label kelas dari data ke-I	γ	: Parameter Gamma
$x_k^T.x_k$	: Hasil perkalian transpose vektor fitur x_k		

2.8. Evaluasi Model

Penelitian ini melakukan evaluasi model dengan memanfaatkan *confusion matrix* dengan memiliki tujuan mengevaluasi kinerja metode klasifikasi baik secara aktual maupun prediktif (Musfiroh dkk., 2021). *Confusion matrix* terdiri dari nilai *accuracy*, nilai *precision*, nilai *recall*, dan nilai *f1-score* yang dapat membantu memastikan kemampuan model atau algoritma dapat memprediksikan sentimen yang tepat dari teks berdasarkan data yang dianalisis. Rumus persamaan dari *confusion matrix* adalah sebagai berikut:

a. Accuracy

$$\frac{TP}{TP+TN+FP+FN} \quad (10)$$

b. Precision

$$\frac{TP}{TP+FP} \quad (11)$$

c. Recall

$$\frac{TP}{TP+FN} \quad (12)$$

d. F1-score

$$\frac{2\times Precision\times Recall}{Precision+Recall} \quad (13)$$

3. HASIL DAN PEMBAHASAN

3.1. Pengumpulan Data

Data untuk penelitian diambil dari platform media sosial Twitter dengan menerapkan teknik *crawling* data. Pengumpulan data dilakukan dari tanggal 3 Februari hingga 11 November 2023 dengan menggunakan beberapa *keyword* atau kata kunci yang terdiri dari "galon BPA", "galon non-BPA", "galon isi ulang", dan "galon sekali pakai", maupun komentar dalam *thread*. Hasil dari proses *crawling* data keseluruhan diperoleh sebanyak 4564 baris data yang menampilkan 12 kolom. Akan tetapi, kolom yang digunakan dalam penelitian hanya menggunakan kolom full_text dan dilakukan *filtering* data atau penyaringan data untuk menghapus data *tweet* yang tidak relevan dengan penelitian sehingga diperoleh 1257 data yang siap digunakan untuk analisis lebih lanjut. Hasil dari pengumpulan data terlampir pada Tabel 1.

Tabel 1. Hasil Pengumpulan Data

full_text
@Pai_C1 Galon guna ulang udah dipake dari lama dan blm ada tuh keluhan kesehatan karena kemasan BPA.
Salah satu alesan knpa aku masih bertahan pake galon isi ulang tuh krna kandungan BPA nya masih aman banget dan lagi kita membantu mengurangi sampah plastik https://t.co/xJBi1ZgB9V
Akusih seneng banget pake galon isi ulang karna selain kandungan bpa nya msh terbilang aman, kita juga bisa membantu mengurangi pencemaran lingkungan dari smpah plastik https://t.co/wYyfrtLky1

3.2. Preprocessing Data

Proses preprocessing data terdiri dari beberapa tahapan dengan tujuan utama dari preprocessing data adalah membersihkan dan menyiapkan data sebelum dilakukan analisis berikutnya. Hasil dari tahapan preprocessing data terlampir dalam Tabel 2.

Tabel 2. Hasil Proses Preprocessing Data

full_text	case folding	cleaning	normalization	tokenizing	stopword removal	stemming
@FilkarZakaria @ririn_said Emang parah sih, danone aqua tidak merawat kebersihan galon dan tidak menjaga keamanan air galon nya dari BPA	@filkarzakaria @ririn_said emang parah sih, danone aqua tidak merawat kebersihan galon dan tidak menjaga keamanan air galon nya dari bpa	emang parah sih danone aqua tidak merawat kebersihan galon dan tidak menjaga keamanan air galon nya dari bpa	memang parah danone aqua tidak merawat kebersihan galon dan tidak menjaga keamanan air galon dari bpa	['memang', 'parah', 'danone', 'aqua', 'tidak', 'merawat', 'kebersihan', 'galon', 'dan', 'tidak', 'menjaga', 'keamanan', 'air', 'galon', 'dari', 'bpa']	['parah', 'danone', 'aqua', 'merawat', 'kebersihan', 'galon', 'menjaga', 'keamanan', 'air', 'galon', 'bpa']	['parah', 'danone', 'aqua', 'rawat', 'bersih', 'galon', 'jaga', 'aman', 'air', 'galon', 'bpa']

3.3. Pelabelan Data

Pelabelan data dilakukan secara manual pada kolom *normalization* untuk menentukan sentimen positif dan sentimen negatif berdasarkan perdebatan masyarakat terhadap penggunaan galon yang mengandung BPA dan non-BPA. Dalam proses pelabelan data ini, sentimen positif diberikan nilai 1 untuk komentar yang mendukung penggunaan galon non-BPA dengan alasan bahwa galon berbahan *Polycarbonate* memiliki kandungan BPA yang berbahaya dan tidak aman dikonsumsi dalam jangka waktu yang panjang. Sentimen negatif diberikan nilai 0 untuk komentar yang tidak mendukung penggunaan galon non-BPA dengan alasan bahwa penggunaan galon *Polycarbonate* tidak berbahaya dan mengikuti ketentuan yang berlaku dari Badan Pengawas Obat dan Makanan (BPOM). Hasil dari tahapan pelabelan data terlampir dalam Tabel 3.

Tabel 3. Hasil Pelabelan Data Sentimen

full_Text	sentiment_label	label_score
@FilkarZakaria @ririn_said Emang parah sih, danone aqua tidak merawat kebersihan galon dan tidak menjaga keamanan air galon nya dari BPA	Positif	1
@Pai_C1 Galon guna ulang udah dipake dari lama dan blm ada tuh keluhan kesehatan karena kemasan BPA.	Negatif	0

Hasil pelabelan data menunjukkan bahwa terdapat 807 data berlabel positif dan 450 data berlabel negatif dari total 1257 data.

3.4. Pembobotan Kata TF-IDF

Pembobotan kata dalam penelitian ini memanfaatkan *Term Frequency* dan *Inverse Document Frequency* dengan tujuan tiap-tiap kata akan diberikan nilai pembobotan yang terlampir dalam Tabel 4.

Tabel 4. Hasil Proses Pembobotan Kata TF-IDF

	aman	bpa	galon	isi	kemasan	pakai	ulang
0	0.0000	0.1187	0.1170	0.0000	0.2857	0.2511	0.1489
1	0.1434	0.0700	0.0690	0.0929	0.0000	0.1480	0.0877
…	…	…	…	…	…	…	…
1255	0.0000	0.0000	0.1948	0.0000	0.0000	0.0000	0.2478
1256	0.1053	0.0000	0.0507	0.0000	0.0000	0.1087	0.0645

Pada Tabel 4 menunjukkan beberapa nilai bobot dari 1050 fitur yang dihasilkan untuk setiap kata dalam tahap pembobotan kata TF-IDF. Bobot tersebut menunjukkan tingkat pentingnya setiap kata dalam setiap dokumen. Kata "galon" memiliki bobot 0.1948 pada dokumen 1256 menunjukkan bahwa kata tersebut memiliki tingkat kepentingan yang tinggi dalam dokumen tersebut. Selain itu, kata "bpa" memiliki bobot 0.1187 pada dokumen pertama menunjukkan tingkat kepentingan kata yang lebih tinggi.

3.5. Seleksi Fitur *Chi-Square*

Pada tahap seleksi fitur *Chi-Square*, fitur kata keseluruhan telah diperoleh dari proses pembobotan kata TF-IDF akan diseleksi dan diurutkan berdasarkan nilai *Chi-Square* tertinggi hingga terendah dengan tujuan memperoleh fitur kata yang relevan terhadap penelitian. Dalam proses pemilihan fitur kata ini mencari fitur kata berdasarkan kondisi apabila nilai p-values yang dihasilkan lebih kecil dari nilai taraf nyata α, maka fitur kata tersebut dipilih begitupun sebaliknya. Hasil dari seleksi fitur *Chi-Square* terlampir dalam Tabel 5.

Tabel 5. Hasil Seleksi Fitur *Chi-Square*

Nilai Taraf Nyata	Nilai Kritis	Jumlah Fitur Terpilih	Jumlah Fitur Terbuang
0.95	0.004	1033	17
0.90	0.016	1004	46
0.80	0.064	953	97
0.75	0.102	932	118

Hasil seleksi fitur *Chi-Square* pada Tabel 5 menunjukkan bahwa penggunaan nilai taraf nyata berpengaruh pada jumlah fitur yang terpilih dan fitur yang terbuang. Hasil seleksi fitur *Chi-Square* menunjukkan bahwa terdapat tiga nilai taraf nyata, yaitu 0.95, 0.80, dan 0.70 dengan nilai kritis yang sesuai. Dari hasil ini menunjukkan bahwa pada taraf nyata 0.95 terdapat 1033 fitur yang terpilih dengan 17 fitur yang terbuang. Pada taraf nyata 0.90 menghasilkan 1004 fitur dipilih dengan 46 fitur. Kemudian, pada taraf nyata 0.80 menghasilkan 953 fitur dipilih dengan 97 fitur yang terbuang, sementara pada taraf nyata 0.75 menghasilkan 932 fitur terpilih dengan 118 fitur yang terbuang. Hal ini menunjukkan nilai taraf nyata yang dipilih semakin kecil, maka fitur kata yang terbuang semakin banyak.

3.6. Klasifikasi Data

Pada klasifikasi data membandingkan tiga skema *kernel* dari algoritma *Support Vector Machine*, yaitu *kernel Linear*, *kernel Polynomial*, dan *kernel* RBF dengan memanfaatkan parameter dari C = [0.01, 0.1, 0.5, 1, 10, 20, 30, 40, 50, 100], degree = [1, 2, 3, 4, 5], Gamma = [0.001, 0.01, 0.1, 1, 10]. Proses klasifikasi melibatkan data yang dilatih 80% dari 1005 data dan data yang diujikan 20% dari 252 data dengan menggunakan hasil fitur kata yang telah terseleksi melalui metode *Chi-Square* serta data asli yang telah dibobotkan menggunakan metode TF-IDF.

Tabel 6. Klasifikasi Data SVM Tanpa Seleksi Fitur *Chi-Square*

Kernel	Parameter			Akurasi
	C	Gamma	Degree	
Linear	0.5	-	-	0.94
Polynomial	10	-	2	0.90
RBF	10	1	-	0.94

Hasil yang diperoleh dari klasifikasi data *Support Vector Machine* tanpa *Chi-Square* pada Tabel 6 menunjukkan bahwa ketiga jenis *kernel* yang diuji, yaitu *kernel Linear*, *Polynomial*, dan RBF dengan memiliki parameter-parameter tertentu. Pada *kernel Linear* menghasilkan akurasi sebesar 0.94 dengan parameter C = 0.5. Sedangkan *kernel Polynomial* menghasilkan akurasi sebesar 0.90 dengan parameter C = 10 dan Degree = 2. Sementara itu, *kernel* RBF menghasilkan akurasi sebesar 0.94 dengan parameter C = 10 dan Gamma = 1. Dari hasil klasifikasi data *Support Vector Machine* tanpa *Chi-Square* menunjukkan bahwa *kernel Linear* dan RBF memiliki akurasi tertinggi, sementara *kernel Polynomial* menunjukkan akurasi yang lebih rendah.

Tabel 7. Klasifikasi Data SVM Dengan Seleksi Fitur *Chi-Square*

Nilai Taraf Nyata	Nilai Kritis	Kernel	Parameter			Akurasi
			C	Gamma	Degree	
0.95	0.004	*Linear*	1	-	-	0.95
		Polynomial	10	-	2	0.90
		RBF	10	1	-	0.94
0.90	0.016	*Linear*	1	-	-	0.95
		Polynomial	10	-	2	0.90
		RBF	10	1	-	0.95
0.80	0.064	*Linear*	1	-	-	0.95
		Polynomial	10	-	2	0.89
		RBF	10	1	-	0.95
0.75	0.102	*Linear*	1	-	-	0.94
		Polynomial	50	-	2	0.83
		RBF	10	1	-	0.93

Hasil yang diperoleh dari klasifikasi data *Support Vector Machine* dengan *Chi-Square* pada Tabel 7 menunjukkan bahwa ketiga nilai taraf nyata 0.95, 0.80, dan 0.70. Pada nilai taraf nyata 0.95 menunjukkan bahwa *kernel Linear* menghasilkan akurasi tertinggi 0.95 atau 95% dengan parameter C = 1. Kemudian, pada nilai taraf nyata 0.90 menunjukkan bahwa *kernel Linear* memiliki parameter terbaik dari C adalah 1 dan *kernel* RBF memiliki parameter C adalah 10 dan parameter dari Gamma adalah 1 memperoleh hasil akurasi tertinggi hingga 0.95. Sedangkan, nilai tarif nyata 0.80 menunjukkan bahwa *kernel Linear* dengan parameter terbaik dari C adalah 1 dan *kernel* pada RBF parameter C adalah 10 dan parameter terbaik dari Gamma adalah 1 memperoleh hasil akurasi tertinggi hingga 0.95. Sementara itu, pada nilai tarif nyata 0.75 menunjukkan bahwa pada *kernel* pada *Linear* memiliki parameter C adalah 1 menghasilkan akurasi paling tinggi sebesar 0.94. Berdasarkan parameter yang diperoleh, untuk menentukan *hyperplane* pada *kernel* diperoleh fungsi h(x) sebagai berikut:

a. *Kernel Linear*

$$h(x) = \sum_{i=0}^{N} a_i y_i K(x, x_i) - 0.2763818$$

b. *Kernel Polynomial*

$$h(x) = \sum_{i=0}^{N} a_i y_i K(x, x_i) + 0.08704533$$

c. *Kernel* RBF

$$h(x) = \sum_{i=0}^{N} a_i y_i K(x, x_i) - 0.21912869$$

Berikut merupakan hasil prediksi dengan menggunakan data *testing* sebanyak 252 data yang dihasilkan dari nilai tarif nyata 0.90:

Tabel 8. Hasil Prediksi Data *Testing*

Kernel	Jumlah Benar	Jumlah Prediksi Salah
Linear	239	13
Polynomial	227	25
RBF	239	13

Dalam hasil prediksi pada Tabel 8 menunjukkan bahwa pada *kernel Linear* dan *kernel* RBF diperoleh total prediksi yang tepat terdapat 239 dan total prediksi yang tidak tepat terdapat 13. Sementara itu, pada *kernel Polynomial* menghasilkan total prediksi yang tepat adalah 227 dan total prediksi yang tidak tepat terdapat 25. Melalui hasil prediksi ini dapat diketahui bahwa *Support Vector Machine* dengan *kernel Linear* dan *kernel* RBF memiliki kinerja yang optimal dalam memprediksi klasifikasi data *testing*, sementara *kernel Polynomial* lebih banyak melakukan kesalahan pada saat melakukan prediksi dengan menggunakan data *testing*.

3.7. Evaluasi Model

Setelah memperoleh hasil klasifikasi model *Support Vector Machine* dengan seleksi fitur *Chi-Square* dari nilai tarif nyata 0.90. Kemudian, dilakukan evaluasi model dengan *confusion matrix* yang bertujuan mengevaluasi performa model yang melibatkan *accuracy*, *precision*, *recall*, dan *f1-score* setiap kelas sentimen bernilai positif dan sentimen bernilai negatif. Hasil dari evaluasi model terlampir dalam Tabel 9.

Tabel 9. Hasil Evaluasi Model

	kernel	accuracy	precision	recall	f1-score
SVM tanpa seleksi fitur *Chi-Square*	*Linear*	0.94	0.94	0.93	0.93
	Polynomial	0.90	0.92	0.87	0.89
	RBF	0.94	0.96	0.91	0.93
SVM dengan seleksi fitur *Chi-Square* α = 0.90	*Linear*	0.95	0.95	0.94	0.94
	Polynomial	0.90	0.94	0.87	0.89
	RBF	0.95	0.96	0.93	0.94

Hasil evaluasi model pada Tabel 9 menunjukkan perbandingan kinerja antara SVM tanpa *Chi-Square* dan SVM dengan *Chi-Square* dengan nilai taraf nyata 0.90. Dari hasil evaluasi model tersebut menunjukkan bahwa penggunaan seleksi fitur telah meningkatkan akurasi model untuk setiap jenis *kernel* yang diuji. Dalam skenario pertama, SVM tanpa *Chi-Square* menunjukkan bahwa *kernel Linear* dan *kernel* RBF menghasilkan akurasi tertinggi sebesar 0.94. Namun, *kernel* RBF menunjukkan hasil terbaik dengan menghasilkan nilai rata-rata *precision* 0.96, *recall* 0.91, dan *f1-score* 0.93. Pada skenario kedua, SVM dengan seleksi fitur *Chi-Square* α = 0.90 telah meningkatkan kinerja model SVM pada *kernel Linear* dan RBF. *Kernel Linear* mencapai akurasi 0.95 dengan rata-rata nilai *precision* diperoleh hasil 0.95, *recall* diperoleh hasil 0.94, dan *f1-score* diperoleh hasil 0.94. Sementara itu, *kernel* RBF mencapai akurasi yang sama dengan rata-rata nilai *precision* lebih tinggi, yaitu 0.96, *recall* diperoleh hasil 0.93, dan *f1-score* diperoleh hasil 0.94.

4. KESIMPULAN

Berdasarkan penelitian ini, menunjukkan bahwa hasil sentimen masyarakat terhadap penggunaan galon BPA di media sosial Twitter adalah positif. Hal ini menandakan bahwa masyarakat lebih mendukung penggunaan galon non-BPA dengan alasan galon berbahan *Polycarbonate* memiliki kandungan BPA yang berbahaya dan tidak aman dikonsumsi dalam jangka waktu yang panjang. Penambahan seleksi fitur *Chi-Square* dengan α 0.90 pada algoritma *Support Vector Machine* menghasilkan akurasi tertinggi, yaitu 0.95 untuk *kernel Linear* dengan parameter C adalah 1 dan *kernel* pada RBF dengan parameter dari C adalah 10 dan parameter dari Gamma adalah 1, serta menghasilkan total prediksi yang tepat sebanyak 239 dan total prediksi yang tidak tepat sebanyak 13. Dari penemuan ini, dapat disimpulkan bahwa menyeleksi fitur kata yang tidak penting dengan metode *Chi-Square* mampu meningkatkan hasil akurasi dengan baik. Untuk penelitian selanjutnya, disarankan dapat menggunakan metode seleksi fitur lainnya.

DAFTAR RUJUKAN

Abror, D. (2023). Analisis Sentimen Review Aplikasi PeduliLindungi Menggunakan Seleksi Fitur Information Gain Berbasis SVM. *Indonesian Journal on Software Engineering (IJSE)*, *9*(1), 1–8. http://ejournal.bsi.ac.id/ejurnal/index.php/ijse

Agustina, D. A., Subanti, S., & Zukhronah, E. (2020). Implementasi Text Mining Pada Analisis Sentimen Pengguna Twitter Terhadap Marketplace di Indonesia Menggunakan Algoritma *Support Vector Machine*. *Indonesian Journal of Applied Statistics*, *3*(2), 109. https://doi.org/10.13057/ijas.v3i2.44337

Alrajak, M. S., Ernawati, I., & Nurlaili, I. (2020). Analisis sentimen serhadap pelayanan PT PLN di jakarta pada twitter dengan algoritma k- nearest neighbor (k-nn). *Seminar Nasional Mahasiswa Ilmu Komputer Dan Aplikasinya (SENAMIKA)*, *1*(2), 110–122.

Anggoro, D. A., & Permatasari, D. (2023). Performance Comparison of the *Kernel*s of *Support Vector Machine* Algorithm for Diabetes Mellitus Classification. *International Journal of Advanced Computer Science and Applications*, *14*(1), 580–585. https://doi.org/10.14569/IJACSA.2023.0140163

Aulia, G., & Mita, S. R. (2023). Review Artikel: Pengaruh Bisphenol-A (BPA) dalam Kemasan Pangan Terhadap Kesehatan. *Farmaka*, *21*(1), 43–49.

Aziz, M. M., & Purbolaksono, M. D. (2023). *Method comparison of Naïve Bayes , Logistic Regression , and SVM for Analyzing Movie Reviews*. *4*(4), 1714–1720. https://doi.org/10.47065/bits.v4i4.2644

I Gede Susrama Mas Diyasa, Ikbar Athallah Taufik, Dimas Dzaky Daniswara, & Ahmad Adiib Aminullah. (2023). Implementation Of Natural Language Processing for Spam Email Detection in Outcome Based Education (OBE) Application. *IJEBD (International Journal of Entrepreneurship and Business Development)*, *6*(6), 1166–1171. https://doi.org/10.29138/ijebd.v6i6.2587

Kementerian Kesehatan RI. (2022). *Direktorat Jenderal Pelayanan Kesehatan*. Kementerian Kesehatan RI. https://yankes.kemkes.go.id/view_artikel/2838/pentingnya-air-dan-status-hidrasi-untuk-kesehatan-jantung

Kompas.com. (2023). *Terjamin BPA Free, Galon Le Minerale Aman untuk Anak, Ibu Hamil, dan Keluarga*. https://lifestyle.kompas.com/read/2023/06/15/123000020/terjamin-bpa-*free*-galon-le-minerale-aman-untuk-anak-ibu-hamil-dan-keluarga

Kurniawan, D., Yasir, M., & Venna, F. C. (2022). Optimization of Sentiment Analysis using Naive Bayes with Features Selection *Chi-Square* and Information Gain for *Accuracy* Improvement. *International Conference on Electrical Engineering, Computer Science and Informatics (EECSI)*, 153–160. https://doi.org/10.23919/EECSI56542.2022.9946510

Mala Olhang, M. M., Achmadi, S., & Wibisono, F. . A. (2020). Analisis Sentimen Pengguna Twitter Terhadap Covid-19 Di Indonesia Menggunakan Metode Naive Bayes Classifier (Nbc). *JATI (Jurnal Mahasiswa Teknik Informatika)*, *4*(2), 214–221. https://doi.org/10.36040/jati.v4i2.2695

Musfiroh, D., Khaira, U., Utomo, P. E. P., & Suratno, T. (2021). Analisis Sentimen terhadap Perkuliahan Daring di Indonesia dari Twitter Dataset Menggunakan InSet Lexicon. *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, *1*(1), 24–33. https://doi.org/10.57152/malcom.v1i1.20

Muttaqin, M. N., & Kharisudin, I. (2021). Analisis Sentimen Pada Ulasan Aplikasi Gojek Menggunakan Metode *Support Vector Machine* dan K Nearest Neighbor. *UNNES Journal of Mathematics*, *10*(2), 22–27. http://journal.unnes.ac.id/sju/index.php/ujm

Negara, S. (2020). Peraturan Pemerintah Republik Indonesia Nomor 27 Tahun 2020 Tentang Pengelolaan Sampah Spesifik. *Peraturan Pemerintah*, *4*(039247), 39247–39267.

Nuraliza, H., Pratiwi, O. N., & Hamami, F. (2022). Analisis Sentimen IMBd Film Review Dataset Menggunakan *Support Vector Machine* (SVM) dan Seleksi Feature Importance. *Jurnal Mirai Manajemen*, *7*(1), 1–17.

Rahmawati, S., & Habibi, M. (2020). Public Sentiments Analysis about Indonesian Social Insurance Administration Organization on Twitter. *IJID (International Journal on Informatics for Development)*, *9*(2), 87–93. https://doi.org/10.14421/ijid.2020.09205

Santyningtya, A. C., Wahjuni, E., & Fajri, F. B. (2023). Legal Protection For Refillable Gallon Consumers Due To Bisphenol A (BPA) Content. *MIDA: Majalah Ilmiah Dinamika Administrasi*, *20*(April), 285–299. https://e-journal.unwiku.ac.id/isip/index.php/DA/article/download/109/84

Septian, J. A., Fachrudin, T. M., & Nugroho, A. (2019). Analisis Sentimen Pengguna Twitter Terhadap Polemik Persepakbolaan Indonesia Menggunakan Pembobotan TF-IDF dan K-Nearest Neighbor. *Journal of Intelligent System and Computation*, *1*(1), 43–49. https://doi.org/10.52985/insyst.v1i1.36

Septiana, R. D., Susanto, A. B., & Tukiyat, T. (2021). Analisis Sentimen Vaksinasi Covid-19 Pada Twitter Menggunakan Naive Bayes Classifier Dengan Feature Selection *Chi-Square*d Statistic dan Particle Swarm Optimization. *Jurnal SISKOM-KB (Sistem Komputer Dan Kecerdasan Buatan)*, *5*(1), 49–56. https://doi.org/10.47970/siskom-kb.v5i1.228

Zulfa, N., & Mulyawati, I. (2023). Higiene Sanitasi dan Uji Pemeriksaan Mikrobiologi Depot Air Minum Isi Ulang. *HIGEIA (Journal of Public Health Research and Development)*, *7*(1), 44–54. https://journal.unnes.ac.id/sju/index.php/higeia/article/view/61441