1215 ANALISIS SENTIMEN MASYARAKAT TERHADAP PROGRAM KARTU PRAKERJA MENGGUNAKAN METODE KNN

By Cecep M Zakariya

WORD COUNT 3748
TIME SUBMITTED 21-JUN-2024 01:00PM
PAPER ID 109764436

Template ini berlaku untuk terbitan Volume 6 Nomor 1 April 2023

ANALISIS SENTIMEN MASYARAKAT TERHADAP PROGRAM KARTU PRAKERJA MENGGUNAKAN METODE KNN

Cecep M Zakariya

Abstract

Unemployment rate is a serious issue in a country's economy. According to data released by the Central Statistics Agency, the Open Unemployment Rate (TPT) in February 2022 reached 5.83%. The government-launched Kartu Prakerja program aimed at reducing unemployment has sparked considerable debate among the public. Twitter, as a social media platform, allows users to express their views on various issues, including the Prakerja program. Sentiment analysis can be used to identify whether public opinions are positive, negative, or neutral. This study aims to analyze public sentiment regarding the Kartu Prakerja program using the K-Nearest Neighbors (KNN) algorithm, building on previous research that showed the Naive Bayes Classifier method achieved good accuracy with a G-Mean of 80.1% and AUC of 81.2%. Testing results using KNN and Information Gain feature selection with parameters K=1, and a 70% training data and 30% testing data split, showed an accuracy of 93% and a G-Mean of 94%.

Keywords : *KNN, Sentiment Analysis, Twitter, Kartu Prakerja, Information Gain*

Abstrak

Tingkat pengangguran adalah masalah serius dalam ekonomi suatu negara. Menurut data yang dirilis oleh Badan Pusat Statistik, Tingkat Pengangguran Terbuka (TPT) pada bulan Februari 2022 mencapai 5,83%.Program Kartu Prakerja yang diluncurkan pemerintah untuk mengurangi pengangguran telah menimbulkan banyak perdebatan di kalangan masyarakat. Twitter sebagai platform media sosial, memungkinkan pengguna untuk menyampaikan pandangan mereka tentang berbagai isu, termasuk program prakerja. Analisis sentimen dapat digunakan untuk mengidentifikasi apakah pendapat masyarakat bersifat positif, negatif, dan netral. Penelitian ini memiliki untuk menganalisis sentimen masyarakat mengenai program Kartu Prakerja menggunakan algoritma K-Nearest Neighbors (KNN), dengan mengacu pada penelitian sebelumnya yang menunjukkan bahwa metode Naive Bayes Classifier memberikan hasil akurasi yang baik dengan G-Mean sebesar 80,1% dan AUC 81,2%. Hasil pengujian menggunakan KNN dan seleksi fitur Information Gain dengan parameter K=1, serta rasio data latih 70% dan data uji 30%, menunjukkan akurasi sebesar 93% dan G-Mean sebesar 94%.

Kata kunci : *KNN, Analisis Sentimen, Twitter, Kartu Prakerja, Information Gain*

1. PENDAHULUAN

Salah satu masalah makroekonomi yang menghambat kemajuan suatu negara adalah tingkat pengangguran, yang dapat digunakan sebagai tolak ukur kondisi perekonomian negara[1]. Berdasarkan informasi dari Badan Pusat Statistik, bahwa Tingkat Pengangguran Terbuka (TPT) Februari 2022 sebesar 5,83 %, terjadi penurunan sebesar 0,43 % dibandingkan Februari 2021[2]. Satu program yang direncanakan Pemerintah memulai implementasi program Kartu Prakerja setelah Presiden Joko Widodo mengusulkan program tersebut selama kampanye Pilpres 2019 untuk menangani permasalahan ketenagakerjaan serta untuk meningkatkan Sumber Daya Manusia. Program ini kemudian direncanakan dan dilaksanakan oleh Presiden Joko Widodo setelah terpilih. Pada bulan Februari 2020, Perpres No. 36 Tahun 2020 secara resmi menetapkan dasar hukum untuk pelaksanaan Program Kartu Prakerja dengan tujuan meningkatkan kemampuan kerja masyarakat[3].

Mereka yang sedang mencari pekerjaan, dan individu yang membutuhkan peningkatan

keterampilan dapat memanfaatkan program prakerja, Program ini juga diperuntukkan bagi pelaku umkm. Walaupun program ini merupakan salah satu inisiatif kerja pemerintah yang beroperasi sejak April 2020, tetap ada beberapa masalah yang terkait dengannya, sama seperti pembayaran insentif terlambat, banyaknya kandidat peserta yang membuat website pendaftaran tidak dapat diakses dan salah sasaran [4].

Dari permasalah tersebut menimbulkan banyak opini pro dan kontra dari masyarakat yang mendukung, menentang, dan tidak peduli dengan inisiatif terobosan pemerintah melalui Twitter. Analisis sentimen dapat digunakan untuk mengevaluasi seberapa efektif kebijakan pemerintah untuk masyarakat, sehingga pemerintah dapat memperbaiki kekurangan mereka, dan untuk mengetahui pendapat positif dan pendapat negatif yang terjadi di masyarakat, terutama pengguna Twitter, terhadap kebijakan kartu prakerja ini [1].

Pada penelitian terdahulu [3], yang berjudul "Analisis Sentimen Pengguna Twitter terhadap Program Kartu Prakerja di Tengah Pandemi Covid-19 Menggunakan Metode Naive Bayes Classifier" mendapatkan hasil bahwa sentimen tentang program kartu prakerja sebagian besar berkategori negatif. Sentimen dengan aspek negatif mencerminkan bahwa sejumlah orang mengalami kesulitan saat mendaftar, sementara sentimen dengan kategori positif menunjukkan bahwa program ini memberikan bantuan yang signifikan bagi banyak individu. Hasil klasifikasi yang diperoleh dengan metode naive bayes classifier. pada data latih menunjukan bahwa nilai G-mean sebesar 80,1% dan nilai AUC sebesar 81,2%. Sedangkan pada data uji menunjukan nilai G-mean sebesar 69,2% dan nilai AUC sebesar 73,4%.

Berdasarkan penelitian terdahulu, peneliti ingin melakukan analisis sentimen masyarakat tentang program kartu prakerja pada Twitter dengan metode K-Nearest Neighbors (KNN) dan seleksi fitur Information Gain.

Metode ini dipilih sebagai alternatif untuk menganalisis sentimen setelah penelitian sebelumnya menggunakan metode Naive Bayes Clasifier. Dengan menggunakan KNN dan seleksi fitur Information Gain, peneliti berharap dapat menghasilkan sentimen yang lebih akurat atau bahkan meningkatkan hasil klasifikasi dari penelitian sebelumnya.

TINJUAN LITERATUR

2.1. Analisis Sentimen

Analisis sentimen adalah langkah evaluasi dan pengukuran perasaan, pandangan, atau respon dari individu atau kelompok terhadap suatu topik atau entitas. Proses ini umumnya dilakukan dengan menggunakan teknik komputasional atau metode analisis teks untuk memutuskan apakah sentimen tersebut cenderung positif, negatif, atau netral. Analisis sentimen, yang kadang disebut sebagai opinion mining, mengindikasikan bahwa terdapat emosi yang tersirat di dalam kata-kata yang dipakai oleh individu. Saat ini, masyarakat cenderung menggunakan platform seperti media sosial dan situs web sebagai sarana untuk mengungkapkan pendapat dan perasaan mereka. Analisis sentimen memiliki kemampuan untuk dilakukan secara otomatis, yang menghemat waktu dan sumber daya. Algoritma khusus untuk analisis data telah menggabungkan banyak alat[5]. Salah satu tugas utama Analisis sentimen adalah proses mengkategorikan teks dalam dokumen atau kalimat dan kemudian menentukan apakah kalimat tersebut bersifat positif atau negatif. Selain itu, analisis perasaan dapat menunjukkan perasaan seperti marah, gembira, atau sedih. Di website, Anda dapat menemukan ulasan tentang merek, barang, atau orang dan mengetahui apakah mereka dianggap baik atau buruk[3].

2.2. *Text Mining*

Text mining adalah proses mengekstraksi informasi dari teks dengan menganalisis kecenderungan data serta melacak tren dan pola tertentu. Pada proses pemrosesan teks, terjadi pembobotan kata dengan tujuan memberikan nilai atau signifikansi pada term-term yang terdapat dalam dokumen. Penilaian yang diberikan pada setiap istilah dapat bervariasi tergantung pada teknik yang digunakan. Banyak orang berusaha untuk menggabungkan atau mengubah berbagai metode pembobotan kata guna menciptakan pendekatan yang baru dan inovatif[6].

2.3. Twitter

Twitter adalah sebuah *Platform* yang menawarkan layanan media sosial berbentuk microblogging, memungkinkan pengguna untuk membaca dan mengirim pesan singkat yang disebut tweet.[7]. Pada awal tahun 2013, Twitter sebuah platform media sosial, membolehkan pengguna untuk mengirim dan membaca pesan dengan batasan karakter maksimum sebanyak 140. Setiap hari, lebih dari 500 juta ulasan dikirimkan melalui platform ini. Kepopuleran twitter telah membuatnya digunakan untuk berbagai tujuan, seperti protes, kampanye politik, pendidikan, dan komunikasi darurat. [8]. Trending topic Twitter adalah masalah yang

sering dibahas dihitung oleh pengguna Twitter menggunakan hashtag (#)[3]. Penggunaan hashtag adalah untuk menandai topik-topik yang berkaitan atau agar orang lain dapat mencari topik yang serupa.

2.4. Kartu Prakerja

Program kartu prakerja dapat dimanfaatkan oleh individu yang sedang mencari pekerjaan atau pekerja/buruh yang mengalami pemutusan hubungan kerja, serta individu yang ingin meningkatkan keterampilan mereka. Selain itu, program ini juga ditujukan untuk pelaku bisnis dengan skala UMK[9]. Misi dari program kartu prakerja adalah meningkatkan kompetensi karyawan, produktivitas, dan keterampilan wirausaha[3]. Individu yang merupakan Warga Negara Indonesia (WNI) berusia 18 tahun ke atas dan tidak sedang aktif dalam proses pendidikan. resmi dapat menggunakan program kartu prakerja. Peserta program akan menerima bantuan pelatihan, dana insentif setelah pelatihan, dan akan diminta untuk berpartisipasi dalam survei evaluasi.

2.5. Preprocessing

Preprocessing adalah suatu proses untuk membersihkan data dari masalah yang bisa mengganggu hasil pemrosesan data. Tujuan dari preprocessing adalah untuk meningkatkan kualitas data yang akan digunakan[10]. Langkah-langkah dalam preprocessing data melibatkan beberapa tahap tertentu :

- Pembersihan (*Cleaning*) merupakan langkah yang dilakukan untuk mengeliminasi elemen yang tidak dibutuhkan dari data ulasan. Ini mencakup penghapusan tanda baca, numerik, url, pengguna, penanda, tanda pagar, dan retweet dengan menerapkan pola khusus, seperti ('"~&?!><#%{}([0-9]+:;')[1122].
- Case folding adalah proses mengubah semua huruf dalam teks menjadi huruf kecil atau huruf besar, tanpa memperhatikan format aslinya.
- Normalisasi melibatkan konversi dan perbaikan kata yang disingkat menjadi kata yang mengandung makna yang setara berdasarkan Kamus Besar Bahasa Indonesia, Sehingga memungkinkan untuk diproses dengan lebih mudah. Contohnya, mengubah "g" menjadi "yang", dan sejenisnya.
- Tokenisasi adalah proses memecah data teks menjadi bagian-bagian yang lebih kecil, seperti kata atau frasa. [11]. sehingga memungkinkan untuk mendapatkan nilai dari setiap kata.
- Penghapusan Stopword merupakan langkah di mana kata-kata penghubung seperti "yang," "untuk," "adalah," "dari," dan sejenisnya dihilangkan atau dibuang dari teks[12].
- Stemming melibatkan penghapusan semua afiks (prefix, suffix, dan konfix) yang terdapat dalam data tweet[13].

2.6. Pelabelan

Dalam proses pelabelan ini, data akan melalui pemrosesan otomatis yang mencakup perhitungan skor dengan menggunakan kamus lexicon, kamus tersebut termasuk kedalam metode klasifikasi dengan memakai kamus opini untuk menemukan perasaan kelas positif dan negatif dalam teks [13]. Nilai sentimen akan dikurangkan dengan nilai sentimen kategori negatif pada setiap komentar[14]. Skor yang dihasilkan akan menentukan kategori sentimen sebuah kalimat. Jika skornya lebih dari 0, kalimat tersebut dianggap memiliki sentimen positif; jika skornya kurang dari 0, kalimat tersebut dianggap memiliki sentimen negatif; dan jika skornya sama dengan 0, kalimat tersebut dianggap netral.[13].

2.7. Transformasi TF-IDF

TF-IDF merupakan teknik dalam pemrosesan teks yang digunakan untuk menilai tingkat kepentingan suatu kata dalam sebuah dokumen. TF-IDF menekankan hubungan antara kata dan dokumen dengan memberikan bobot pada frekuensi kemunculan kata dalam dokumen tersebut. Frekuensi kemunculan kata dalam sebuah dokumen disebut sebagai *Term Frequency* (TF)[15]. Inverse Document Frequency (IDF) berusaha untuk mengevaluasi kesesuaian term yang dicari dengan kata kunci yang diinginkan; term yang sering muncul akan memiliki dampak yang minim dalam menentukan keterkaitan antara kata kunci dan dokumen[16]. Nilai pembobotan dihasilkan dari perkalian dari pembobotan term frequency (TF) dan inverse document frequency (IDF).

Berikut merupakan persamaan TF yang dapat dilihat pada persamaan (1)

$$tf_{(t,d)} = \frac{f_{t,d}}{\sum_{t \in d} f_{t,d}} \qquad (1)$$

Berikut merupakan persamaan IDF yang dapat dilihat pada persamaan (2)

$$idf_{(t,D)} = log\ \frac{D}{d_j} \quad (2)$$

Berikut merupakan persamaan TF-IDF yang dapat dilihat pada persamaan (3)

$$tfidf_{(t,d,D)} = tf_{(t,d)} \times idf_{(t,D)} \quad (3)$$

2.8. Seleksi Fitur Information Gain

Pemilihan fitur digunakan untuk mengidentifikasi fitur-fitur yang paling informatif serta mengurangi dimensi ruang dokumen dengan menghapus fitur yang tidak relevan. Ini membantu meningkatkan akurasi algoritma klasifikasi dengan mempertahankan hanya fitur-fitur yang penting[17].

Seleksi fitur, juga dikenal sebagai pemilihan variabel, pemilihan atribut, atau pemilihan subset fitur, adalah proses dimana fitur-fitur yang relevan dipilih untuk menjadi target dalam pembelajaran data suatu masalah[18].

Information Gain adalah suatu metode seleksi fitur yang menilai seberapa besar dampak kehadiran atau ketiadaan suatu fitur dalam pengambilan keputusan klasifikasi yang akurat pada kelas data tertentu [19]. Tujuanya adalah untuk mengurangi kompleksitas data dengan mengurangi jumlah fitur dan meningkatkan ketepatan klasifikasi. Information Gain mengukur seberapa banyak informasi yang diberikan oleh keberadaan atau ketiadaan suatu kata, yang signifikan dalam membuat keputusan klasifikasi yang akurat pada berbagai kelas [20].

Menghitung nilai Entropy menggunakan persamaan (4)

$$Entrophy(S) = -\sum_{i=1}^{n} p_i\ log_2\ (pi) \quad (4)$$

Menghitung Information Gain dari sebuah fitur A dalam dataset S, itu menggunakan persamaan (5)

$$Gain(S,A) = Entropy(S) - \sum values(A)\ \frac{|S_v|}{|S|}\ Entropy(S_v) \quad (5)$$

2.9. K-Nearest Neighbors

Algoritma KNN adalah metode pembelajaran yang dapat digunakan untuk klasifikasi. Algoritma ini bekerja dengan cara mencari nilai K titik data terdekat dari sampel yang diberikan, kemudian menggunakan label kelas atau nilai dari titik-titik data tersebut untuk memprediksi label kelas atau nilai dari sampel yang sedang diproses. Dalam konteks klasifikasi, KNN menggunakan prosedur matematis untuk menilai kriteria-kriteria tersebut dan kemudian mengkategorikan sampel ke dalam label kelas yang paling umum di antara K tetangga terdekat.[21]. Metode ini juga memiliki kemampuan untuk mengklasifikasikan data berdasarkan data uji dan data latih, serta memungkinkan untuk Untuk menerjemahkan hasil dan akurasi prediksi, nilai k terdekat harus dipilah dengan tepat [22] Perhitungan jarak menggunakan Euclidean distance dapat dilakukan menggunakan rumus persamaan (6)

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - x_{2)}{}^2} \quad (6)$$

2.10. Confusion Matrix

Metode confusion matrix, yang digunakan untuk mengevaluasi hasil, biasanya digunakan untuk mengukur kinerja metode klasifikasi, menghitung dan menarik kesimpulan dari hasil penelitian [13]. Precision, Accuracy, Recall, dan f-measure, G-Mean yang dibu berdasarkan persamaan akan dihitung dalam confusion matrix.

TABEL 1 CONFUSION MATRIX

		Nilai Aktual	
		Potisif	Negatif
Nilai Predik	Positif	True Positif (TP)	False Positif (FP)
	Negatif	False Negatif (FN)	True Negatif (TN)

Persamaan yang digunakan untuk mengukur akurasi didefinisikan pada persamaan (7)

$$Accuracy = \frac{TP+TN}{TP+FN+FP+TN} * 100\% \quad (7)$$

Persamaan yang digunakan untuk mendapatkan nilai precision didefinisikan pada persamaan (8)

$$Precision = \frac{TP}{TP+FP} * 100\% \quad (8)$$

Persamaan yang digunakan untuk mendapatkan nilai recall didefinisikan pada persamaan (9)

$$Recall = \frac{TP}{TP+FN} * 100\% \quad (9)$$

Persamaan yang digunakan untuk mendapatkan nilai F1-score didefinisikan pada persamaan (10)

$$F1 - Score = \frac{2*\text{precision}*\text{recall}}{\text{precision}+\text{recall}} * 100\% \quad (10)$$

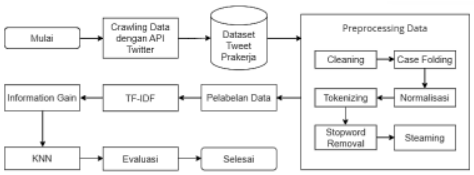
Persamaan yang digunakan untuk mendapatkan nilai G-Mean didefinisikan pada persamaan (11)

$$G - Mean = \sqrt[3]{\text{recall}_{\text{Negatif}} * \text{recall}_{\text{Netral}} * \text{recall}_{\text{Positif}}} \quad (11)$$

3. METODOLOGI PENELITIAN

3.1. Skema Alur Penelitian

Penelitian ini memiliki tahapan-tahapan yang disusun secara struktural dan divisualisasikan dalam bentuk gambar dengan urutan yang berkesinambungan. Alur penelitian tersebut dapat dilihat pada Gambar 1.



Gambar 1 Metode Penelitian

2.1. *Crawling* Data

Data yang digunakan pada penelitian ini didapatkan dari ulasan masyarakat pada sosial media Twitter atau X yang memiliki kata Kartu Prakerja. Data yang terkumpul dari bulan Januari 2023 sampai dengan bulan Desember 2023 terdiri dari 1595 data dengan menggunakan teknik pengumpulan data scrapping atau crawling.

2.2. Dataset Prakerja

Dataset *Tweet* Kartu Prakerja merupakan dataset ulasan masyarakat terhadap program kartu prakerja yang berhasil di ambil pada platform twitter, dataset ini masih kotor atau mentah. Data kotor atau data mentah adalah data yang belum diolah dan masih dalam bentuk aslinya. Berikut merupakan sampel dataset yang ditujukan pada Tabel

TABEL 2 DATASET

No.	Tanggal	Username	Teks
1	30-Des-2023	_geekydad	@hrdbacot di 2023 ini banyak roller coaster nya tapi yang paling gue banggain dari diri gue adalah bisa ikut dalam program prakerja jadi asisten tenaga pelatih. sarjana perikanan ini ngajarin gudang dan logistik

2.3. *Preprocessing*

Dataset yang diperoleh dari hasil crawling akan diproses pada tahap preprocessing, yang meliputi:

1. *Cleaning*

Cleaning dilakukan untuk penghapusan tautan, emoticon, username, tanda baca, hastag, dan karakter khusus lainya yang tidak diperlukan dalam proses pengolahan data. Berikut hasil cleaning dapat dilihat pada Tabel 3

TABEL 3 CLEANING

Full_Text	Cleaning
@hrdbacot di 2023 ini banyak roller coaster nya tapi yang paling gue banggain dari diri gue adalah bisa ikut dalam program prakerja jadi asisten tenaga pelatih. sarjana perikanan ini ngajarin gudang dan logistik	di ini banyak roller coaster nya tapi yang paling gue banggain dari diri gue adalah bisa ikut dalam program prakerja jadi asisten tenaga pelatih sarjana perikanan ini ngajarin gudang dan logistik

2. *Case Folding*

Case Folding dilakukan untuk mengkonversi semua huruf dalam *text* menjadi huruf kecil (*lowercase*). Berikut hasil *case folding* dapat dilihat pada Tabel 4

TABEL 4 CASE FOLDING

Cleaning	Case Folding

di ini banyak roller coaster nya tapi yang paling gue banggain dari diri gue adalah bisa ikut dalam program prakerja jadi asisten tenaga pelatih sarjana perikanan ini ngajarin gudang dan logistik	di ini banyak roller coaster nya tapi yang paling gue banggain dari diri gue adalah bisa ikut dalam program prakerja jadi asisten tenaga pelatih sarjana perikanan ini ngajarin gudang dan logistik

3. Normalisasi Kata

Normalisasi dilakukan untuk memperbaiki kata yang disingkat menjadi kata yang sebenarnya. Contohnya seperti kata yg akan di ubah menjadi yang. Berikut hasil Normalisasi Kata dapat dilihat pada Tabel 5

TABEL 5 NORMALISASI

Case Folding	Normalisasi Kata
di ini banyak roller coaster nya tapi yang paling gue banggain dari diri gue adalah bisa ikut dalam program prakerja jadi asisten tenaga pelatih sarjana perikanan ini ngajarin gudang dan logistik	di ini banyak roller coaster tapi yang mungkin saya banggain dari diri saya adalah bisa ikut dalam program prakerja jadi asisten tenaga pelatih sarjana perikanan ini ngajar gudang dan logistik

4. *Tokenizing*

Tokenizing dilakukan untuk membagi sebuah teks menjadi sebuah token-token kata. Berikut hasil *tokenizing* dapat dilihat pada Tabel 6

TABEL 6 TOKENIZING

Normalisasi Kata	Tokenizing
di ini banyak roller coaster tapi yang mungkin saya banggain dari diri saya adalah bisa ikut dalam program prakerja jadi asisten tenaga pelatih sarjana perikanan ini ngajar gudang dan logistik	['di', 'ini', 'banyak', 'roller', 'coaster', 'tapi', 'yang', 'mungkin', 'saya', 'banggain', 'dari', 'diri', 'saya', 'adalah', 'bisa', 'ikut', 'dalam', 'program', 'prakerja', 'jadi', 'asisten', 'tenaga', 'pelatih', 'sarjana', 'perikanan', 'ini', 'ngajar', 'gudang', 'dan', 'logistik']

5. *Stopword Removal*

Stopword Removal dilakukan untuk menghapus kata penghubung seperti di, ke, dan, dari, dsb. *Stopword* diambil berdasarkan Stopword Bahasa Indonesia yang tersedia di library Sastrawi. Berikut hasil *Stopword Removal* dapat dilihat pada Tabel 7

TABEL 7 STOPWORD REMOVAL

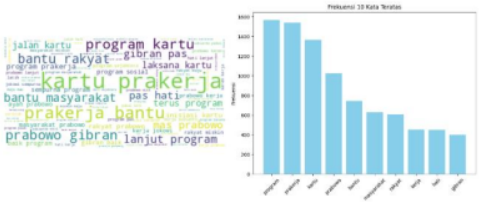
Tokenizing	Stopword Removal
['di', 'ini', 'banyak', 'roller', 'coaster', 'tapi', 'yang', 'mungkin', 'saya', 'banggain', 'dari', 'diri', 'saya', 'adalah', 'bisa', 'ikut', 'dalam', 'program', 'prakerja', 'jadi', 'asisten', 'tenaga', 'pelatih', 'sarjana', 'perikanan', 'ini', 'ngajar', 'gudang', 'dan', 'logistik']	['roller', 'coaster', 'banggain', 'program', 'prakerja', 'asisten', 'tenaga', 'pelatih', 'sarjana', 'perikanan', 'ngajar', 'gudang', 'logistik']

6. *Stemming*

Stemming dilakukan untuk mengubah kata-kata menjadi kata dasar dan mengembalikan kembali token-token kata menjadi teks yang telah di preproses. Berikut hasil *Stemming* dapat dilihat pada Tabel 8

TABEL 8 STEMMING

Stopword Removal	Stemming
['roller', 'coaster', 'banggain', 'program', 'prakerja', 'asisten', 'tenaga', 'pelatih', 'sarjana', 'perikanan', 'ngajar', 'gudang', 'logistik']	roller coaster banggain program prakerja asisten tenaga latih sarjana ikan ngajar gudang logistik



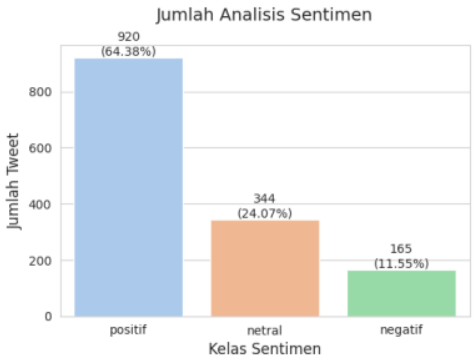
Gambar 2 Wordcloud dan Frekuensi Kata

2.4. Pelabelan Data

Pelabelan data salah satu proses yang digunakan untuk pemberian label atau sentimen pada data dengan perhitungan skor menggunakan kamus lexicon based untuk keperluan pembelajaran mesin atau analisis data.

TABEL 9 LABELING DATA

Tweet	Sentimen
roller coaster banggain program prakerja asisten tenaga latih sarjana ikan ngajar gudang logistik	Positif
...	...
pungut biaya peser syarat lolos program prakerja sedia ikut webinar terima instruksi instruktur mimin usia gitu isi data	Negatif
...	...
prabowo tegas program makan siang minum susu gratis sekolah ganti program program kartu indonesia sehat kartu indonesia sehat kartu indonesia pintar kartu indonesia pintar kartu sembako kartu prakerja program keluarga harap program keluarga harap	Netral



Gambar 3 Grafik Labeling

2.5. Transformasi TF-IDF

Teknik TF-IDF (Term Frequency-Inverse Document Frequecy) digunakan untuk mengukur atau mengevaluasi seberapa pentingnya kemunculan suatu kata dalam sebuah dokumen. Berikut adalah matriks hasil perhitungan TF-IDF yang merupakan hasil dari perkalian TF dan IDF dapat dilihat pada tabel 10

TABEL 10 TRANFORMASI TF-IDF

Fitur	TF-IDF		
	Teks 1	Teks 2	Teks 3
program	0	0	0
kartu	0	0	0
prakerja	0	0	0
skema	0.477	0	0
normal	0.477	0	0
insentif	0.477	0	0
turun	0.477	0	0
ya	0	0.477	0
daftar	0	0.176	0.176
mandiri	0	0.477	0
mudah	0	0	0.477

2.6. Information Gain

Setelah proses TF-IDF dilakukan, langkah berikutnya adalah melakukan seleksi fitur menggunakan information gain. Information Gain digunakan untuk pemilihan fitur untuk memilih subset fitur yang paling informatif dan relevan sambil mengurangi dimensi data dan kompleksitas model.

2.7. KNN

Dalam penelitian ini, metode KNN diterapkan untuk mengklasifikasikan sentimen masyarakat terhadap program Kartu Prakerja. Kinerja model dapat dievaluasi berdasarkan nilai akurasi yang dihasilkan, sehingga diperlukan pengujian untuk menilai tingkat akurasi tersebut. Dalam pengujian model K-Nearest Neighbors, akan menggunakan dua skenario yaitu pengujian model tanpa Information Gain dan pengujian model dengan menggunakan Information Gain. Pengujian dilakukan dengan dataset dengan perbandingan 80%:20%, 70%:30%, dan 60%:30% dan dengan nilai K yang di variasikan antara 1, 3, 5, dan 7.

2.8. Evaluasi

Setelah dilakukan klasifikasi menggunakan model algoritma KNN, selanjutnya menguji performa model dengan menggunakan confusion matrix untuk melihat nilai accuracy, precission, recall, F-1 Score, dan G-Mean yang dihasilkan dari model K-Nearest Neighbors.

4. HASIL DAN PEMBAHASAN

Pengujian klasifikasi sentimen masyarakat terhadap program Kartu Prakerja menggunakan metode K-Nearest Neighbors (KNN) telah dilakukan. Pengujian ini menggunakan dataset dengan perbandingan split data 80% : 20%, 70% : 30%, dan 60% : 40%, dengan variasi nilai K antara 1, 3, 5, dan 7.

a) Pengujian 1 dataset 80% : 20%

Hasil pengujian pertama dilakukan menggunakan dataset dengan perbandingan 80%:20% dan dengan variasi nilai K = 1, 3, 5, dan 7. Performa model dapat dilihat pada Tabel 11

TABEL 11 PENGUJIAN SKENARIO 1

Rasio Split Dataset 80% : 20%								
Performa	KNN				KNN + InfoGain			
	1	3	5	7	1	3	5	7
Akurasi	0.88	0.83	0.79	0.77	0.92	0.87	0.85	0.82
Presisi	0.89	0.85	0.82	0.80	0.93	0.87	0.85	0.82
Recall	0.87	0.88	0.79	0.77	0.92	0.87	0.85	0.83
F1-Score	0.87	0.82	0.78	0.76	0.92	0.87	0.85	0.82
G-Mean	0.90	0.87	0.84	0.82	0.94	0.90	0.89	0.87

Pada dataset rasio 80% : 20% menunjukan bahwa nilai akurasi menunjukan peningkatan dari sekitar 88% tanpa InfoGain menjadi 92% dengan InfoGain.

b) Pengujian 2 dataset 70% : 30%

Hasil pengujian kedua dilakukan menggunakan dataset dengan perbandingan 70%:30% dan dengan variasi nilai K = 1, 3, 5, dan 7. Performa model dapat dilihat pada Tabel 12

TABEL 12 PENGUJIAN SKENARIO 2

Rasio Split Dataset 70% : 30%								
Performa	KNN				KNN + InfoGain			
	1	3	5	7	1	3	5	7
Akurasi	0.87	0.82	0.79	0.78	0.93	0.87	0.85	0.83
Presisi	0.89	0.84	0.82	0.81	0.93	0.87	0.85	0.83
Recall	0.87	0.82	0.78	0.77	0.93	0.87	0.85	0.83
F1 – Score	0.87	0.81	0.77	0.76	0.93	0.87	0.85	0.83
G-Mean	0.90	0.86	0.84	0.83	0.94	0.84	0.88	0.87

Pada dataset rasio 70% : 30% menunjukan bahwa nilai akurasi menunjukan peningkatan dari sekitar 87% tanpa InfoGain menjadi 93% dengan InfoGain.

c) Pengujian 3 dataset 60% : 40%

Hasil pengujian kedua dilakukan menggunakan dataset dengan perbandingan 60%:40% dan dengan variasi nilai K = 1, 3, 5, dan 7. Performa model dapat dilihat pada Tabel 13

TABEL 13 PENGUJIAN SKENARIO 3

Performa	KNN				KNN + InfoGain			
	1	3	5	7	1	3	5	7
Akurasi	0.86	0.81	0.78	0.76	0.91	0.86	0.84	0.82
Presisi	0.87	0.83	0.81	0.79	0.91	0.86	0.85	0.83
Recall	0.86	0.81	0.79	0.76	0.91	0.87	0.84	0.82
F1 – Score	0.85	0.80	0.77	0.75	0.91	0.86	0.84	0.82
G-Mean	0.89	0.86	0.84	0.82	0.93	0.90	0.88	0.87

Pada dataset rasio 60% : 40% menunjukan bahwa nilai akurasi menunjukan peningkatan dari sekitar 86% tanpa InfoGain menjadi hingga 91% dengan InfoGain.

Hasil pengujian menunjukkan bahwa model KNN dengan rasio data pelatihan dan pengujian 80%:20%, 70%:30, dan 60%:40 memiliki performa terbaik pada nilai K=1. Dalam semua skenario, akurasi, presisi, recall, F1-score, dan G-Mean menurun signifikan seiring peningkatan nilai K. Penggunaan Infomation Gain sebagai metode seleksi fitur terbukti meningkatkan performa model pada semua metrik evaluasi, terutama pada nilai K yang lebih kecil. InfoGain secara konsisten meningkatkan akurasi, presisi, recall, F1-Score, dan G-Mean dalam semua rasio data.

Dari ketiga skenario pengujian yang dilakukan, model KNN + InfoGain dengan nilai K=1 secara konsisten memberikan hasil terbaik di semua metrik evaluasi. Skema dataset 70:30 dengan KNN + InfoGain pada K=1 menunjukkan sedikit keunggulan dengan nilai akurasi 93% dibandingkan skenario lainnya. Namun, secara umum, model KNN + InfoGain dengan K=1 adalah pilihan yang paling optimal untuk mencapai performa terbaik di berbagai rasio dataset yang diuji.

5. Kesimpulan dan Saran

Penelitian ini mengumpulkan 1.595 ulasan pengguna tentang program Kartu Prakerja dari platform Twitter (X) selama periode Januari 2023 hingga Desember 2023. Ulasan-ulasan ini kemudian diproses melalui beberapa tahap, yaitu preprocessing, pelabelan, transformasi TF-IDF, seleksi fitur menggunakan Information Gain, dan klasifikasi menggunakan model K-Nearest Neighbors (KNN). Evaluasi model dilakukan menggunakan confusion matrix untuk mengukur akurasi.

Hasil penelitian menunjukkan bahwa model KNN dengan seleksi fitur menggunakan Information Gain memberikan performa terbaik, dengan akurasi mencapai 93% menggunakan parameter nilai K=1 dan rasio split dataset 70% untuk data pelatihan dan 30% untuk data pengujian.

1215 ANALISIS SENTIMEN MASYARAKAT TERHADAP PROGRAM KARTU PRAKERJA MENGGUNAKAN METODE KNN

ORIGINALITY REPORT

17%
SIMILARITY INDEX

PRIMARY SOURCES

1	rathyo2smind.blogspot.com Internet	48 words — 1%
2	ojs.unida.ac.id Internet	42 words — 1%
3	ojs.unud.ac.id Internet	38 words — 1%
4	www.researchgate.net Internet	35 words — 1%
5	Syamsul Mujahidin, Muhamad Nur Hasyim, Boby Mugi Pratama. "Implementasi Analisis Sentimen Opini Publik Mengenai Sirkuit Internasional Mandalika Pada Twitter Menggunakan Metode Multinomial Naïve Bayes Classifier", Bianglala Informatika, 2022 Crossref	27 words — 1%
6	garuda.kemdikbud.go.id Internet	27 words — 1%
7	ejurnal.stmik-budidarma.ac.id Internet	26 words — 1%
8	repository.uin-suska.ac.id Internet	26 words — 1%

9	softscients.com Internet	26 words — 1%
10	123dok.com Internet	25 words — 1%
11	ejurnal.its.ac.id Internet	24 words — 1%
12	ujsds.ppj.unp.ac.id Internet	21 words — 1%
13	Rahadi Ramlan, Neva Satyahadewi, Wirda Andani. "Analisis Sentimen Pengguna Twitter Menggunakan Support Vector Machine Pada Kasus Kenaikan Harga BBM", Jambura Journal of Mathematics, 2023 Crossref	18 words — < 1%
14	repository.ub.ac.id Internet	18 words — < 1%
15	www.cnbcindonesia.com Internet	18 words — < 1%
16	core.ac.uk Internet	17 words — < 1%
17	repository.uncp.ac.id Internet	17 words — < 1%
18	kemnaker.go.id Internet	14 words — < 1%
19	www.cnnindonesia.com Internet	13 words — < 1%
	elshinta.com	

20	Internet	11 words — < 1%
21	j-ptiik.ub.ac.id Internet	11 words — < 1%
22	repository.its.ac.id Internet	11 words — < 1%
23	jabarekspres.com Internet	10 words — < 1%
24	Riska Dwi Handayani, Kusrini Kusrini, Hanif Al Fatta. "Perbandingan Fitur Ekstraksi Untuk Klasifikasi Emosi Pada Sosial Media", Jurnal Ilmiah SINUS, 2020 Crossref	9 words — < 1%
25	conference.upnvj.ac.id Internet	9 words — < 1%
26	docplayer.info Internet	9 words — < 1%
27	e-journal.stmiklombok.ac.id Internet	9 words — < 1%
28	ejournal-binainsani.ac.id Internet	9 words — < 1%
29	text-id.123dok.com Internet	9 words — < 1%
30	web-spacer.blogspot.com Internet	9 words — < 1%
31	www.idxchannel.com Internet	9 words — < 1%

32	www.scribd.com Internet	9 words — < 1%
33	Diana Nurfitriana, Taufik Ridwan, Apriade Voutama. "Analisis Opini Terhadap Aplikasi Riliv di Twitter Menggunakan Algoritma Naïve Bayes dan Random Forest", Jurnal SAINTEKOM, 2024 Crossref	8 words — < 1%
34	Iustisia Natalia Simbolon. "PREDIKSI KUALITAS AIR SUNGAI DI JAKARTA MENGGUNAKAN KNN YANG DIOPTIMALISASI DENGAN PSO", Jurnal Informatika dan Teknik Elektro Terapan, 2024 Crossref	8 words — < 1%
35	eprints.akakom.ac.id Internet	8 words — < 1%
36	eprints.upnyk.ac.id Internet	8 words — < 1%
37	ijcs.stmikindonesia.ac.id Internet	8 words — < 1%
38	ijns.org Internet	8 words — < 1%
39	lib.unnes.ac.id Internet	8 words — < 1%
40	repository.unri.ac.id Internet	8 words — < 1%
41	Hendriyo Mokodompit, Nurnaningsih Nico Abdul, Elvie Fatmah Mokodongan. "PONDOK PESANTREN MODERN DARUL MADINAH WONOSARI KABUPATEN	6 words — < 1%

BOALEMO DENGAN PENDEKATAN ARSITEKTUR TROPIS", JAMBURA Journal of Architecture, 2024
Crossref

42 Nursuci Putri Husain. "BEST FIRST FEATURE SELECTION DAN RADIAL BASIS FUNCTION UNTUK KLASIFIKASI PENYAKIT DIABETES", ILTEK : Jurnal Teknologi, 2021 — 6 words — < 1%
Crossref

43 etheses.uin-malang.ac.id — 6 words — < 1%
Internet

44 jurnal.umk.ac.id — 6 words — < 1%
Internet

EXCLUDE QUOTES	OFF	EXCLUDE SOURCES	OFF
EXCLUDE BIBLIOGRAPHY	OFF	EXCLUDE MATCHES	OFF