ANALISIS SENTIMEN TERKAIT PROGRAM KARTU PRAKERJA MENGGUNAKAN METODE *K-NEAREST NEIGHBORS*

Cecep M Zakariya[1], Yulison Herry Chrisnanto[2], Gunawan Abdillah[3]

[123,]Informatika, Fakultas Sains & Informatika, Universitas Jenderal Achmad Yani

Jln. Terusan Jend. Sudirman, Cimahi, Kota Cimahi, Jawa Barat 40525
[1] cecepmzakaria20@if.unjani.ac.id , [2]yhc@if.unjani.ac.id , [3] gunawanabdillah03@gmail.com

Abstract

One of the macroeconomic problems that hinder the development of a country is the unemployment rate. Based on data from the Central Bureau of Statistics, in February 2022 the open unemployment rate reached 5.58%. There are many opinions about the government's pre-employment card program to address unemployment. This program received pro and con responses that resulted in a variety of opinions. Twitter as a social media platform allows the delivery of opinions on various issues, including the pre-employment card program. Sentiment analysis approaches are currently widely used to assess opinions on various topics. In this study, sentiment analysis was conducted using the K-Nearest Neighbors (KNN) method and Information Gain feature selection to see the level of accuracy produced by the model and determine whether opinions about the pre-employment card program are positive, negative, or neutral. the results showed that this method succeeded in achieving an accuracy rate of 93% as well as high precision, recall, F1-Score, and G-Mean values, with a dataset ratio of 70% training data and 30% test data with a parameter value of K = 1.

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

Abstrak

Salah satu masalah makroekonomi yang menghambat perkembangan suatu negara adalah tingkat pengangguran. Berdasarkan data dari Badan Pusat Statistik, pada Februari 2022 tingkat pengangguran terbuka mencapai 5,58%. Banyak pendapat tentang program kartu prakerja pemerintah untuk mengatasi pengangguran. Program ini mendapat tanggapan pro dan kontra yang menghasilkan beragam pendapat. Twitter sebagai *platform* media sosial memungkinkan penyampaian opini tentang berbagai isu, termasuk program kartu prakerja. Pendekatan analisis sentimen saat ini banyak digunakan untuk menilai opini tentang berbagai topik. Dalam penelitian ini, dilakukan analisis sentimen menggunakan metode *K-Nearest Neighbors* (KNN) dan seleksi fitur Information Gain untuk melihat tingkat akurasi yang dihasilkan oleh model dan menentukan apakah opini tentang program kartu prakerja bersifat positif, negatif, atau netral. hasil penelitian menunjukan bahwa metode ini berhasil mencapai tingkat akurasi sebesar 93% serta nilai *precision, recall, F1-Score, dan G-Mean* yang tinggi, dengan rasio dataset 70% data latih dan 30% data uji dengan parameter nilai K=1.

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].

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 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.

2. 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. 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 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. Program ini juga ditujukan untuk pelaku bisnis dengan skala UMK[9]. Misi dari program kartu prakerja adalah meningkatkan kompetensi karyawan, produktivitas, dan keterampilan berwirausaha[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 teknik mengubah setiap 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 "yg" menjadi "yang", dan sejenisnya.
- Tokenisasi adalah proses memecah teks menjadi 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 adalah teknik pemrosesan teks yang menilai tingkat kepentingan suatu kata dalam dokumen. Teknik ini menekankan hubungan antara kata dan dokumen dengan memberikan bobot pada frekuensi kata yang muncul dalam dokumen. 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)

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

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].

Informasi Gain adalah suatu teknik untuk mengukur seberapa besar dampak kehadiran atau ketiadaan suatu fitur pada pengambilan keputusan klasifikasi yang tepat untuk 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 yang tepat tentang klasifikasi 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 menggunakan prosedur pembelajaran matematis untuk menilai kriteria sampel dan kemudian mengkategorikan sampel ke dalam kumpulan tertentu. Algoritma ini bekerja dengan menemukan nilai K dari titik data terdekat dan kemudian menggunakan label kelas atau nilai dari titik data tersebut untuk memprediksi label kelas atau nilai dari sampel yang sedang diproses[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. Conflusion 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 dibuat berdasarkan persamaan akan dihitung dalam confusion matrix.

TABEL I 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)

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

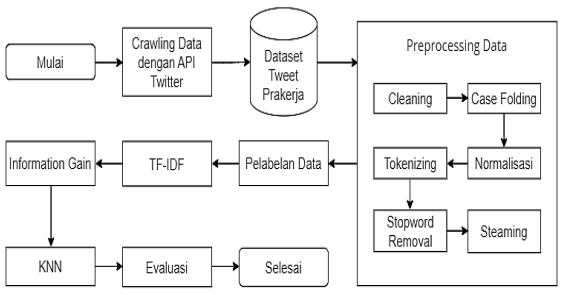
Persamaan yang digunakan untuk mendapatkan nilai G-Mean didefinisikan pada persamaan (11)

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

3. METODOLOGI PENELITIAN

3.1. Skema Alur Penelitian

Penelitian ini memiliki langkah-langkah yang disusun secara struktural dan divisualisasikan dalam urutan diagram. 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 mentah, juga disebut data kotor, adalah data yang belum diolah dan tetap dalam bentuk aslinya. Berikut merupakan sampel dataset yang ditujukan pada Tabel

TABEL II 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 III 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*

Semua huruf dalam teks diubah menjadi huruf kecil (*lowercase*) dengan metode case folding. Berikut hasil *case folding* dapat dilihat pada Tabel 4

TABEL IV 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

Kata yang disingkat dapat diubah menjadi kata yang sebenarnya melalui normalisasi. Berikut hasil Normalisasi Kata dapat dilihat pada Tabel 5

TABEL V 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 VI 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 VII STOPWORD REMOVAL

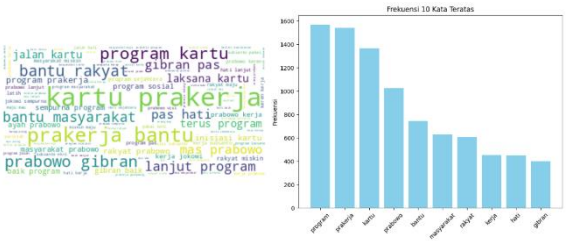
Tokenizing	Stopword Removal
['di', 'ini', 'banyak', 'roller', 'coaster', 'tapi', 'yang', 'mungkin', 'saya', 'banggain', 'dari', 'diri', 'saya', 'adalah', 'bisa', 'ikut', 'dalam', 'program', 'prakerja', 'jadi', 'asisten', 'tenaga', 'pelatih', 'sarjana', 'perikanan', 'ini', 'ngajar', 'gudang', 'dan', 'logistik']	['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 menggabungkan kembali token-token kata menjadi teks yang telah di preproses. Berikut hasil *Stemming* dapat dilihat pada Tabel 8

TABEL VII 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



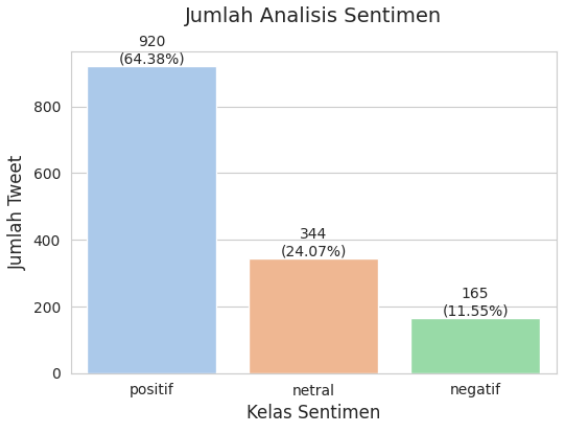
Gambar 2. Wordcloud dan Frekuensi Kata

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 IX 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 X 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, precision, recall, F-1 Score, dan G-Mean* yang dihasilkan dari algoritma *K-Nearest Neighbors* (KNN).

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

TABELXI 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 XII 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 XIII 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

5.1. Kesimpulan

Penelitian ini berhasil menganalisis sentimen yang berkaitan dengan program Kartu Prakerja dengan menggunakan algoritma *K-Nearest Neighbors* (KNN) dan Seleksi Fitur Informasi Gain.. Hasil analisis menunjukkan bahwa metode KNN mencapai tingkat keakuratan yang tinggi dalam mengklasifikasi ulasan atau tweet yang mengandung kata kunci "prakerja". Akurasi, recall, precision, dan F-1 Score mencapai 93%, sementara G-Mean mencapai 94%. Penggunaan data dengan rasio 70% untuk data latihan dan 30% untuk data uji, serta pengaturan parameter K=1, menunjukkan bahwa kombinasi KNN dan seleksi fitur Information Gain mampu menghasilkan klasifikasi yang sangat efektif dan akurat terhadap ulasan mengenai program prakerja tersebut.

5.2. Saran

Beberapa rekomendasi dapat diterapkan pada penelitian berikutnya agar hasilnya lebih baik.

1. Bisa menggunakan data dari sumber lain selain Twitter, seperti Instagram, Facebook, atau platform lainnya.
2. Bisa mencoba menggunakan metode seleksi fitur selain Information Gain untuk mengetahui seberapa akurat setiap metode.

Daftar Pustaka:

[1] W. P. Anggraini and M. S. Utami, "Klasifikasi Sentimen Masyarakat Terhadap Kebijakan Kartu Pekerja Di Indonesia," *Fakt. Exacta*, vol. 13, no. 4, p. 255, 2021, doi: 10.30998/faktorexacta.v13i4.7964.

[2] BPS.go.id, "Februari 2022: Tingkat Pengangguran Terbuka (TPT) sebesar 5,83 persen dan Rata-rata upah buruh sebesar 2,89 juta rupiah per bulan," *Badan Pus. Stat.*, no. 36, p. 1, 2022, [Online]. Available: https://www.bps.go.id/pressrelease/2022/05/09/1915/februari-2022--tingkat-pengangguran-terbuka--tpt--sebesar-5-83-persen.html

[3] ela wahyu novianti dan wahyu wibowondemi Covid-, "Analisis Sentimen Pengguna Twitter terhadap Program Kartu Prakerja di Tengah Pandemi Covid-19 Menggunakan Metode Naïve Bayes Classifier," *J. Sains dan senin ITS*, vol. 11, no. 1, pp. 136–142, 2022.

[4] R. Sanusi, F. D. Astuti, and I. Y. Buryadi, "Sentiment analysis on twitter towards pre-employment card program with recurrent neural network," *JIKO (Jurnal Inform. dan Komputer)*, vol. 5, no. 2, pp. 89–99, 2021.

[5] S. Hasanah, I. Purwasih, and ..., "Analisis Sentimen Terhadap Masyarakat Adanya Uang Kertas Baru Menggunakan Algoritma K-Nearest Neighbor (Knn)," *IKRA-ITH Inform. …*, vol. 7, no. 2, pp. 105–114, 2023, [Online]. Available: http://journals.upi-yai.ac.id/index.php/ikraith-informatika/article/view/2813%0Ahttps://journals.upi-yai.ac.id/index.php/ikraith-informatika/article/download/2813/2065

[6] A. Deolika, K. Kusrini, and E. T. Luthfi, "Analisis Pembobotan Kata Pada Klasifikasi Text Mining," *J. Teknol. Inf.*, vol. 3, no. 2, p. 179, 2019, doi: 10.36294/jurti.v3i2.1077.

[7] M. Iqbal Zakasih and W. Tri Handoko, "Analisis Sentimen Pengguna Twitter Tentang Nft (Non Fungible Token) Dengan Metode Naive Bayes Classifier," *J. Inform. dan Rekayasa Elektron.*, vol. 5, no. 2, pp. 221–229, 2022, doi: 10.36595/jire.v5i2.694.

[8] D. Duei Putri, G. F. Nama, and W. E.

Sulistiono, "Analisis Sentimen Kinerja Dewan Perwakilan Rakyat (DPR) Pada Twitter Menggunakan Metode Naive Bayes Classifier," *J. Inform. dan Tek. Elektro Terap.*, vol. 10, no. 1, pp. 34–40, 2022, doi: 10.23960/jitet.v10i1.2262.

[9] P. A. Nugroho, N. Sucahyo, and I. Kurniati, "Sentimen Analisis pada Sosial Media Twitter untuk Menilai Respon Masyarakat terhadap Seleksi Kartu Prakerja," *J. Teknlogi Inform. dan Komput. MH. Thamrin*, vol. 9, no. 1, pp. 72–83, 2023, [Online]. Available: http://journal.thamrin.ac.id/index.php/jtik/article/view/862/pdf

[10] I. Febriansyah, M. Fikry, and Yusra, "Analisis Sentiment di Twitter terhadap Anies Baswedan sebagai Bakal Calon Presiden 2024 Menggunakan Metode K-Nearest Neighbor," *G-Tech J. Teknol. Terap.*, vol. 7, no. 3, pp. 1061–1070, 2023, doi: 10.33379/gtech.v7i4.2723.

[11] A. Baita and N. Cahyono, "Analisis Sentimen Mengenai Vaksin Sinovac Menggunakan Algoritma Support Vector Machine (Svm) Dan K-Nearest Neighbor (Knn)," *Infos*, vol. 4, no. 2, pp. 42–42, 2021.

[12] S. A. Azzahra and A. Wibowo, "Analisis Sentimen Multi-Aspek Berbasis Konversi Ikon Emosi dengan Algoritme Naïve Bayes untuk Ulasan Wisata Kuliner Pada Web Tripadvisor," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 7, no. 4, p. 737, 2020, doi: 10.25126/jtiik.2020731907.

[13] M. Furqan, S. Sriani, and S. M. Sari, "Analisis Sentimen Menggunakan K-Nearest Neighbor Terhadap New Normal Masa Covid-19 Di Indonesia," *Techno.Com*, vol. 21, no. 1, pp. 51–60, 2022, doi: 10.33633/tc.v21i1.5446.

[14] R. Mahendrajaya, G. A. Buntoro, and M. B. Setyawan, "Analisis Sentimen Pengguna Gopay Menggunakan Metode Lexicon Based Dan Support Vector Machine," *Komputek*, vol. 3, no. 2, p. 52, 2019, doi: 10.24269/jkt.v3i2.270.

[15] P. D. Batlayeri and W. Gatta, "Analisis Sentimen Pejualan Jafra Dalam Pandemi Covid-19 Dengan Algoritma Klasifikasi," *J. Inform. dan Rekayasa Elektron.*, vol. 5, no. 1, pp. 11–18, 2022, doi: 10.36595/jire.v5i1.569.

[16] N. Fitriyah, B. Warsito, and D. A. I. Maruddani, "Analisis Sentimen Gojek Pada Media Sosial Twitter Dengan Klasifikasi Support Vector Machine (Svm)," *J. Gaussian*, vol. 9, no. 3, pp. 376–390, 2020, doi: 10.14710/j.gauss.v9i3.28932.

[17] F. Septianingrum, J. H. Jaman, and U. Enri, "Analisis Sentimen Pada Isu Vaksin Covid-19 di Indonesia dengan Metode Naive Bayes Classifier," *J. Media Inform. Budidarma*, vol. 5, no. 4, p. 1431, 2021, doi: 10.30865/mib.v5i4.3260.

[18] R. I. Pristiyanti, M. A. Fauzi, and L. Muflikhah, "Sentiment Analysis Peringkasan Review Film Menggunakan Metode Information Gain dan K-Nearest Neighbor," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 2, no. 3, pp. 1179–1186, 2018, [Online]. Available: http://j-ptiik.ub.ac.id

[19] H. Hidayatullah, P. Purwantoro, and Y. Umaidah, "Penerapan Naïve Bayes Dengan Optimasi Information Gain Dan Smote Untuk Analisis Sentimen Pengguna Aplikasi Chatgpt," *JATI (Jurnal Mhs. Tek. Inform.*, vol. 7, no. 3, pp. 1546–1553, 2023, doi: 10.36040/jati.v7i3.6887.

[20] A. Bijaksana, P. Negara, H. Muhardi, and I. M. Putri, "Analisis Sentimen Maskapai Penerbangan Menggunakan Metode Naive Bayes Dan Seleksi Fitur Information Gain Sentiment Analysis on Airlines Using Naïve Bayes Method and Feature Selection Information Gain," *Jtiik*, vol. 7, no. 3, pp. 599–606, 2020, doi: 10.25126/jtiik.202071947.

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

[22] M. J. Tursina, "Sentimen Analisis Sistem Zonasi Sekolah Pada Media Sosial Youtube Menggunakan Metode K-Nearest Neighbor Dengan Algoritma Levenshtein Distance," *Univ. Islam Negeri Syarif Hidayatullah Jakarta*, pp. 1–99, 2019.