Analisis Sentimen Pengguna *Twitter* terhadap Program Kartu Prakerja di Tengah Pandemi Covid-19 Menggunakan Metode *Naïve Bayes Classifier*

Ela Wahyu Novianti dan Wahyu Wibowo
Departemen Statistika Bisnis, Institut Teknologi Sepuluh Nopember (ITS)
e-mail: wahyu_w@statistika.its.ac.id

***Abstrak*—Indonesia mengkonfirmasi virus corona penyebab COVID-19 masuk pertama kali pada awal Maret 2020. Sejak itu seluruh sektor terdampak dari pandemi COVID-19 tak hanya kesehatan, sektor ekonomi juga menagalami dampak serius akibat pandemi ini. Pemerintah melakukan berbagai upaya penanggulangan salah satunya adalah dengan melakukan Pembatasann Aktivitas Berskala Besar (PSBB). Kebijakan PSBB berpengaruh pada aktivitas bisnis yang berimbas pada perekonomian sehingga berdampak pada situasi ketenagakerjaan di Indonesia. Dalam mengatasi masalah ketenagakerjaan pemerintah membuat kebijkan program Kartu Prakerja. Masalahnya muncul persepsi bahwa ditengah pandemi COVID-19 ini, logika Kartu Prakerja tidak tepat digunakan sebab tak ada jaminan bahwa pekerja yang telah dilatih mendapatkan pekerjaan baru, apalagi ditengah kondisi ekonomi yang sedang terpuruk. Akibatnya timbul pro dan kontra dari masyarakat terkait Kartu Prakerja yang sempat menjadi *trending topic* di *Twitter*. Hasil analisis sentimen program kartu prakerja kebanyakan bersifat negatif. Sentimen negatif disini menunjukkan kritik masyarakat mengenai kesulitan saat proses pendaftaran. Sentimen positif menunjukkan bahwa banyak yang mendapatkan manfaat dengan adanya program kartu prakerja. Hasil klasifikasi menggunakan metode *naïve bayes classifier* didapatkan nilai nilai *G-mean* sebesar 80,1% dan nilai AUC sebesar 81,2%. Sedangkana pada data *testing* nilai *G-mean* sebesar 69,2% dan nilai AUC sebesar 73,4%.**

Kata Kunci*—Analisis Sentimen, Kartu Prakerja, *Naïve Bayes Classifier.

I. PENDAHULUAN

INDONESIA merupakan salah satu negara yang terdampak pandemi COVID-19. Indonesia mengkonfirmasi virus corona penyebab COVID-19 masuk pertama kali pada awal Maret 2020. Pemerintah melakukan berbagai upaya untuk menanggulangi serta meredam dampak dari pandemi virus corona. Salah satu upaya pemerintah adalah dengan melakukan Pembatasan Aktivitas Berskala Besar (PSBB). Kebijakan Pembatasan Aktivitas Berskala Besar (PSBB) yang mengharuskan untuk melakukan *physical distancing* berpengaruh pada aktivitas bisnis yang berimbas pada perekonomian dan melemahnya kinerja ekonomi sehingga berdampak pada situasi ketenagakerjaan di Indonesia.

Situasi ini menjadikan jumlah pengangguran semakin meningkat akibat banyak pekerja yang dirumahkan atau bahkan diberhentikan (PHK) untuk menekan kerugian bagi pelaku bisnis karena terhambatnya aktivitas perekonomian. Pemerintah membuat kebijakan program Kartu Prakerja yang telah digagas Presiden Joko Widodo dalam debat kampanye Pilpres 2019 untuk mengatasi masalah ketenagakerjaan dengan pengembangan Sumber Daya Manusia. Presiden Joko Widodo setelah terpilih, membuat rencana program tersebut untuk merealisasikannya. Pada bulan Februari 2020, program ini resmi memiliki landasan hukum dengan disahkannnya Perpres No. 36 Tahun 2020 tentang Pengembangan Kompetensi Kerja melalui Program Kartu Prakerja.

Program Kartu Prakerja program pengembangan kompetensi kerja dan kewirausahaan yang ditujukan untuk pencari kerja, pekerja/buruh yang terkena pemutusan hubungan kerja, dan/atau pekerja/buruh yang membutuhkan peningkatan kompetensi, termasuk pelaku usaha mikro dan kecil. Realisasi program Kartu Prakerja dalam rencana awal akan dilakukan serangkaian pelatihan secara *offline* akan tetapi rencana yang telah dirumuskan berubah menjadi *online* akibat wabah pandemi COVID-19. Program Kartu Prakerja diprioritaskan untuk karyawan yang terkena PHK, pekerja informal dan pelaku UMKM yang terdampak COVID-19.

Timbul pro dan kontra dari masyarakat terkait Kartu Prakerja yang sempat menjadi *trending topic* di *Twitter* yang beranggapan bahwa program ini tidak tepat digunakan di tengan pandemi Covid-19 sebab tak ada jaminan bahwa pekerja yang telah dilatih mendapatkan pekerjaan baru ditengah kondisi ekonomi yang sedang terpuruk.. *Twitter* adalah salah satu *microblog* yang memiliki banyak pengguna di dunia.

Pengguna *twitter* di Indonesia menempati peringkat 5 terbesar di dunia dibawah USA, Brazil, Jepang, dan Inggris yaitu mencapai angka 19,5 juta pengguna twitter dari total 300 juta pengguna global. *Tweet* pada setiap pengguna *twitter* dapat berpengaruh dalam pembentukan citra suatu produk atau program, semakin banyak suatu topik tertentu diulas dalam *tweet* pengguna maka topik tersebut dapat menjadi *trending topic* di *twitter*.

Dari ulasan di atas, maka peneliti tertarik melakukan penelitian untuk mengetahui tanggapan masyarakat berdasarkan sentimen di media sosial (*twitter*) terhadap program Kartu Prakerja. Analisis sentimen digunakan untuk melihat kecenderungan suatu sentimen atau pendapat, apakah pendapat tersebut cenderung beropini positif atau negatif. *Twitter* telah menyediakan *Application Programming Interface* (API) yaitu sekumpulan fungsi atau protokol yang disediakan untuk pengguna dalam rangka mengembangkan sebuah aplikasi.

Twitter API memungkinkan pengguna untuk mengakses dan mendapatkan informasi mengenai *tweet*, profil pengguna, data *follower*, dan lainnya. Analisis sentimen atau *opinion mining* merupakan metode analisis berbasis komputasi

Tabel 1.
Confusion Matrix

Kelas Aktual	*Kelas Prediksi*	
	Positif	Negatif
Positif	*TP*	*FN*
Negatif	*FP*	*TN*

Tabel 2.
Ilustrasi *K-fold Cross Validation*

Iteration 1	*Iteration 2*	*Iteration 3*	⋯	*Iteration K*
Fold 1	*Fold* 1	*Fold* 1	⋯	*Fold* 1
Fold 2	*Fold* 2	*Fold* 2	⋯	*Fold* 2
Fold 3	*Fold* 3	*Fold* 3	⋯	*Fold* 3
⋮	⋮	⋮		⋮
Fold K	*Fold K*	*Fold K*	⋯	*Fold K*

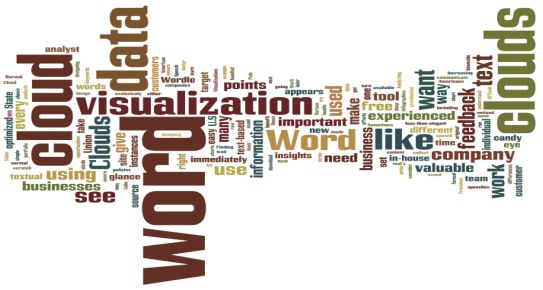
Tabel 3.
Data *Tweet* Sebelum *Preprocessing Text*

No	*Tweet*
1	*"Suka, setuju atau tidak pada krisis ekonomi dan pandemi saat ini, kebijakan pemerintah yg sangat hebat itu, menurut saya adalah penyaluran bantuan BLT, Prakerja, bantuan Covid-19 dan Bantuan UMKM"*
2	*"Dapet BLT Prakerja cair Gaji cair Wajah tersenyum dengan lingkaran cahaya #NovemberWish"*
3	*"YG DIMANJAKAN BUKAN YG DEMO DI JLN, TAPI STAFFSUS MILLENIAL YG DIGAJI PULUHAN JT/BLN TAPI KERJANYA NGGAK JELAS, YG DPT MUNTAHAN PROYEK RATUSAN MLYRD PROYEK PRAKERJA DI STARTUPNYA, YG CUMA BISA MALAKIN CAMAT2 DGN KOP SURAT ISTANA DAN BUZZERP YG KECIPRATAN DOKU 90 M LEBIH"*
4	*"sebuah kesalahan mengupload prakerja, langsung di palak orang orang"*
5	*"Di masa pandemi begini perekonomian anjlok, banyak yg kena phk. Nyari kerja juga susah, mau usaha juga susah. Pemerintah manjain anak mudanya dimana? Gimana? Ngasih bantuan prakerja juga ga terlalu efektif."*
⋮	⋮
3890	*"upload masih gagal.... akan saya coba seharian penuh... #prakerja @itsborneo @prakerjagoid"*

mengenai pendapat, sentimen, dan emosi [1]. Penelitian sebelumnya yang berkaitan adalah oleh Kurniawan (2017) tentang analisis sentimen pengguna *twitter* terhadap media *mainstream* menggunakan metode NBC dan SVM [2]. Klasifikasi menggunakan NBC pada data media TV One dan Kompas TV menghasilkan akurasi sebesar 95,6% dan 97,8%, sedangkan pada media Metro TV menghasilkan nilai *G-mean* dan AUC berturut-turut sebesar 81,3% and 82,36% [2]. Serta penelitian tentang metode panambangan aturan asosiasi untuk identifikasi penggunaan internet. Hasil dari memeriksa satu item menunjukkan bahwa jenis kelamin laki-laki dan perempuan sebagian besar menggunakan internet untuk mengakses media sosial [3].

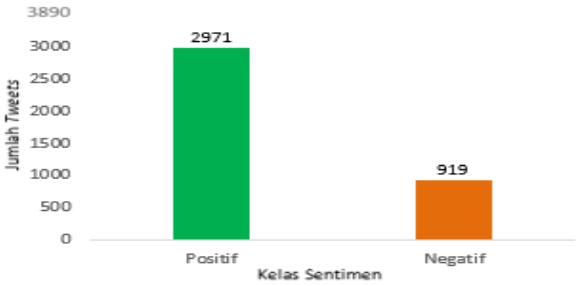
Penelitian ini bertujuan untuk melihat kecenderungan sentimen atau tanggapan masyarakat mengenai program Kartu Prakerja cenderung beropini positif atau negatif menggunakan metode *Naïve Bayes Classifier* (NBC). Metode *Naïve Bayes Classifier* merupakan salah satu metode yang dapat mengklasifikasikan teks. Kelebihan NBC adalah algoritmanya sederhana tetapi memiliki akurasi yang tinggi [4].

Penelitian ini diharapkan dapat membantu mengetahui tanggapan masyarakat mengenai program Kartu Prakerja yang diluncurkan secara objektif serta memberikan informasi mengenai kata yang kerap dibicarakan dari program Kartu Prakerja.



Gambar 1. Ilustrasi *wordcloud.*

Gambar 2. Visualisasi *wordcloud.*



Gambar 3. *Barchart* kelas sentimen.

II. TINJAUAN PUSTAKA

A. Teorema Bayes

Teorema *Bayes* merupakan teorema yang mengacu konsep probabilitas bersyarat [5]. Secara matematis teorema *Bayes* dapat dinotasikan pada persamaan (1).

$$P(A|B) = \frac{P(A)P(B|A)}{P(B)} \quad (1)$$

Keterangan:

A : Sampel data yang label kelasnya belum diketahui
B : Kelas-kelas hasil dari klasifikasi
$P(A|B)$: Probabilitas terjadinya A jika B diketahui. Disebut probabilitas *posterior*, di mana peluang A bergantung pada nilai B tertentu.
$P(B|A)$: Probabilitas terjadinya B jika A diketahui atau disebut *likelihood function.*
$P(A)$: Probabilitas *prior A* mendahului terjadinya B
$P(B)$: Probabilitas *prior B* dan bertindak sebagai *norma-lizing constant.*

B. Naïve Bayes Classifier (NBC)

Algoritma *Naïve Bayes Classifier* (NBC) merupakan algoritma yang digunakan untuk mencari nilai probabilitas tertinggi untuk mengklasifikasi data uji pada kategori yang

Tabel 4.
Data *Tweet* Sebelum *Preprocessing Text*

No	Text
1	*"suka setuju krisis ekonomi pandemi kebijakan pemerintah hebat penyaluran bantuan blt bantuan covid bantuan umkm"*
2	*"blt cair gaji cair wajah senyum lingkaran cahaya"*
3	*"manja demo staffsus millenial gaji puluhan kerja muntahan proyek proyek startup malakin camat kop surat istana buzzerp kecipratan"*
4	*"salah upload langsung palak"*
⋮	⋮
3890	*"upload gagal coba seharian penuh"*

Tebel 5.
Pelabelan Sentimen Data T*weet*

No	Text	*Score*	*Kelas*
1	*"suka setuju krisis ekonomi pandemi kebijakan pemerintah hebat penyaluran bantuan blt bantuan covid bantuan umkm"*	3	Positif
2	*"blt cair gaji cair wajah senyum lingkaran cahaya"*	1	Positif
3	*"manja demo staffsus millenial gaji puluhan kerja muntahan proyek proyek startup malakin camat kop surat istana buzzerp kecipratan"*	-2	Negatif
4	*"salah upload langsung palak"*	-1	Negatif
⋮	⋮	⋮	⋮
3890	*"upload gagal coba seharian penuh"*	-1	Negatif

Tabel 6.
Pembagian Data *Training* dan Data *Testing*

k		*Iterasi Ke -*									
		1	2	3	4	5	6	7	8	9	10
1	Positif	297	297	297	297	297	297	297	297	297	298
	Negatif	92	92	92	92	92	92	92	92	92	91
2	Positif	297	297	297	297	297	297	297	297	297	298
	Negatif	92	92	92	92	92	92	92	92	92	91
3	Positif	297	297	297	297	297	297	297	297	297	298
	Negatif	92	92	92	92	92	92	92	92	92	91
4	Positif	297	297	297	297	297	297	297	297	297	298
	Negatif	92	92	92	92	92	92	92	92	92	91
5	Positif	297	297	297	297	297	297	297	297	297	298
	Negatif	92	92	92	92	92	92	92	92	92	91
6	Positif	297	297	297	297	297	297	297	297	297	298
	Negatif	92	92	92	92	92	92	92	92	92	91
7	Positif	297	297	297	297	297	297	297	297	297	298
	Negatif	92	92	92	92	92	92	92	92	92	91
8	Positif	297	297	297	297	297	297	297	297	297	298
	Negatif	92	92	92	92	92	92	92	92	92	91
9	Positif	297	297	297	297	297	297	297	297	297	298
	Negatif	92	92	92	92	92	92	92	92	92	91
10	Positif	297	297	297	297	297	297	297	297	297	298
	Negatif	92	92	92	92	92	92	92	92	92	91

paling tepat [6]. Metode *Naive Bayes Classification* merupakan salah satu metode yang dapat mengklasifikasikan teks.

Kelebihan NBC adalah algoritmanya sederhana tetapi memiliki akurasi yang tinggi. Terdapat dua tahapan dalam klasifikasi *tweet.* Tahap pertama adalah pelatihan terhadap *tweet* yang telah diketahui kategorinya (*training*), sedangkan tahap kedua adalah proses klasifikasi *tweet* yang belum diketahui kategorinya (*testing*).

Dalam algoritma NBC setiap dokumen direpresentasikan dengan pasangan atribut "a_1, a_2,..., a_n" di mana a_1 adalah kata pertama, a_2 adalah kata kedua hingga kata ke n (a_n). Sedangkan V adalah himpunan kategori *tweet*. Pada saat klasifikasi, algoritma akan mencari probabilitas tertinggi dari semua kategori dokumen yang diujikan (V_{MAP}). Adapun persamaan V_{MAP} adalah persamaan (2).

$$V_{MAP} = \arg\max P(v_j) \prod_i P(a_i|v_j) \quad (2)$$

Nilai $P(v_j)$ dihitung pada saat *training*, didapat dengan persamaan (3):

$$P(v_j) = \frac{|doc\ j|}{|training|} \quad (3)$$

Dimana |*doc j*| merupakan jumlah *tweet* yang memiliki kategori *j* dalam *training*. Sedangkan |*training*| merupakan jumlah *tweet* yang digunakan dalam *training*. Untuk setiap probabilitas kata a_i untuk setiap kategori $P(a_i|v_j)$, dihitung pada saat *training*.

$$P(a_i|v_j) = \frac{n_i+1}{|n+kosakata|} \quad (4)$$

Dimana n_i adalah jumlah kemunculan kata a_i yang berkategori v_j dalam *tweet*, sedangkan n adalah banyaknya seluruh kata dalam *tweet* kategori v_j dan |kosakata| adalah banyaknya kata dalam data *training.*

C. Ketepatan Klasifikasi

Pengukuran ketepatan klasifikasi dilakukan untuk melihat performa klasifikasi yang telah dilakukan.

Dalam mengukur ketepatan klasifikasi, perlu diketahui jumlah pada setiap kelas prediksi dan kelas aktual yang terdiri dari:

1. *TP* (*True Positive*) yaitu jumlah *tweet* bersentimen positif yang tepat terprediksi dalam kelas positif
2. *TN (True Negative)* yaitu *tweet* bersentimen negatif yang tepat terprediksi dalam kelas negatif
3. *FP (False Positive*) yaitu *tweet* bersentimen negatif yang terprediksi dalam kelas positif

Tabel 7.
Confusion Matrix Data *Training* Iterasi 1

Kelas Aktual	Kelas Prediksi	
	Negatif	Positif
Negatif	566	155
Positif	261	2519

Tabel 8.
Ketepatan Klasifikasi NBC

Iterasi	*Training*		*Testing*	
	G-mean	AUC	*G-mean*	AUC
1	0,803	0,813	0,733	0,748
2	**0,806**	**0,816**	0,610	0,664
3	0,791	0,804	0,639	0,693
4	0,800	0,811	0,662	0,697
5	0,803	0,815	0,624	0,672
6	0,794	0,806	0,535	0,635
7	0,804	0,815	0,712	0,743
8	0,799	0,810	0,735	0,753
9	0,805	0,814	**0,890**	**0,890**
10	0,802	0,813	0,839	0,843
Rata-rata	0,801	0,812	0,698	0,734

4. *FN (False Negative)* yaitu *tweet* bersentimen positif yang terprediksi dalam kelas negatif. Tabel 1 merupakan *confusion matrix* nilai tersebut.

Pengukuran yang sering digunakan untuk menghitung ketepatan klasifikasi adalah akurasi, *specificity*, dan *sensitivity* [7]. Akurasi merupakan persentase dokumen yang teridentifikasi secara tepat dari total dokumen dalam proses klasifikasi. Akurasi digunakan untuk menghitung ketepatan klasifikasi sebuah dokumen yang mempunyai data yang *balanced* pada tiap kategorinya. Persamaan (5), persamaan (6) dan persamaan (7) berturut-turut merupakan persamaan yang digunakan untuk akurasi, *specificity* dan *sensitivity*.

$$Akurasi = \frac{TN+TP}{TN+TP+FN+FP} \quad (5)$$

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

$$Specifity = \frac{TN}{TN+FP} \quad (7)$$

Jika rasio suatu kelas data lebih banyak dibandingkan kelas lain, maka data dikatakan tidak seimbang (*imbalanced*). Kelas mayor merupakan kelas data yang lebih banyak, sedangkan kelas minor merupakan kelas data yang lebih sedikit. Jika proporsi sampel kelas minoritas kurang dari 35% dari data, maka dapat dikategorikan data tidak seimbang [5]. Pengukuran ketepatan klasifikasi untuk data *imbalanced* menggunakan *G-Mean*. *G-Mean* atau *geometric mean* merupakan rata-rata geometrik nilai *recall* dari data yang memiliki dua kategori [8]. Dalam mengukur nilai performansi klasifikasi, *G-Mean* memiliki kelebihan yaitu nilai klasifikasi yang dihasilkan *robust*. Persamaan (8) merupakan rumus untuk mendapatkan nilai *G-Mean*.

$$G - mean = \sqrt{Sensitivity \times Specifity} \quad (8)$$

Selain *G-Mean* juga digunakan nilai *Area Under Curve* (AUC), AUC merupakan indikator performansi kurva ROC (*Receiver Operating* Characteristic) yang dapat meringkas kinerja sebuah *classifier* menjadi satu nilai [9].

$$AUC = \frac{1}{2}(Sensitivity + Specificity) \quad (9)$$

D. *Sentiment Analysis*

Sentiment analysis atau *opinion mining* mengacu pada bidang yang luas dari pengolahan bahasa alami, komputasi linguistik dan *Text Mining* yang bertujuan menganalisis pendapat, sentimen, evaluasi, sikap, penilaian dan emosi seseorang pembicara atau penulis berkenaan dengan suatu topik, produk, layanan, organisasi, individu, ataupun kegiatan tertentu lainnya [1]. Tugas dasar dalam analisis sentimen adalah mengelompokkan teks yang ada dalam sebuah kalimat atau dokumen kemudian menentukan kalimat atau dokumen tersebut apakah bersifat positif atau negatif. *Sentiment analysis* juga dapat menyatakan perasaan emosional sedih, gembira, atau marah. Dapat mencari pendapat tentang produk-produk, merek atau orang-orang dan menentukan apakah mereka dilihat positif atau negatif di *website*.

E. *Text Mining*

Text Mining adalah penggalian data untuk menyelesaikan masalah kebutuhan informasi dengan menerapkan teknik *data mining*, *machine learning*, *natural language processing*, pencarian informasi, dan manajemen pengetahuan. *Text mining* melibatkan pra-proses dokumen seperti kategorisasi teks, ekstraksi informasi, dan ekstraksi kata. Metode ini digunakan untuk mengekstraksi informasi dari sumber data melalui identifikasi dan eksplorasi pola yang menarik [6]. Pada dasarnya proses kerja dari *Text Mining* banyak mengapdopsi dari penelitian *data mining,* yang menjadi perbedaan adalah pola yang digunakan oleh *Text Mining* diambil dari sekumpulan bahasa alami yang tidak terstruktur. Sedangkan dalam *data mining* pola yang diambil dari *database* yang terstruktur. Tahap-tahap *Text Mining* secara umum adalah pra-proses teks dan *featureselection*. Langkah-langkah yang dilakukan dalam *text mining* adalah sebagai berikut:

1) *Text Preprocessing*

Tindakan yang dilakukan pada tahap ini adalah *toLowerCase*, yaitu mengubah semua karakter huruf menjadi huruf kecil, dan *Tokenizing* yaitu proses penguraian deskripsi yang semula berupa kalimat-kalimat menjadi kata-kata dan menghilangkan tanda baca seperti tanda titik (.), koma (,), spasi dan karakter angka yang ada pada kata tersebut.

2) *Feature* Selection

Pada tahap ini tindakan yang dilakukan adalah menghilangkan *stoplist (stoplist removal)* dan *stemming* terhadap kata yang berimbuhan [6]. *Stoplist* adalah kosa kata yang bukan merupakan ciri (kata unik) dari suatu data. Misalnya "*di*", "*oleh*", "*pada*", "*sebuah*", "*karena*" dan lain sebagainya. *Stemming* adalah proses pemetaan dan penguraian berbagai bentuk (*variants*) dari suatu kata menjadi bentuk kata dasarnya (*stem*). Tujuan dari proses *stemming* adalah meng-hilangkan imbuhan-imbuhan baik itu berupa prefiks, sufiks, maupun konfiks yang ada pada setiap kata.

F. *Word Cloud*

Word cloud disebut juga *text cloud* atau *tag cloud* merupakan salah satu metode untuk menampilkan data teks secara visual. Grafik ini populer dalam *text mining* karena mudah dipahami. Dengan menggunakan *word cloud*, gambaran frekuensi kata-kata dapat ditampilkan dalam

bentuk yang menarik namun tetap informatif. Semakin sering satu kata digunakan, maka semakin besar pula ukuran kata tersebut ditampilkan dalam *word cloud* (Gambar 1) [10].

G. K-fold Cross Validation

K-fold cross validation adalah salah satu metode yang digunakan untuk mempartisi data menjadi data *training* dan data *testing*. Metode ini banyak digunakan peneliti karena dapat mengurangi bias yang terjadi dalam pengambilan sampel. *K-fold cross validation* secara berulang-ulang membagi data menjadi data *training* dan data *testing*, di mana setiap data mendapat kesempatan menjadi data *testing* [11]. *K* merupakan besar angka partisi data yang digunakan untuk pembagian *training testing.* Hasil beberapa penelitian terdahulu menunjukkan bahwa *10-fold cross validation* merupakan pilihan terbaik untuk mendapatkan estimasi yang akurat. Penggunaan jumlah *fold* terbaik untuk uji validitas, dianjurkan menggunakan *10-fold cross validation* dalam model [12]. Ilustrasi pembagian data menggunakan *K-fold cross validation* ditampilkan pada Tabel 2.

H. Twitter

Twitter adalah sosial media berbasis internet yang memfasilitasi pengguna untuk menulis ide, status, atau gagasan yang bisa dilihat oleh pengguna lainnya dalam bentuk teks yang dikenal dengan '*tweets*'. Berbagai topik yang dibahas di *Twitter* memberikan informasi yang sangat luas bagi para pengguna. Pengguna bisa menulis ataupun mengakses informasi-informasi yang terjadi di seluruh dunia. Para pengguna bisa memilih untuk mengikuti pengguna yang lain untuk didapatkan informasinya [2]. *Trending topic* pada *twitter* merupakan isu yang sedang banyak diulas oleh pengguna *Twitter* yang dihitung dengan penanda *hashtag* (#). *Hashtag* digunakan untuk menandai suatu topik tertentu agar dapat seragam dengan pembahasan pengguna lain atau agar dapat dicari pengguna lain yang tertarik dengan topik yang sama.

I. Kartu Prakerja

Program Kartu Prakerja adalah program pengembangan kompetensi kerja dan kewirausahaan yang ditujukan untuk pencari kerja, pekerja/buruh yang terkena pemutusan hubungan kerja, dan/atau pekerja/buruh yang membutuhkan peningkatan kompetensi, termasuk pelaku usaha mikro dan kecil. Program Kartu Prakerja bertujuan untuk mengembangkan kompetensi angkatan kerja, meningkatkan produktivitas dan daya saing angkatan kerja, serta mengembangkan kewirausahaan. Program Kartu Prakerja terbuka bagi WNI berusia 18 tahun ke atas yang tidak sedang mengikuti pendidikan formal. Peserta akan mendapatkan insentif sesuai ketentuan yang berlaku, seperti bantuan pelatihan, dana insentif pascapelatihan dan insentif saat survei evaluasi.

III. METODOLOGI PENELITIAN

A. Metode Pengumpulan Data

Sumber data yang digunakan pada penelitian ini yaitu data sekunder yang diperoleh dari *tweet* pengguna *twitter* menggunakan *Twitter* API (*Application Programming Interface*) mengenai kartu prakerja dengan kata kunci 'Prakerja'. Data diambil selama 3 bulan pada bulan November 2020 hingga Januari 2021 dan didapatkan 3.890 data *tweet*.

Struktur data yang digunakan pada penelitian ini terdiri dari variabel prediktor dan variabel respon. Variabel prediktor yaitu kata dasar dari setiap *tweet* sedangkan variabel respon yaitu klasifikasi sentimen (positif atau negatif).

B. Langkah Analisis

Langkah analisis yang dilakukan pada penelitian ini adalah sebagai berikut: (1) Menyiapkan data *tweet* yang diperoleh dari *twitter* dengan kata kunci "Prakerja". (2) Melakukan *pre processing text* yaitu; (a) Melakukan pembersihan data (*Cleaning*) yaitu meng-hapus *URL*, *mention*, dan *hashtag*. (b) Melakukan *casefolding* yaitu mengubah semua dengan huruf kecil (*non capital*). (c) Menghapus *punctuation* (tanda baca). (d) Menghapus nomor. (e) Menghapus spasi berlebih. (f) Melakukan proses *stopword removal* yaitu menghapus kata-kata tidak penting Daftar *stopwords*. (g) Melakukan proses *stemming* yaitu menghapus imbuhan sehingga kata menjadi kata dasar. (h) Menyimpan data yang telah dibersihkan. (3) Karakteristik data (Membuat *term document matrix* untuk mengubah struktur data menjadi nominal dan Visualisasi *word cloud*). (4) Melakukan pelabelan kelas pada data *tweet* yang telah bersih menjadi kelas sentimen positif dan negatif menggunakan kamus yang terdiri dari kumpulan kata positif dan negatif. (5) Membagi data *tweet* menjadi data *training* dan data *testing* menggunakan *10-fold cross validation*. (6) Mengklasifikasi data menggunakan metode NBC: (a) Menghitung probabilitas dari V_j pada data *training*, di mana V_j merupakan kategori sentimen, yaitu V_1 = positif, dan V_2 = negatif. (b) Menghitung probabilitas kata a_i pada kategori V_j (c) Membuat model probabilitas NBC disimpan dan digunakan untuk tahap data testing. (d) Menghitung probabilitas tertinggi dari kategori sentimen yang diujikan (V_{MAP}). (e) Mencari nilai V_{MAP} paling maksimum dan memasukkan *tweet* tersebut pada kategori dengan V_{MAP} maksimum. (7) Melakukan evaluasi hasil klasifikasi. (8) Melakukan intepretasi hasil analisis. (9) Menarik kesimpulan dan saran.

IV. ANALISIS DAN PEMBAHASAN

A. Preprocessing dan Karakteristik Data

Data yang telah dikumpulkan dari *tweet* pengguna *twitter* mengenai kartu prakerja yang diambil selama 3 bulan pada bulan November 2020 – bulan Januari 2021 didapatkan sebanyak 3.890 data *tweet*. Data di lakukaan *preprocessing* teks meliputi *case folding*, *stemming*, *stopword.* Berikut merupakan struktur data *tweet* sebelum dilakukan *preprocessing* (Tabel 3).

Data *tweet* yang belum dilakukan *preprocessing* masih memuat *link URL*, *mention, username,* kata-kata tidak penting (*stopwords*), dan simbol-simbol yang tidak menggambarkan isi dari tweet seperti tanda baca dan nomor sehingga perlu dilakukan *preprocessing* untuk mendapatkan data *tweet* yang menggambarkan isi *tweet*. Selain itu, dalam *preprocessing* juga dilakukan proses penghapusan kata imbuhan (*stemming*) sehingga menjadi kata dasar. *Preprocessing* data dilakukan untuk meningkatkan ketepatan

dalam klasifikasi dan mengurangi kesalahan klasifikasi data. Data yang telah dilakukan *preprocessing* ditambilkan dalam Tabel 4.

Proses selanjutnya yaitu memecah *tweet* menjadi kata per kata pada data yang telah dibersihkan dan membuat pembobotan kata untuk melihat kata yang sering muncul dalam *tweet* pengguna *twitter.* Kata yang sering digunakan muncul dalam *tweet* pengguna *twitter* divisualisasikan menggunakan *wordcloud* pada Gambar 2.

Ukuran kata dalam *wordcloud* menggambarkan frekuensi kata tersebur sering muncul dalam *tweet* pengguna *twitter*. Semakin besar ukuran kata, maka semakin tinggi frekuensi kata tersebut digunakan. Gambar 2 terlihat bahwa kata *"gelombang"* merupakan kata yang paling sering muncul diikuti dengan kata *"daftar"* kemudian kata *"bantuan"* lalu kata *"pelatihan"* dan *"kerja".*

B. Pelabelan Tweet Pengguna Twitter Mengenai Kartu Prakerja

Pelabelan pada data *tweet* yang telah dilakukan tahap *preprocessing* terdiri dari dua kelas yaitu kelas sentimen positif dan negatif. Pelabelan dilakukan terlebih dahulu sebelum melakukan tahap klasifikasi untuk mengetahui kelas sentimen dari data *tweet.* Kelas sentimen pada data *tweet* dilakukan menggunakan kamus yang terdiri dari kumpulan kata negatif dan positif. Pelabelan dilakukan secara otomatis dengan menggunakan *software* R. Tabel 5 merupakan hasil pelabelan yang telah dilakukan. Pelabelan dilakukan dengan cara menghitung jumlah skor kata positif dikurangi dengan jumlah skor kata negatif. *Tweet* dengan jumlah skor yang nilainya > 0 berlabel positif, dan jumlah skor yang nilainya < 0 maka berlabel negatif (Tabel 5).

Jumlah sentimen setiap kelas berdasarkan hasil dari pelabelan divisualisasikan menggunakan *barchart* pada Gambar 3.

Hasil pelabelan didapatkan bahwa tweet pengguna twitter menganai program kartu prakerja cenderung bersifat negatif, dengan tweet kelas positif sebanyak 2971 tweet dan kelas negatif sebanyak 919 dari 3890. Sehingga dapat disimpulkan bahwa tweet pengguna twitter mengenai program kartu prakerja cenderung bersifat positif dengan persentase 76,35%. Kategori tweet cenderung imbalance atau tidak seimbang antara kategori sentimen positif dan negatif.

C. Klasifikasi Menggunakan Metode Naïve Bayes Classifier

Tahap yang dilakukan sebelum melakukan klasifikasi yaitu membagi data menjadi data *training* dan data *testing* pada data *tweet.*

1) Pembagian Data Training dan Data Testing

Pembagian dilakukan menggunakan metode *10-fold cross validation.* Data yang digunakan pada tahap ini merupakan data yang telah dilakukan pelabelan data. Hasil pembagian data *training* dan data *testing* ditampilkan pada Tabel 6.

Data tweet sebanyak 3890 yang terdiri dari kelas sentimen positif sebanyak 2971 dan kelas negatif sebanyak 919. Pembagian data training dan data testing menggunakan 10-fold cross validation sehingga data dibagi 10 seperti yang telah digambarkan pada Tabel 6. Iterasi pertama, tabel dengan kotak berwarna abu-abu akan menjadi data testing dan sisanya akan menjadi data training, begitu seterusnya hingga iterasi ke 10. Data testing bukan merupakan bagian dari data training, di mana data testing tidak digunakan untuk membentuk model.

2) Klasifikasi Menggunakan Metode Naïve Bayes Classifier

Metode *Naïve Bayes Classifier* terdiri dari 2 tahap dalam proses klasifikasi data teks yaitu tahap pelatihan dan tahap klasifikasi. Tahap pelatihan yaitu membentuk model dengan melatih model menggunakan data *training.* Tahap kedua yaitu klasifikasi, pertama memperkirakan ketepatan prediksi dari model yang telah dilatih. Model yang telah dibangun digunakan untuk mengklasifikasikan data *training* dan data *testing* ke dalam kelas sentimen positif atau negatif. Pengukuran ketepatan prediksi dilakukan dengan membentuk *confusion matrix* dari hasil prediksi. Tabel 7 menampilkan hasil *confusion matrix* dari data *training Iterasi* 1.

Setelah terbentuk *confusion matrix* pada Tabel 7, langkah selanjutnya adalah melakukan perhitungan ketepatan klasifikasi. Ukuran ketepatan klasifikasi yang akan digunakan adalah *G-mean* dan AUC karena data cenderung tidak seimbang (*imbalance*). Hasil ketepatan klasifikasi menunjukkan besarnya nilai yang tepat terklasifikasi pada kelas aktual. Berikut merupakan hasil pengu kuran ketepatan klasifikasi dari setiap iterasi menggunakan algoritma *Naïve Bayes Classifier* (Tabel 8).

Ketepatan klasifikasi data *training* yang tinggi dihasilkan pada iterasi ke-2 dengan nilai *G-mean* sebesar 80,6% dan nilai AUC sebesar 81,6%. Pada data *testing* ketepatan klasifikasi tertinggi pada iterasi ke-9 dengan nilai *G-mean* dan AUC sebesar 89%. Secara keseluruhan hasil ketepatan klasifikasi metode NBC menggunakan 10 *fold cross validation* pada *tweet* pengguna *twitter* mengenai program kartu prakerja pada data *training* nilai *G-mean* sebesar 80,1% dan nilai AUC sebesar 81,2%. Sedangkana pada data *testing* nilai *G-mean* sebesar 69,8% dan nilai AUC sebesar 73,4%. Hasil ketepatan prediksi pada data *training* mendapatkan hasil yang lebih baik dari data *testing,* karena data yang digunakan untuk mengukur ketepatan prediksi data *training* sama dengan data untuk membentuk model.

V. KESIMPULAN

Hasil analisis dan pembahasan diperoleh kesimpulan bahwa dari data tweet pengguna twitter mengenai program kartu prakerja digunakan 3.890 data tweet. Kata yang sering digunakan dalam tweet yaitu kata "gelombang" kemudian kata "daftar" dan diikuti kata "bantuan", "pelatihan" serta "kerja" yang mengartikan bahwa pendaftaran dan informasi mengenai dibukanya gelombang kartu prakerja yang sering menjadi topik pembahasan mengenai program kartu prakerja. Hasil analisis sentimen program kartu prakerja cenderung bersifat positif. Sentimen positif disini menunjukkan bahwa banyak yang mendapatkan manfaat dengan adanya program kartu prakerja. Sentimen negatif menunjukkan kritik masyarakat mengenai kesulitan saat proses pendaftaran. Hasil klasifikasi menggunakan metode naïve bayes classifier didapatkan ketepatan klasifikasi pada data training nilai G-mean sebesar 80,1% dan nilai AUC sebesar 81,2%. Sedangkana pada data testing nilai G-mean sebesar 69,8% dan nilai AUC sebesar 73,4%.

Pemerintah dapat memperbaiki sistem pendaftaran program kartu prakerja dikarenakan banyaknya kritik masyarakat mengenai kesulitan saat proses pendaftaran serta lebih memperketat proses seleksi penerimaan program kartu prakerja sehingga program ini dapat terus berjalan dan penerima kartu prakerja tepat sasaran.

DAFTAR PUSTAKA

[1] B. Liu, *Handbook of Natural Language Processing*, 2nd ed. New York: CRC Press, 2010.

[2] T. Kurniawan, "Implementasi Text Mining pada Analisis Sentimen Pengguna Twitter Terhadap Media Mainstream Menggunakan Naïve Bayes Classifier dan Support Vector Machine," Institut Teknologi Sepuluh Nopember, 2017.

[3] W. Wibowo, N. P. Sari, R. N. Wilantari, and S. Abdul-Rahman, "Association Rule Mining Method for the Identification of Internet Use," in *Journal of Physics: Conference Series*, 2021, vol. 1874, no. 1, p. 12009, doi: 10.1088/1742-6596/1874/1/012009.

[4] N. Falahah, "Pengembangan Aplikasi Sentiment Analysis Menggunakan Metode Naive Bayes (Studi Kasus Sentiment Analysis dari media Twitter)," in *Seminar Nasional Sistem Informasi Indonesia*, 2015, pp. 2–3.

[5] H. Li and J. Sun, "Forecasting business failure: The use of nearest-neighbour support vectors and correcting imbalanced samples--Evidence from the Chinese hotel industry," *Tour. Manag.*, vol. 33, no. 3, pp. 622–634, 2012, doi: https://doi.org/10.1016/j.tourman.2011.07.004.

[6] R. Feldman, J. Sanger, and others, *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*, 1st ed. United States of Amerika: Cambridge university press, 2007, ISBN: 9780521836579.

[7] D. W. Freeman and W. A. Sistrunk, "Effects of post-harvest storage on the quality of canned snap beans," *J. Food Sci.*, vol. 43, no. 1, pp. 211–214, 1978, doi: https://doi.org/10.1111/j.1365-2621.1978.tb09773.x.

[8] Y. Sun, M. S. Kamel, and Y. Wang, "Boosting for Learning Multiple Classes with Imbalanced Class Distribution," in *Sixth International Conference on Data Mining (ICDM'06)*, 2006, pp. 592–602, doi: 10.1109/ICDM.2006.29.

[9] M. Bekkar, H. K. Djemaa, and T. A. Alitouche, "Evaluation measures for models assessment over imbalanced data sets," *J. Inf. Eng. Appl.*, vol. 3, no. 10, pp. 1–13, 2013.

[10] J. Davis and M. Goadrich, "The Relationship Between Precision-Recall and ROC Curves," in *Proceedings of the 23rd International Conference on Machine Learning*, 2006, pp. 233–240, doi: https://doi.org/10.1145/1143844.1143874.

[11] E. Gokgoz and A. Subasi, "Comparison of decision tree algorithms for EMG signal classification using DWT," *Biomed. Signal Process. Control*, vol. 18, pp. 138–144, 2015, doi: https://doi.org/10.1016/j.bspc.2014.12.005.

[12] R. Kohavi and others, "A Study of Cross-Validation and Bootstrap for Acuracy Estimation and Model Selection," in *International Joint Conference on Arti cial Intelligence (IJCAI)*, 1995, vol. 14, no. 2, pp. 1137–1145.