

Perbandingan Analisis Sentimen Mengenai BPJS pada Media Sosial *Twitter* Menggunakan *Naïve Bayes Classifier* (NBC) dan *Support Vector Machine* (SVM)

Diva Durrotun Nada, Soehardjoepri, dan R. Mohamad Atok
Departemen Aktuaria, Institut Teknologi Sepuluh Nopember (ITS)
e-mail: joepri_its@yahoo.com

Abstrak—Seiring dengan perkembangan teknologi, masyarakat saat ini dapat mengungkapkan perasaan, pendapat, atau pandangannya kepada publik melalui jejaring sosial. Salah satu media sosial terpopuler saat ini adalah *Twitter* yang diluncurkan oleh Jack Dorsey pada tanggal 15 Juli 2006. Media sosial ini merupakan salah satu media sosial utama yang digunakan masyarakat Indonesia untuk memberikan opini kepada pengguna internet. Karena jumlah pengguna *Twitter* yang cukup besar, hal ini sering digunakan oleh pemerintah, pelaku bisnis, maupun masyarakat untuk melihat pendapat pengguna tentang suatu produk atau layanan. Karena sebagian besar masyarakat Indonesia menggunakan BPJS, maka hal ini menyebabkan banyak pengguna media sosial seperti *Twitter* mengunggah ulasan mereka terkait kinerja BPJS. Hal ini dikarenakan hasil penelitian diperoleh langsung dari opini publik atas apa yang mereka alami, maka hasil tersebut dapat digunakan sebagai pengoptimalisasian program kerja, dan peningkatan kualitas pelayanan bagi perusahaan tersebut. Penelitian ini menggunakan dua metode untuk membandingkan tingkat akurasi antara metode *Naïve Bayes Classifier* dan *Support Vector Machine* menggunakan data *Twitter* berupa *tweet* umum mengenai kinerja BPJS dengan kata kunci “BPJS”, “Badan Penyelenggara Jaminan Sosial”, “Klaim” sejak Januari 2019 sampai Desember 2021. Hasil penelitian menunjukkan bahwa metode *Support Vector Machine* Kernel RBF dengan parameter $C = 1000$ dan $\gamma = 100$ memiliki performa ketepatan klasifikasi yang paling baik dibanding *Naïve Bayes Classifier* dan *Support Vector Machine* Kernel Linear. Dengan hasil rata-rata ketepatan klasifikasi SVM Kernel RBF, SVM Kernel Linear, dan *Naïve Bayes Classifier* masing-masing sebesar 97,1%, 92,5%, dan 86,7%.

Kata Kunci—Analisis sentimen, BPJS, *Naïve Bayes Classifier*, Pengajuan Klaim, *Support Vector Machine*, *Twitter*.

I. PENDAHULUAN

TWITTER adalah salah satu jejaring sosial online terpopuler yang memungkinkan pengguna untuk mengirim pesan singkat 280 karakter yang biasa disebut *tweets*. Media sosial yang didirikan oleh Jack Dorsey dan resmi diluncurkan pada 15 Juli 2006 ini merupakan salah satu media sosial utama masyarakat Indonesia untuk memberikan opini terhadap seseorang, pemerintah, barang, atau jasa. Pada kuartal keempat tahun 2020, *twitter* memiliki 192 juta pengguna aktif harian dan pada kuartal kedua tahun 2021, *twitter* memiliki 206 juta pengguna aktif harian. Karena sebagian masyarakat Indonesia adalah pengguna *twitter* dan BPJS, maka hal ini menyebabkan banyak pengguna mengunggah ulasan mereka terkait kinerja BPJS di sosial media *twitter*. Ulasan tersebut dapat berupa ulasan positif maupun negatif [1].

Proses untuk meninjau pendapat masyarakat ini sering disebut analisis sentiment. Dimana analisis sentimen atau bisa juga disebut *opinion mining* adalah bidang ilmu

komputer yang menganalisis pendapat, penilaian, sikap, emosi, sentiment, dan evaluasi terhadap suatu produk, layanan, organisasi, individu, tokoh public, maupun kegiatan tertentu [2]. Manfaat dari analisis sentimen sendiri adalah membantu perkembangan layanan dan aplikasi, karena hasil penelitian diperoleh langsung dari opini publik atas apa yang mereka alami dan hasil tersebut dapat digunakan sebagai optimalisasi produk, program kerja, maupun isu yang beredar.

Ada beberapa metode yang dapat digunakan dalam analisis sentimen antara lain, *Naïve Bayes Classifier*, *Support Vector Machine*, *K-Nearest Neighbor*, *Decision Tree*, dan masih banyak lagi. Pada penelitian ini untuk membandingkan hasil ketepatan klasifikasi pada data *twitter* mengenai kinerja BPJS selama pandemi, maka metode yang akan digunakan adalah *Naïve Bayes Classifier* dan *Support Vector Machine*.

II. TINJAUAN PUSTAKA

A. Preprocessing Data

Preprocessing data merupakan langkah awal dari *text mining* yang bertujuan untuk mempersiapkan dokumen menjadi data yang siap diolah pada tahap selanjutnya [3]. Berikut merupakan tahapan dalam *preprocessing* data [4].

1. Cleaning

Merupakan proses pembersihan *tweet* dari kata yang tidak diperlukan seperti HTML, *emoticons*, *hashtag*, *username*, dan *url*.

2. Case Folding

Merupakan proses mengubah kata ke dalam format yang sama, dalam hal ini menjadi format *lowercase* atau *uppercase*.

3. Stopword Removal

Merupakan proses menghapus kata-umum dan sering muncul tetapi tidak memiliki pengaruh yang signifikan terhadap makna kalimat.

4. Stemming

Merupakan proses mengubah kata menjadi kata dasar dengan menghilangkan awalan, akhiran, sisipan, dan *confixes*.

5. Tokenization

Proses memecah kalimat menjadi kata-kata yang lebih berarti dan bermakna.

B. Term Frequency-Invers Document Frequency

Term Frequency-Invers Document Frequency merupakan pengukuran yang digunakan untuk mengukur seberapa penting suatu kata terhadap kumpulan dokumen. *Term*

Frequency (TF) berfungsi untuk meringkas kemunculan sebuah kata pada suatu dokumen, sedangkan *Invers Document Frequency* (IDF) berfungsi menghitung banyaknya dokumen yang mengandung suatu kata tersebut. Rumus untuk menentukan pembobotan TF-IDF adalah sebagai berikut [3].

$$TF_{ij} = \frac{f_{ij}}{\sum_{i \in j} f_{ij}} \tag{1}$$

$$IDF = \log \left(\frac{N}{DF_{ij}} \right) \tag{2}$$

$$W_{ij} = TF_{ij} \times IDF \tag{3}$$

Dimana W_{ij} adalah bobot TF-IDF pada kata kunci ke- i dan *tweet* ke- j , TF_{ij} adalah jumlah kemunculan kata i pada *tweet* ke- j , f_{ij} adalah jumlah kata i untuk setiap *tweet*, $\sum_{i \in j} f_{ij}$ adalah jumlah *tweet* yang mengandung kata i , DF_{ij} adalah jumlah *tweet* ke- j yang mengandung kata i , N adalah jumlah seluruh *tweet*.

C. Oversampling

Oversampling adalah salah satu metode yang bertujuan untuk menambahkan jumlah data pada kelas minoritas dengan memanfaatkan Teknik *sampling* pada data kelas minoritas sehingga diharapkan rasio antar kelas dapat lebih seimbang [5]. Konsep penambahan data pada *oversampling* dibagi menjadi dua, yaitu *oversampling* menggunakan data asli seperti metode *Random Oversampling* dan penambahan menggunakan data sintetik seperti *Synthetic Minority Oversampling Technique* (SMOTE) [6]. Pada algoritma *Random Oversampling* (ROS), data kelas minoritas dipilih secara acak kemudian ditambahkan ke dalam data *training* sampai jumlah data kelas minoritas sama dengan kelas mayoritas.

D. K-Fold Cross Validation

K-Fold Cross Validation adalah Teknik memvalidasi keakuratan sebuah model yang dibentuk berdasarkan dataset tertentu. Metode ini banyak digunakan peneliti karena dapat mengurangi bias yang terjadi dalam pengambilan sampel [7]. Masing-masing *fold* memiliki jumlah data dengan ukuran yang sama atau mendekati sama. Menurut Kohavi (1995), penggunaan jumlah *fold* terbaik untuk uji validitas dianjurkan menggunakan 10 *fold* [8].

E. Naïve Bayes Classifier

Naïve Bayes Classifier merupakan salah satu teknik *data mining* yang sering digunakan untuk mengklasifikasikan data dalam jumlah yang besar dan dapat untuk memprediksi probabilitas keanggotaan suatu *class* [3]. Kelebihan dari *Naive Bayes Classifier* adalah algoritmanya sederhana tetapi memiliki tingkat akurasi yang relatif tinggi. Secara umum model probabilitas untuk *Naive Bayes Classifier* adalah sebagai berikut.

$$P(C_j|X_1, X_2, \dots, X_n) = \frac{P(X_1, X_2, \dots, X_n|C_j)P(C_j)}{P(X_1, X_2, \dots, X_n)} \tag{4}$$

Dimana $P(C_j|X_1, X_2, \dots, X_n)$ adalah probabilitas kelas C_j pada X_1, X_2, \dots, X_n (*posterior*), $P(X_1, X_2, \dots, X_n|C_j)$ adalah probabilitas X_1, X_2, \dots, X_n pada kelas C_j , $P(C_j)$ adalah probabilitas dari kelas C_j , $P(X_1, X_2, \dots, X_n)$ adalah probabilitas dari X_1, X_2, \dots, X_n .

Nilai dari $P(X_1, X_2, \dots, X_n)$ selalu tetap untuk tiap kelas pada suatu ulasan. Sehingga persamaan tersebut menjadi:

Tabel 1.
Fungsi kernel SVM

Fungsi Kernel	Rumus K ($xm1, xm2$)	Parameter
Linear	$x_i^T x_j$	C
RBF	$exp(-\gamma \ x_i - x_j\ ^2), \gamma > 0$	γ dan C

$$P(C_j|X_1, X_2, \dots, X_n) = P(X_1, X_2, \dots, X_n|C_j) \times P(C_j) \tag{5}$$

Nilai dari *posterior* ini nantinya akan dibandingkan dengan nilai *posterior* lainnya untuk menentukan klasifikasi kelas ulasan tersebut. Kelas terbaik ditentukan dengan mencari nilai *maximum a posterior* (MAP) dengan persamaan berikut.

$$C_{MAP} = \underset{C_j=1}{argmax} P(C_j|X_1, X_2, \dots, X_n) \tag{6}$$

Menurut persamaan 5 maka persamaan 6 dapat disesuaikan menjadi persamaan 7.

$$C_{MAP} = \underset{C_j=1}{argmax} P(X_1, X_2, \dots, X_n|C_j) \times P(C_j) \tag{7}$$

F. Support Vector Machine

Support vector machine adalah teknik prediksi yang digunakan untuk klasifikasi dan regresi. SVM merupakan salah satu dari metode yang dikembangkan untuk mengatasi permasalahan yang tidak bisa diselesaikan dengan metode statistika klasik terutama pada kasus klasifikasi dan prediksi. Metode ini mampu menyelesaikan permasalahan linear maupun non linear. Pada metode ini memerlukan parameter yaitu parameter *Cost* (C) dan *gamma* (γ). Tidak ada patokan nilai untuk kedua parameter ini sehingga harus melalui proses *trial and error*.

1. SVM pada *Linearly Separable Data*

Merupakan penerapan metode SVM pada data yang dapat dipisahkan secara linier. Konsep dari SVM pada *linear separable data* adalah menemukan *hyperplane hyperplane* yang optimum pada input *space* yang berfungsi sebagai pemisah dua buah kelas yang sering disimbolkan dengan -1 dan +1 [4].

2. SVM pada *Non-Linearly Separable Data*

Pada data riil, sangat jarang ditemukan masalah yang berifat *linear separable*. Oleh karena itu SVM membutuhkan fungsi yang mampu membuat pemisah yang tidak linier. Fungsi yang sering digunakan untuk mengatasi hal tersebut adalah fungsi kernel. Fungsi kernel yang umum digunakan diberikan pada Tabel 1 [4].

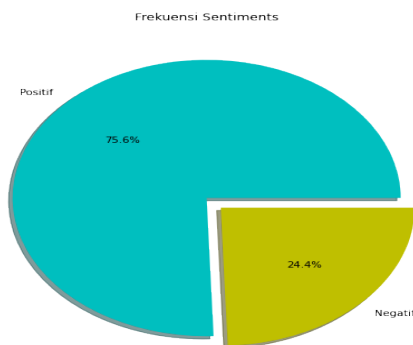
Fungsi kernel *Gaussian RBF* memiliki kelebihan yaitu secara otomatis menentukan nilai, lokasi dari *center* serta nilai pembobot dan dapat mencakup nilai rentang tak terhingga.

G. Klasifikasi

Hasil klasifikasi dan prediksi dapat dievaluasi menggunakan pengukuran ketepatan klasifikasi. Dalam mengukur ketepatan klasifikasi, perlu diketahui jumlah pada setiap kelas prediksi dan actual yang terdiri dari *True Positive* (TP) yaitu jumlah ulasan bersentimen positif yang tepat terprediksi dalam kelas positif, *True Negative* (TN) yaitu jumlah ulasan yang bersentimen negatif tepat terprediksi dalam kelas negatif, *False Positive* (FP) yaitu jumlah ulasan

Tabel 2.
Confusion matrix

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



Gambar 1. Diagram presentase jumlah sentimen.

bersentimen negatif yang terprediksi dalam kelas positif, dan *False Negative* (FN) yaitu jumlah ulasan bersentimen positif yang terprediksi dalam kelas negatif. Semua nilai tersebut kemudian ditulis ke dalam *Confusion matrix*, diberikan pada Tabel 2, yang merupakan salah satu *tools* penting dalam metode visualisasi [9]. Dengan akurasi diberikan pada Persamaan 4.

$$Accuracy = \frac{TN+TP}{TN+TP+FP+FN} \tag{4}$$

H. Word Cloud

Merupakan sebuah sistem yang memunculkan kata sebagai citra visual terkait frekuensi kemunculan kata dalam suatu teks. Visualisasi *word cloud* akan memudahkan pengamat dalam melihat gagasan sehingga dapat menjadi alat bantu dalam melakukan analisis terhadap sebuah wacana tertulis [10].

III. METODOLOGI PENELITIAN

A. Sumber Data

Data yang digunakan pada penelitian ini adalah data sekunder berupa kumpulan *tweet* mengenai ulasan atau tanggapan pengguna *twitter* mengenai kinerja BPJS. Data tersebut berupa *tweet* berbahasa Indonesia dengan kata kunci “BPJS” yang diambil menggunakan teknik *scrapping* dengan Bahasa pemrograman *Python*.

B. Variabel Penelitian

Data terdiri dari dua variabel, yaitu variabel predictor (*X*) dan variabel respon (*Y*). Dimana *X* adalah frekuensi kata di setiap *tweet* dan *Y* adalah klasifikasi sentimen *tweet* positif dan negatif.

C. Metode Analisis

Langkah pertama dalam penelitian ini adalah mengambil data *tweet* yang mengandung kata ‘BPJS’ pada social media *twitter*. Lalu data yang telah diambil tersebut dilakukan *preprocessing text* untuk menghapus HTML, emoticons, hashtag, username, url, slang, singkatan, kata-kata yang sering muncul tapi tidak memiliki makna yang berarti, mengubah kata imbuhan menjadi kata dasar, merubah

Tabel 3.

Beberapa *tweet* setelah *preprocessing text*

Sebelum	Sesudah
Disnaker Lampung mengusulkan petugas Komisi Pemilihan Umum tingkat desa hingga provinsi dilindungi BPJS Ketenagakerjaan.#disnakerlampung #KPU #KPUMelayani #bpjskesehatan #lampostco https://t.co/2nUaexGu5i	'disnaker', 'lampung', 'usul', 'tugas', 'komisi', 'pilih', 'tingkat', 'desa', 'provinsi', 'lindung', 'bpjs', 'ketenagakerjaan'
BPJS Kesehatan mendapatkan 3 penghargaan dalam Penghargaan Top Digital Awards 2021 yang diselenggarakan oleh Majalah ItWorks. https://t.co/Lx0ss3567d	'bpjs', 'sehat', 'harga', 'harga', 'top', 'digital', 'awards', 'selenggara', 'majalah', 'itworks'
Premi BPJS Kesehatan Bakal Ditentukan sesuai Jumlah Harta Kekayaan https://t.co/toea1sk714	'premi', 'bpjs', 'sehat', 'tentu', 'sesuai', 'harta', 'kaya'
Kesabaran tampaknya sudah menjadi syarat wajib buat peserta BPJS Kesehatan jika ingin berobat. #Visi #IuranBPJS https://t.co/eIITzBqVh3	'sabar', 'syarat', 'wajib', 'serta', 'bpjs', 'sehat', 'obat'
Disnaker Lampung mengusulkan petugas Komisi Pemilihan Umum tingkat desa hingga provinsi dilindungi BPJS Ketenagakerjaan.#disnakerlampung #KPU #KPUMelayani #bpjskesehatan #lampostco https://t.co/2nUaexGu5i	'disnaker', 'lampung', 'usul', 'tugas', 'komisi', 'pilih', 'tingkat', 'desa', 'provinsi', 'lindung', 'bpjs', 'ketenagakerjaan'

kalimat menjadi huruf kecil, dan memecah ulasan menjadi per kata. Setelah melakukan *preprocessing text*, langkah selanjutnya adalah melakukan pelabelan kelas pada setiap ulasan dan melakukan pembobotan kata menggunakan metode TF-IDF. Karena pada penelitian ini jumlah kategori sentiment tidak seimbang, maka dilakukan metode *oversampling* untuk mengatasi ketidakseimbangan tersebut. Setelah melakukan proses *oversampling* selanjutnya adalah membagi data *testing* dan *training* menggunakan *k-fold cross validation* dalam penelitian ini menggunakan *10-fold*. Selanjutnya melakukan pemodelan klasifikasi menggunakan *Naïve Bayess Classifier* dan *Support Vector Machine*. Untuk pemodelan klasifikasi pada *Support vector machine* parameter *C* dan γ yang digunakan mulai dari 10^{-3} sampai 10^3 . Setelah melakukan pemodelan, langkah selanjutnya adalah menghitung ketepatan klasifikasi menggunakan *confusion matrix*. Setelah itu membandingkan dan menginterpretasi hasil ketepatan klasifikasi antara *naïve bayess classifier* dengan *support vector machine*. Setelah membandingkan dan menginterpretasi hasil, selanjutnya adalah membuat visualisasi *word cloud*. Dan yang terakhir, menarik kesimpulan dan saran untuk penelitian.

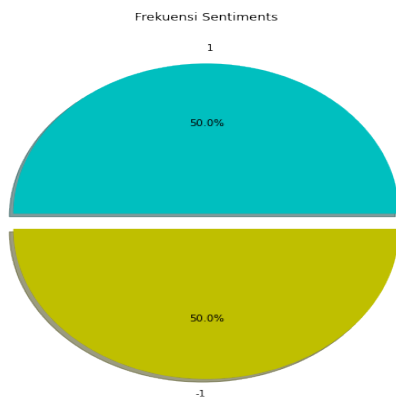
IV. HASIL DAN PEMBAHASAN

A. Pre-processing Text

Data *tweet* yang telah dikumpulkan kemudian dilakukan *preprocessing text* yang meliputi *cleaning*, *case folding*, *stemming*, *stopwords removal*, dan *tokenizing*. Tabel 3 merupakan beberapa *tweet* setelah *preprocessing text*.

B. Pelabelan Kelas Sentimen

Setelah dilakukan *preprocessing text* langkah selanjutnya adalah memberi label pada kelas sentimen. Pelabelan ini dilakukan secara manual dengan melihat banyak kata negatif atau positif dalam satu *tweet*. Untuk sentimen positif akan diberi label 1 dan untuk kelas negative akan diberi label -1. Gambar 1 merupakan diagram frekuensi jumlah sentimen.



Gambar 2. Diagram presentase jumlah sentiment setelah *oversampling*.

Tabel 4.
Frekuensi tiap kata pada *tweet*

<i>Tweet</i>	desa	disnaker	dosis	...	tugas	turun	usul
1	1	1	0		1	0	1
2	0	0	1		0	0	0
3	0	0	0	...	0	0	0
4	0	0	0		0	1	0

Tabel 5.
Pembobotan *tweet* menggunakan TF-IDF

<i>Tweet</i>	desa	disnaker	dosis	...	tugas	turun	usul
1	0,32	0,44	0,00		0,32	0,00	0,30
2	0,00	0,00	0,41		0,00	0,00	0,00
3	0,00	0,00	0,00	...	0,00	0,00	0,00
4	0,00	0,00	0,00		0,00	0,55	0,00

Dapat dilihat pada gambar sebanyak 75,6% sentimen merupakan sentimen positif dan 24,4% merupakan sentimen negatif. Karena proporsi kategori kelas negatif kurang dari 35%, maka hal ini menunjukkan bahwa data cenderung tidak seimbang atau *imbalance*. Keadaan ini akan berpengaruh pada perhitungan ketepatan klasifikasi dan menghasilkan hasil yang bias. Maka untuk mengatasi hal tersebut perlu dilakukan *oversampling* agar kelas antar kategori dapat lebih seimbang.

C. *Oversampling*

Dengan menambahkan jumlah data pada kelas minoritas, dimana dalam penelitian ini adalah kelas negatif sehingga diharapkan rasio antar kelas dapat lebih seimbang. Gambar 2 merupakan hasil setelah dilakukannya *oversampling*.

Dapat dilihat pada gambar bahwa proporsi yang tadinya sangat tidak seimbang setelah dilakukan *oversampling* kini menjadi sama besar. Hal ini sejalan dengan konsep algoritma dari *oversampling* itu sendiri. Sehingga kini data yang semula hanya 2662 data kini menjadi 4024 data.

D. *Term Frequency-Invers Document Frequency*

Setelah diberi label, maka data *tweet* akan dilakukan pembobotan dengan metode TF-IDF. Sebelum dibobot dengan TF-IDF, kata-kata yang terdapat pada *tweet* akan dihitung frekuensinya lebih dulu. Tabel 4 merupakan perhitungan frekuensi untuk tiap kata pada *tweet*.

Perhitungan frekuensi ini dilakukan sampai *tweet* terakhir kemudian dari hasil perhitungan frekuensi ini didapatkan perhitungan bobot untuk tiap kata tersebut. Tabel 5 merupakan pembobotan dari tiap kata yang telah dihitung frekuensi nya.

Tabel 6.
Hasil probabilitas *naive bayes classifier*

Negatif	Positif	Keputusan
0,605983	0,394017	Negatif
0,513188	0,486812	Negatif
0,196549	0,803451	Positif
0,327767	0,672233	Positif
0,882198	...	Negatif
0,868786	0,117802	Negatif
0,602315	0,131214	Negatif
0,397685	0,397685	Negatif
0,647539	0,352461	Negatif

Tabel 7.
Pengukuran ketepatan klasifikasi *naive bayes classifier*

Iterasi	Accuracy (%)
1	89,6
2	87,3
3	84,6
4	86,1
5	87,1
6	86,6
7	86,6
8	88,1
9	86,8
10	84,6
Rata-rata	86,7

Tabel 8.
Confusion matrix naive bayes classifier

Kelas Prediksi	Kelas Aktual	
	Positif	Negatif
Positif	1869	143
Negatif	391	1621

E. *Naive Bayes Classifier*

Setelah membagi data menjadi data *training* dan *testing* dengan *stratified 10-fold cross validation* dimana terdapat 10 *subset* yang tiap *subset* nya terdiri dari 10% data *testing* dan 90 data *training* yang berbeda. Pada pengklasifikasian dengan *Naive Bayes* ini akan menghasilkan probabilitas yang dapat digunakan untuk menentukan apakah *tweet* tersebut masuk ke dalam kategori positif atau negatif. Tabel 6 merupakan beberapa nilai probabilitas yang dihasilkan dari model *Naive Bayes Classifier*.

Nilai probabilitas pada Tabel 6 menunjukkan peluang suatu *tweet* akan masuk ke dalam salah satu kategori dengan nilai probabilitas terbesar. Sehingga apabila suatu *tweet* memiliki nilai probabilitas dengan kategori positif lebih besar maka *tweet* tersebut masuk ke dalam kategori positif. Setelah mendapat nilai probabilitas dari setiap *tweet*, selanjutnya adalah menghitung ketepatan klasifikasi dari model. Tabel 7 merupakan tabel ketepatan klasifikasi dari *Naive Bayes Classifier*.

Nilai pada Tabel 7 menunjukkan bahwa pengukuran ketepatan klasifikasi dari subset sebanyak 10 kali menghasilkan nilai rata-rata sebesar 86,7%. Nilai *accuracy* ini menunjukkan rasio prediksi benar untuk kelas positif dan negatif dengan seluruh data. Tabel 8 merupakan tabel *Confusion matrix* dari model *Naive Bayes Classifier*.

Dapat dilihat pada Tabel 6, bahwa dari seluruh jumlah sentimen terdapat 1869 sentimen positif yang terklasifikasi benar dan terdapat kesalahan prediksi sebanyak 143 sentimen yang masuk ke dalam sentiment negatif. Dan sebanyak 1621

Tabel 11.
Pengukuran ketepatan klasifikasi SVM kernel linear $C = 100$

Iterasi	Accuracy (%)
1	91,3
2	91,8
3	90,8
4	91,8
5	93,0
6	93,8
7	95,0
8	93,5
9	92,5
10	91,0
Rata-rata	92,5

Tabel 12.
Confusion matrix SVM kernel linear $C = 100$

Kelas Prediksi	Kelas Aktual	
	Positif	Negatif
Positif	1965	47
Negatif	256	1756

sentimen negatif terklasifikasi benar dan terdapat kesalahan prediksi sebanyak 391 sentimen yang masuk kedalam sentiment positif. Berdasarkan Tabel 8 *confusion matrix* tersebut, diperoleh hasil ketepatan klasifikasi untuk model *Naïve Bayes Classifier* adalah sebesar 86,7% yang berarti sebanyak 3490 sentimen terklasifikasi dengan benar. Hal ini dapat dikatakan bahwa metode *Naïve Bayes Classifier* baik digunakan dalam mengklasifikasi sentimen data *twitter* mengenai BPJS.

F. Support Vector Machine

Pada penelitian ini terdapat dua jenis kernel yang digunakan, pertama SVM Kernel Linear, dan SVM Kernel RBF. Tahap yang dilakukan pada metode ini hampir sama dengan *Naïve Bayes Classifier*, hanya saja pada SVM membutuhkan parameter C dan γ .

1. SVM Kernel Linear

Pada metode SVM Kernel Linear, parameter C yang digunakan akan bernilai 10^{-3} sampai 10^3 . Setelah dilakukan pemodelan, dan dihitung ketepatan klasifikasi, didapatkan bahwa nilai ketepatan klasifikasi untuk parameter $C = 100$ lebih baik dibanding parameter lainnya. Tabel 9 merupakan *confusion matrix* untuk SVM Kernel Linear dengan parameter $C = 100$.

Nilai pada Tabel 9 menunjukkan bahwa pengukuran ketepatan klasifikasi SVM Kernel Linear dengan subset 10 kali menghasilkan nilai rata-rata sebesar 92,5%. Tabel 10 merupakan tabel *Confusion Matrix* dari model SVM Kernel Linear.

Dari Tabel 10, diketahui bahwa dari seluruh jumlah sentimen terdapat 1965 sentimen positif yang terklasifikasi benar dan terdapat kesalahan prediksi sebanyak 47 sentimen yang masuk ke dalam sentimen negatif. Dan sebanyak 1756 sentimen negatif terklasifikasi benar dan terdapat kesalahan prediksi sebanyak 256 sentimen yang masuk kedalam sentimen positif. Berdasarkan Tabel 10 *confusion matrix* tersebut, diperoleh hasil ketepatan klasifikasi untuk model *Support Vector Machine* Kernel Linear dengan parameter $C = 100$ adalah sebesar 92,5%, yang berarti terdapat 3721 sentimen yang terklasifikasi dengan benar. Hal ini dapat dikatakan bahwa metode *Support Vector Machine* Kernel

Tabel 9.
Ketepatan klasifikasi SVM kernel RBF $C = 1000 \gamma = 100$

Iterasi	Accuracy (%)
1	92,8
2	94,0
3	93,5
4	97,8
5	99,0
6	99,5
7	98,8
8	98,3
9	99,0
10	98,5
Rata-rata	97,1

Tabel 10.
Confusion matrix SVM kernel RBF $C = 1000$ dan $\gamma = 100$

Kelas Prediksi	Kelas Aktual	
	Positif	Negatif
Positif	1922	90
Negatif	26	1986

Linear dengan parameter $C = 100$ baik digunakan dalam mengklasifikasi sentimen data *twitter* mengenai BPJS.

2. SVM Kernel RBF

Pada SVM Kernel RBF ada dua parameter yang digunakan pada metode ini, yaitu parameter C dan γ . Rentang nilai yang digunakan untuk kedua parameter ini adalah 10^{-3} sampai 10^3 . Setelah dilakukan pemodelan dan perhitungan ketepatan klasifikasi didapatkan bahwa parameter $C = 1000$ dan $\gamma = 100$ memiliki hasil ketepatan klasifikasi tertinggi diantara parameter lainnya. Tabel 11 merupakan *confusion matrix* untuk SVM Kernel RBF dengan parameter $C = 1000$ dan $\gamma = 100$.

Nilai pada Tabel 11 menunjukkan bahwa pengukuran ketepatan klasifikasi dengan SVM Kernel RBF dengan subset 10 kali menghasilkan nilai rata-rata ketepatan klasifikasi sebesar 97,1%. Tabel 12 merupakan tabel *confusion matrix* SVM Kernel RBF.

Dari Tabel 12, diketahui bahwa dari seluruh jumlah sentimen terdapat 1922 sentimen positif yang terklasifikasi benar dan terdapat kesalahan prediksi sebanyak 90 sentimen yang masuk ke dalam sentimen negatif. Dan sebanyak 1986 sentimen negatif terklasifikasi benar dan terdapat kesalahan prediksi sebanyak 26 sentimen yang masuk kedalam sentimen positif. Berdasarkan tabel *confusion matrix* tersebut, diperoleh hasil ketepatan klasifikasi untuk model *Support Vector Machine* Kernel RBF dengan parameter $C = 1000$ dan $\gamma = 100$ adalah sebesar 97,1%, yang berarti terdapat 3908 sentimen yang terklasifikasi dengan benar. Hal ini dapat dikatakan bahwa metode *Support Vector Machine* Kernel RBF dengan parameter $C = 1000$ dan $\gamma = 100$ baik digunakan dalam mengklasifikasi sentimen data *twitter* mengenai BPJS.

G. Perbandingan Naïve Bayes Classifier dan Support Vector Machine

Setelah memperoleh hasil ketepatan klasifikasi ketiga metode, maka langkah selanjutnya adalah membandingkan hasil ketepatan klasifikasi dari kedua metode tersebut.

Pada Tabel 13 diketahui bahwa secara keseluruhan hasil ketepatan klasifikasi dengan menggunakan metode SVM Kernel Linear maupun Kernel RBF lebih baik dibandingkan metode *Naïve Bayes Classifier*. Hal tersebut dilihat dari nilai

