Analisis Sentimen Pelaporan Masyarakat di Situs Media *Centre* Surabaya dengan *Naïve Bayes Classifier*

Kukuh Tri Pamungkas, Lucia Aridinanti, dan Wahyu Wibowo
Departemen Statistika Bisnis, Institut Teknologi Sepuluh Nopember (ITS)
e-mail: lucia_a@statistika.its.ac.id

***Abstrak*—Media *center* merupakan sebuah sistem pelayanan terintegrasi bagi masyarakat Surabaya. Melalui media *center* masyarakat dapat berpartisipasi memberikan opini atau melaporkan hal-hal yang terkait dengan pembangunan kota Surabaya. Informasi pelaporan masyarakat bisa negatif bisa positif untuk itu perlu dilakukan pengelompokan. Selama ini pengelompokan dilakukan secara manual. Hal ini membutuhkan waktu yang lama, untuk itu dibutuhkan metode pengelompokan yang lebih cepat dan akurat. Dengan menggunakan analisis sentimen dan *Naïve Bayes Classifier* (NBC) diperoleh pelaporan masyarakat Surabaya tahun 2020 memiliki pelaporan yang bersifat negatif 56,03% dan kelas kategori netral 16,22% serta kelas kategori positif 27,75%, dan hasil klasifikasi menghasilkan ketepatan klasifikasi yang cukup tinggi pada data testing dengan tingkat sensitifitas dan *specifity* yang ditunjukkan oleh *G-Mean* dan AUC sebesar 53,14% dan 55,12%.**

Kata Kunci*— Analisis Sentimen, Media *Center, Naive Bayes Classifier,* Pelaporan Masyarakat, *Sensitifitas, Specificity.

I. PENDAHULUAN

KOTA Surabaya merupakan kota terbesar kedua di Indonesia, kota Surabaya juga telah menerapkan teknologi informasi dan komunikasi dalam pemerintahan atau disebut *e-government*. Masyarakat dapat memberikan masukan terhadap kebijakan yang dibuat oleh pemerintah sehingga dapat meningkatkan kinerja pemerintah. *E-government* di kota Surabaya meliputi berbagai macam bidang salah satunya media *center* kota Surabaya. Pemerintah kota Surabaya menyadari perlu adanya keterbukaan informasi kepada masyarakat dengan tidak adanya pembatasan kepada masyarakat perihal memberikan aspirasi, keluhan atau informasi terkait kota Surabaya sehingga akhirnya Dinkominfo meluncurkan media *center* kota Surabaya oleh Dinas Komunikasi dan Informatika kota Surabaya (Dinkominfo). Media *center* ini adalah sistem pelayanan terintegrasi bagi masyarakat Surabaya yang ingin berpartisipasi dalam perkembangan pembangunan kota Surabaya. Melalui media *center* kota Surabaya, masyarakat Surabaya dapat mengetahui sejauh mana tahapan pembangunan yang disusun oleh pemerintah kota Surabaya. Sejak pertama kali di-*launching* pada 28 November 2011, media *center* telah banyak dimanfaatkan oleh masyarakat Surabaya. Berdasarkan data dari Dinas Komunikasi dan Informatika kota Surabaya, pelaporan yang masuk pada tahun pertama sebanyak 698.

Saat ini pengelompokan pelaporan masyarakat yang dilakukan oleh Dinkominfo kota Surabaya masih dilakukan secara manual yang membutuhkan waktu yang lama saat mengelompokkan pelaporan masyarakat. Informasi terkait dari pelaporan masyarakat harus segera disampaikan kepada masyarakat, sedangkan pelaporan masyarakat yang masuk dari tahun ke tahun semakin banyak. Dokumen yang umumnya memiliki data dengan jumlah besar dan bervariasi dapat menyulitkan dalam membuat model klasifikasi. Sehingga dokumen-dokumen tersebut dapat dikelompokkan menurut kelas kategori yang sama agar dapat dengan mudah diklasifikasi, untuk mengetahui pelaporan masyarakat apa yang sering terjadi dalam pembangunan yang disusun oleh pemerintah Kota Surabaya [1]. Dalam hal ini diperlukan metode alternatif yang dapat mempercepat pengelompokan dengan tingkat akurasi yang tinggi. Analisis sentimen adalah salah satu metode alternatif pengelompokan dan NBC akan meningkatkan tingkat akurasi pengelompokan.

Penelitian ini menggunakan data teks pelaporan masyarakat Surabaya tahun 2020. Metode analisis yang digunakan adalah analisis sentimen untuk mengelompokkan dan NBC untuk penetapan klasifikasi dari pengelompokan. Diharapkan, hasil penelitian ini dapat menjadi masukan bagi pemerintah kota Surabaya dalam mengambil tindakan dan kebijakan yang tepat dalam pelaporan masyarakat.

II. TINJAUAN PUSTAKA

A. Media Center

Media *center* adalah sistem pelayanan terintegrasi bagi masyarakat Surabaya yang ingin berpartisipasi dalam perkembangan pembangunan kota Surabaya. Melalui media *center*, masyarakat juga bisa mengetahui sejauh mana tahapan pembangunan yang disusun oleh pemerintah, dapat dilaksanakan sesuai sasaran. Media *center* terdapat dalam berbagai media diantaranya *facebook*, *twitter*, *instagram*, layanan sms, *e-wadul*, *whatsapp* dan lain-lainnya. Bahkan masyarakat dapat langsung datang ke kantor media *center* yang bertempat di Jalan Jimerto.

B. Analisis Sentimen

Analisis sentimen adalah studi komputasi mengenai sikap, emosi, pendapat, penilaian pandangan dari sekumpulan teks yang fokusnya dalam mengekstraksi, mengidentifikasi atau menemukan karakteristik sentimen dalam unit teks menggunakan metode NLP (*Natural Language Processing*), statistik atau *machine learning* [2]. Pada penelitian ini menggunakan kamus sentimen positif dan negatif dengan judul *"ID-Opinion Word"*. Berikut beberapa isi dari kamus *"ID-Opinion Word"* yang dapat dilihat pada Tabel 1. Langkah-langkah analisis sentimen adalah sebagai berikut:

1) Memanggil Kamus Sentimen

Memanggil kamus sentimen positif dan negatif dengan judul *"ID-Opinion Word"* ke *sofware,* dengan format kamus dalam bentuk text.

Tabel 1.
Kamus Sentimen

Kamus Sentimen		
No	Kategori Positif	Kategori Negatif
1	Ahli	Ancam
2	Absolut	Apatis
3	Akademisi	Aneh
⋮	⋮	⋮
4536	Tidak ada tandingan	Tidak punya biaya

2) Menghitung Skor Sentimen.

Menghitung skor sentimen digunakan untuk menentukan kelas kategori dari setiap kelas sentimen dimana skor sentimen dihitung dengan mencari jumlah kata yang memiliki sentimen positif dan sentimen negatif. Jika X_i adalah kata positif ke-i dan X_j adalah kata *negative* ke-j maka dapat dihitung skor untuk suatu komentar dengan persamaan 1.

$$Skor = (\sum_{i=1}^{n} X_i) - (\sum_{i=1}^{n} X_j) \quad (1)$$

3) Pemberian Kelas Kategori Sentimen

Pemberian kelas kategori sentimen ini bertujuan untuk menentukan kelas kategori berdasarkan skor yang telah diperoleh. Jika skor bernilai lebih besar dari nol maka akan dikelompokkan dalam kelas kategori positif, sedangkan jika skor bernilai sama dengan nol maka akan dikelompokkan dalam kelas kategori netral, dan jika skor bernilai kurang dari nol maka akan dikelompokkan dalam kelas kategori negatif [3]. Ilustrasi proses pemberian kelas kategori dapat dilihat pada Gambar 1.

C. Text Mining

Text mining merupakan bidang interdisiplin yang mengacu pada perolehan informasi (*information retrieval*), data *mining*, pembelajaran mesin (*machine learning*), statistik dan komputasi *linguistic* [4], sedangkan data *mining* adalah proses yang menggunakan teknik statistik dan matematika, buatan kecerdasan, serta pembelajaran mesin untuk mengekstrak dan mengidentifikasi informasi yang berguna dan terkait pengetahuan dari berbagai *database* besar [5]. Langkah-langkah yang dilakukan dalam *text mining* adalah sebagai berikut:

1) Text Pre-processing

Langkah-langkah dalam *text pre-processing* terbagi menjadi:

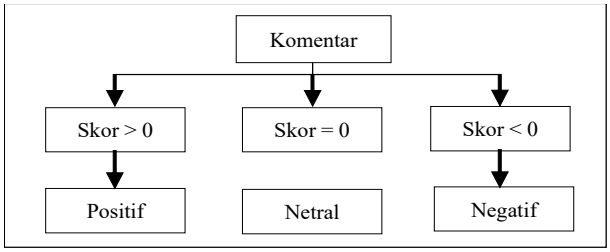
a. *Case folding* dan *Tokenizing*

Case folding adalah mengubah semua huruf dalam dokumen menjadi huruf kecil atau tidak kapital. Hanya huruf "a" sampai dengan "z" yang diterima. Karakter selain huruf dihilangkan. Tahap *tokenizing/parsing* adalah tahap pemotongan *string input* berdasarkan tiap kata yang menyusunnya. Proses pemisahan teks menjadi potongan kalimat dan kata yang disebut token [6].

b. *Spelling Normalization*

Spelling Normalization merupakan proses perbaikan atau subtitusi kata-kata yang salah eja atau disingkat dalam bentuk tertentu. Subtitusi kata dilakukan untuk menghindari jumlah perhitungan dimensi kata yang melebar. Perhitungan dimensi kata akan melebar jika kata yang salah eja atau disingkat tidak diubah karena kata tersebut sebenarnya mempunyai maksud dan arti yang sama tetapi akan dianggap sebagai entitas yang berbeda pada saat proses penyusunan matriks [7].

c. *Filtering*



Gambar 1. Dendogram Rasio Tenaga Kesehatan terhadap Jumlah Penduduk di Provinsi Papua.

Filtering adalah tahap mengambil kata-kata penting dari hasil token. Bisa menggunakan algoritma *stoplist* (membuang kata yang kurang penting) atau *wordlist* (menyimpan kata penting). *Stoplist/stopword* adalah kata-kata yang tidak deskriptif yang dapat dibuang dalam pendekatan *bag-of-words*. Contoh *stopwords* adalah **"**yang**"**, **"**dan**"**, **"**di**"**, **"**dari**"**, dan seterusnya [6].

d. *Stemming*

Tahap *stemming* adalah tahap mencari *root* kata dari tiap kata hasil *filtering*. Pada tahap ini dilakukan proses pengembalian berbagai bentukan kata ke dalam suatu representasi yang sama. Tahap ini kebanyakan dipakai untuk teks berbahasa Inggris dan lebih sulit diterapkan pada teks berbahasa Indonesia. Hal ini dikarenakan bahasa Indonesia tidak memiliki rumus bentuk baku yang permanen [8]. Sebagai contoh, kata bersama, kebersamaan, menyamai, akan di-*stem* ke *root* katanya yaitu **"**sama**"**. Proses *stemming* pada teks berbahasa Indonesia berbeda dengan *stemming* pada teks berbahasa Inggris. Pada teks berbahasa Inggris, proses yang diperlukan hanya proses menghilangkan *sufiks*. Sedangkan pada teks berbahasa Indonesia, selain *sufiks*, *prefiks*, dan *konfiks* juga dihilangkan [9].

2) Term Frequency-Inverse Document Frequency (TF-IDF)

Metode TF-IDF merupakan suatu cara untuk memberikan bobot hubungan suatu kata (token) terhadap suatu dokumen. Metode ini menggunakan dua konsep dalam perhitungan bobot yaitu frekuensi kemunculan sebuah kata di dalam sebuah dokumen dan *inverse* dari frekuensi dokumen yang mengandung kata tersebut. Frekuensi dokumen yang mengandung kata tersebut menunjukkan seberapa umum kata tersebut [10]. Jika w_{td} adalah bobot kata/token t_t terhadap dokumen d_d, sedangkan tf_{td} adalah jumlah kemunculan kata/token t_t dalam dokumen d_d dan N adalah jumlah semua dokumen, serta df_t adalah jumlah dokumen yang mengandung kata/token t_t. Maka bobot kata terhadap dokumen, w_{td} dapat ditentukan dengan persamaan 2.

$$w_{td} = tf_{td} * idf \quad (2)$$

$$w_{td} = tf_{td} * log\left(\frac{N}{df_t}\right) \quad (3)$$

3) Feature Selection

Pemilihan fitur (kata) merupakan tahap lanjut dari pengurangan dimensi pada proses transformasi teks. Pemilihan hanya dilakukan terhadap kata-kata yang relevan yang benar-benar mempresentasikan isi dari suatu dokumen. Algoritma yang digunakan pada *text mining*, biasanya tidak hanya melakukan perhitungan pada dokumen saja tetapi juga pada *feature* [11].

D. K-fold Cross Validation

K-fold cross validation adalah metode yang digunakan untuk membagi data menjadi data *training* dan data *testing*. Metode ini digunakan karena dapat mengurangi bias yang terjadi dalam pengambilan sampel. *K-fold cross validation* secara berulang-ulang membagi data menjadi data *training* dan data *testing*, dimana setiap data mendapat kesempatan menjadi data *testing* [12].

E. Naïve Bayes Classifier

Metode *Naïve Bayes Classifier* merupakan metode klasifikasi yang berdasar kepada teorema *bayes*, sebuah teorema yang terkenal di dalam bidang ilmu probabilitas. Selain itu, metode ini turut didukung oleh ilmu statistika khususnya dalam penggunaan data petunjuk untuk mendukung keputusan pengklasifikasian. Metode ini sangat luas dipakai dalam berbagai bidang, khususnya dalam proses klasifikasi dokumen. Jika A adalah hipotesis data merupakan suatu *class* spesifik, sedangkan B adalah data dengan kelas yang masih belum diketahui dan P(A|B) adalah probabilitas hipotesis A berdasar kondisi B, sedangkan P(B|A) adalah probabilitas B berdasarkan kondisi pada hipotesis A, serta P(A) adalah probabilitas hipotesis A, sedangkan P(B) adalah probabilitas B. Maka Teori dari *Naïve Bayes* dapat dinotasikan dengan persamaan 4 [13]:

$$P(A|B) = \frac{P(B|A)\times P(A)}{P(B)} \quad (4)$$

Kelebihan NBC adalah algoritmanya sederhana tetapi memiliki akurasi yang tinggi [14]. Dalam algoritma NBC setiap dokumen direpresentasikan dengan pasangan atribut "a_1, a_2, a_3,..., a_n" dimana a_1 adalah kata pertama, a_2 adalah kata kedua dan seterusnya. Sedangkan v adalah himpunan kategori data teks. Pada saat klasifikasi algoritma akan mencari probabilitas tertinggi dari semua kategori dokumen yang diujikan (V_{MAP}). Adapun V_{MAP} dapat dilihat pada persamaan 5 [15].

$$V_{MAP} = argmax\, P\,(v_j|a_1, a_2, \dots a_n) \quad (5)$$

Berdasarkan persamaan 4 dan 5, maka Persamaan 6 dapat ditulis sebagai berikut.

$$V_{MAP} = \underset{v_j \in V}{argmax} \frac{P(a_1,a_2,\dots a_n|v_j)P(v_j)}{P(a_1,a_2,\dots a_n)} \quad (6)$$

$P(a_1,a_2,\dots a_n)$ konstan, sehingga dapat dihilangkan menjadi persamaan 7.

$$V_{MAP} = \underset{v_j \in V}{argmax} P(a_1,a_2,\dots a_n|v_j)P(v_j) \quad (7)$$

$P(a_1,a_2,\dots a_n|v_j)P(v_j)$ sulit untuk dihitung, maka akan diasumsikan bahwa setiap kata pada dokumen tidak mempunyai keterkaitan dan dapat dilihat pada persamaan 8.

$$V_{MAP} = \underset{v_j \in V}{argmax} P(v_j) \prod_i P(a_i|v_j) \quad (8)$$

Nilai P(vj) dihitung pada saat *training*, jika $P(v_j)$ adalah probabilitas setiap dokumen terhadap sekumpulan dokumen dan $|docs\,j|$ adalah frekuensi dokumen pada setiap kategori dalam *training*, sedangkan $|training|$ adalah jumlah dokumen dalam contoh yang digunakan untuk *training*. Maka dapat dilihat pada persamaan 9.

$$P(v_j) = \frac{|docsj|}{|training|} \quad (9)$$

Probabilitas kemunculan kata a_i untuk setiap kategori P(a_i $|v_j$) dan n_i adalah frekuensi kemunculan kata ke-i pada setiap kategori, sedangkan N adalah frekuensi ke seluruh kata pada suatu dokumen serta |kosakata| adalah jumlah kata pada dokumen dalam data *training*, maka dihitung pada saat dalam data *training* dapat dilihat pada Persamaan 10.

$$P(a_i|v_j) = \frac{n_i+1}{N+|kosakata|} \quad (10)$$

Metode NBC memiliki dua tahap dalam proses klasifikasi teks, yaitu tahap pelatihan dan tahap klasifikasi [16].

1) Tahap Pelatihan

Pada tahap pelatihan dilakukan proses analisis terhadap sampel dokumen berupa pemilihan *vocabulary*, yaitu kata yang mungkin muncul dalam koleksi dokumen sampel yang sedapat mungkin dapat menjadi representasi dokumen. Selanjutnya adalah penentuan probabilitas *prior* bagi tiap kategori berdasarkan sampel dokumen. Langkah 1: Bentuk *vocabulary* pada setiap dokumen data latih. Langkah 2: Hitung probabilitas pada setiap kategori $P(v_j)$. Langkah 3: Tentukan frekuensi setiap kata (a_i) pada setiap kategori $P(a_i|v_j)$.

2) Tahap Klasifikasi

Pada tahap klasifikasi ditentukan nilai kategori dari suatu dokumen berdasarkan kata yang muncul dalam dokumen yang diklasifikasi. Sedangkan untuk menentukan klasifikasi pada data uji, digunakan persamaan 9. Langkah 1: Hitung $P(v_j) \prod P(a_i|v_j)$ untuk setiap kategori. Langkah 2: Tentukan kategori dengan nilai $P(v_j) \prod P(a_i|v_j)$ maksimal.

F. Ketetapan Klasifikasi

Ketetapan klasifikasi dilakukan untuk melihat performa klasifikasi yang telah dilakukan. Dalam mengukur ketepatan klasifikasi, perlu diketahui jumlah pada setiap kelas prediksi dan kelas aktual terdiri dari *True Positive* (TP), *True Negative* (TN), *True Neutral* (TNe), *False Positive* (FP), *False Negative* (FN) dan *False Neutral* (FNe). Nilai *True Negative* (TN) adalah jumlah data negatif yang terdeteksi dengan benar, sedangkan *False Positive* (FP) adalah jumlah data positif namun terdeteksi salah. Sementara itu, *True Positive* (TP) adalah jumlah data positif yang terdeteksi dengan benar, sedangkan *False Negative* (FN) adalah jumlah data negatif namun terdeteksi salah. *True Neutral* (TNe) adalah jumlah data netral yang terdeteksi dengan benar, sedangkan *False Neutral* (FNe) adalah jumlah data netral namun terdeteksi salah. Berikut klasifikasi *multi-class* dengan tabel *confusion matrix* dapat dilihat pada Tabel 2.

Hasil tabel *confusion matrix* dapat digunakan untuk mencari nilai akurasi*, precission, recall, specificity* dan F1 *score.* Akurasi merupakan persentase dokumen yang teridentifikasi secara tepat dari total dokumen dalam proses klasifikasi sebuah dokumen yang mempunyai data yang *balanced* pada tiap kategorinya, sedangkan *precission* adalah perbandingan jumlah dokumen teks relevan yang terambil diantara semua dokumen teks yang terpilih oleh sistem. Sementara itu, *recall (sensitivity)* adalah perbandingan jumlah dokumen teks relevan yang terambil dari keseluruhan

Tabel 2.
Tabel *Confusion Matrix*

	Keterangan	Hasil Prediksi		
		Negatif	Netral	Positif
Hasil Aktual	Negatif	TN (*True Negative*)	FN (*False Negative*)	FN (*False Negative*)
	Netral	FNe (*False Neutral*)	TNe (*True Neutral*)	FNe (*False Neutral*)
	Positif	FP (*False Positive*)	FP (*False Positive*)	TP (*True Positive*)

Tabel 3.
Variabel Penelitian

Variabel	Keterangan	Skala
X_1	Jumlah Dokumen pada Pelaporan Masyarakat Surabaya	Rasio
X_2	Frekuensi Kemunculan Kata pada Pelaporan Masyarakat Surabaya	Rasio
X_3	Kelas Kategori Sentimen pada Pelaporan Masyarakat Surabaya Terdiri dari Positif, Negatif dan Netral	Nominal

Tabel 4.
Struktur Data

Komentar (X_1)	$X_{2;1}$	$X_{2;2}$	$X_{2;3}$	⋯	$X_{2;n}$	X_3
1	$X_{2;1;1}$	$X_{2;2;1}$	$X_{2;3;1}$	⋯	$X_{2;n;1}$	$X_{3;1}$
2	$X_{2;1;2}$	$X_{2;2;2}$	$X_{2;3;2}$	⋯	$X_{2;n;2}$	$X_{3;2}$
3	$X_{2;1;3}$	$X_{2;2;3}$	$X_{2;3;3}$	⋯	$X_{2;n;3}$	$X_{3;3}$
⋮	⋮	⋮	⋮	⋱	⋮	⋮
n	$X_{2;1;n}$	$X_{2;2;n}$	$X_{2;3;n}$	⋯	$X_{2;n;n}$	$X_{3;n}$

dokumen teks relevan yang ada pada dokumen, sedangkan *specificity* merupakan kebenaran memprediksi negatif dibandingkan dengan keseluruhan data negatif dan F1 *score* merupakan perbandingan rata-rata *precission* dan *recall* yang dibobotkan [17]. Berikut merupakan rumus dalam menghitung akurasi, *precission, recall, specificity dan* F1 *score.*

$$Akurasi = \frac{\frac{TP+TN+TNe}{TP+TN+TNe+FP+FN+FNe}}{l} \times 100\% \quad (11)$$

Rumus untuk nilai *precission* berdasarkan kelas kategori bisa dilihat pada persamaan 12 sampai 14.

$$Precission_{Positif} = \frac{TP}{(FP+TP)} \times 100\% \quad (12)$$

$$Precission_{Negatif} = \frac{TN}{(FN+TN)} \times 100\% \quad (13)$$

$$Precission_{Netral} = \frac{TNe}{(FNe+TNe)} \times 100\% \quad (14)$$

Rumus untuk nilai *recall* berdasarkan kelas kategori bisa dilihat pada persamaan 15 sampai 17.

$$Recall_{Positif} = \frac{TP}{(FN+FNe+TP)} \times 100\% \quad (15)$$

$$Recall_{Negatif} = \frac{TN}{(FP+FNe+TN)} \times 100\% \quad (16)$$

$$Recall_{Netral} = \frac{TNe}{(FN+FP+TNe)} \times 100\% \quad (17)$$

Rumus untuk nilai *specificity* berdasarkan kelas kategori bisa dilihat pada persamaan 18 sampai 21.

$$Specificity_{Positif} = \frac{TN+TNe+FN+FNe}{FP+FP+TN+TNe+FN+FNe} \times 100\% \quad (18)$$

$$Specificity_{Negatif} = \frac{TNe+TP+FNe+FP}{FN+FN+TNe+TP+FNe+FP} \times 100\% \quad (19)$$

$$Specificity_{Netral} = \frac{TN+TP+FN+FP}{FNe+FNe+TN+TP+FN+FP} \times 100\% \quad (20)$$

Setelah dilakukan perhitungan *recall* berdasarkan kelas kategori maka hasil akan dijumlahkan dan dibagi sebanyak kelas kategorinya dan seterusnya untuk *precission* dan *sensitivity*, selanjutnya hasil *recall* dan *precission* akan digunakan untuk memperoleh hasil *F1Score* bisa dilihat pada persamaan 21.

$$F1Score = 2 \times \frac{(Recall \times Precission)}{(Recall + Precission)} \quad (21)$$

Sedangkan untuk data *imbalanced*, ketepatan klasifikasi yang digunakan adalah *G-Mean*. *G-Mean* atau *geometric mean* merupakan nilai rata-rata *geometric* yang diperoleh dengan mengalikan nilai *recall* dan *specificity*. Berikut merupakan rumus untuk mendapatkan nilai *G-Mean*. Selain *G- Mean* juga dapat digunakan nilai *Area Under Curve* (AUC). AUC merupakan indikator performansi kurva ROC (*Receiver Operating Characteristic*) yang dapat meringkas kinerja sebuah klasifikasi menjadi satu nilai, dikatakan ketetapan klasifikasi tinggi jika nilai 80% keatas, sedangkan ketetapan klasifikasi cukup tinggi jika nilai 50% ketas dan ketetapan klasifikasi rendah jika nilai 50% kebawah [18].

$$G - mean = \sqrt{Sensitivity \times Specificity} \quad (22)$$

$$AUC = \frac{1}{2}(Sensitivity + Specificity) \quad (23)$$

III. METODOLOGI PENELITIAN

A. Variabel Penelitian

Variabel Penelitian yang digunakan dalam penelitian ini adalah jumlah data teks pelaporan masyarakat Surabaya dalam rentang waktu Januari 2020 hingga Desember 2020. Frekuensi kemunculan kata pada data pelaporan masyarakat Surabaya sebanyak 8899 kata yang telah muncul dengan kata yang dipakai sebanyak 39 kata. Kelas kategori sentimen positif, negatif dan netral pada pelaporan masyarakat masing-masing berjumlah 2094, 1037 dan 606. Berikut variabel penelitian yang akan digunakan pada penelitian ini dapat dilihat pada Tabel 3.

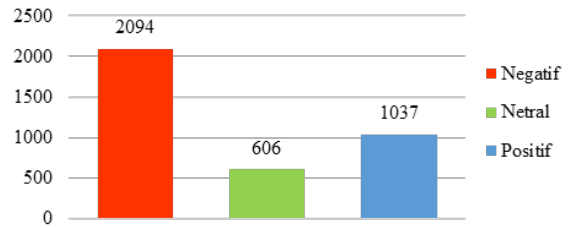
Stuktur data pada penelitian ini bisa didapatkan sebelumnya dengan cara melakukan tahap analisis sentimen, *text pre-processing*, *Term Frequency* (TF), struktur data dapat dilihat pada Tabel 4.

Keterangan :

1. X_1 : Jumlah dokumen pada pelaporan masyarakat Surabaya.
2. $X_{2;1}$, $X_{2;2}$, $X_{2;3}$, ...,$X_{2;n}$: Kata-kata yang muncul pada pelaporan masyarakat melalui tahap *text pre-processing* dan *Term Frequency* (TF).
3. $X_{2;i;n}$: Frekuensi kemunculan kata ke-i pada ulasan pelaporan masyarakat ke-n.
4. X_3 : Kelas kategori sentimen pada pelaporan masyarakat.

B. Metode Pengambilan Sampel

Pengambilan data pada penelitian ini berasal dari data sekunder yaitu teks pelaporan masyarakat Surabaya sebanyak 3737 dokumen teks, yang diperoleh dari situs media *center*



Gambar 2. Pengelompokan Kelas Kategori Menggunakan *Bar Chart*.

Tabel 5.
Stuktur Data Pelaporan Masyarakat Surabaya

Komentar	Jalan	Rumah	Lokasi	⋯	BLT	Kelas Kategori
1	1	2	0	⋯	0	Negatif
2	1	3	0	⋯	0	Negatif
3	2	1	0	⋯	0	Netral
4	1	1	0	⋯	0	Negatif
5	3	0	0	⋯	0	Positif
⋮	⋮	⋮	⋮	⋱	⋮	⋮
3737	0	0	0	⋯	0	Netral

Tabel 6.
Perhitungan Frekuensi Kemunculan Kata Tertinggi

No	Kata	Jumlah
1	Jalan	766
2	Rumah	633
3	Surat	548
4	Kartu	524
5	KTP	458
6	Covid	418
7	Keluarga	409
8	Daftar	376
9	Admin	301
10	Pemkot	280

kota Surabaya, data diambil pada tanggal 2 Maret 2021 di Dinas Komunikasi dan Informatika kota Surabaya.

C. *Langkah Analisis*

Langkah-langkah dalam pelaksanaan penelitian ini adalah sebagai berikut.

1. Mengumpulkan data pelaporan masyarakat kota Surabaya serta melakukan analisis sentimen data pelaporan masyarakat kota Surabaya meliputi: (a)Melakukan *text pre-processing* data yaitu: *case folding, tokenizing,* menghapus *URL,* menghapus *mention,* menghapus *hastag,* menghapus tanda baca, menghapus angka dan *spelling normalization.* (b)Selanjutnya, melakukan analisis sentimen untuk memberikan kelas kategori yaitu: positif, negatif dan netral dan pengelompokan berdasarkan kelas kategori yang sama.
2. Melanjutkan tahapan *text pre-processing* data sebelumnya yaitu *remove*, *filltering, stemming* dan menghapus spasi yang berlebihan, sehingga didapatkan data teks bersih pelaporan masyarakat kota Surabaya.
3. Melakukan perhitungan frekuensi kemunculan kata menggunakan *Term Frequency* (TF) dan pembobotan kata menggunakan TF-IDF (*Term Frequency-Inverse Document Frequency*) dari data pelaporan masyarakat kota Surabaya.
4. Melakukan klasifikasi dari hasil analisis sentimen data pelaporan masyarakat kota Surabaya meliputi: (a)Pembagian data menjadi data *testing* dan data *training* dengan perbandingan proporsi 20%:80% dengan menggunakan 10-*Fold Cross Validation.* (b)Melakukan klasifikasi menggunakan *Naive Bayes Classifier* dari masing-masing perbandingan proporsi.
5. Melakukan pengukuran ketetapan klasifikasi pada data pelaporan masyarakat kota Surabaya menggunakan akurasi, *precission, recall, specificity,* F1 *score, G-mean* dan luas AUC.
6. Menarik kesimpulan.

IV. ANALISIS DAN PEMBAHASAN

A. *Analisis Sentimen Pelaporan Masyarakat Surabaya Tahun 2020*

Analisis sentimen pada pelaporan masyarakat Surabaya digunakan untuk mengidentifikasi dan mengelompokkan pelaporan masyarakat apakah bersifat positif, negatif dan netral, sebelum dilakukan analisis sentimen akan dilakukan terlebih dahulu *text pre-processing* meliputi *case folding, remove* dan *spelling normalization.* Berikut hasil analisis sentimen dari pelaporan masyarakat Surabaya tahun 2020 dapat ditampilkan dalam grafik *bar chart* pada Gambar 2.

Gambar 2 menunjukkan bahwa pengelompokan kelas kategori dari hasil analisis sentimen pelaporan masyarakat Surabaya tahun 2020, didapatkan kelas kategori negatif sebesar 56,03% dan kelas kategori netral sebesar 16,22% serta kelas kategori positif sebesar 27,75%, dapat disimpulkan bahwa pelaporan masyarakat Surabaya tahun 2020 memiliki jumlah tertinggi dengan komentar atau pelaporan yang bersifat negatif, sehingga pelaporan masyarakat Surabaya tahun 2020 perlu ditindak lanjuti dan dievaluasi.

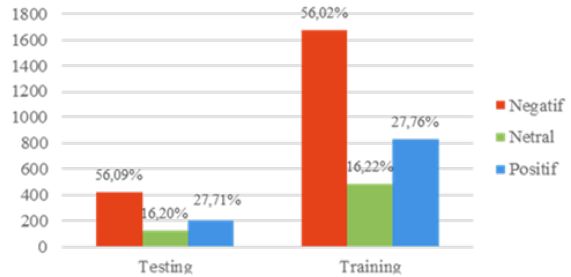
B. *Klasifikasi Pelaporan Masyarakat Surabaya Tahun 2020 Menggunakan Naïve Bayes Classifier*

Klasifikasi pelaporan masyarakat Surabaya tahun 2020 menggunakan *Naïve Bayes Classifier* akan dilakukan terlebih dahulu yaitu *text pre-processing lanjutan meliputi: remove, filltering, stemming* dan *whitespase,* perhitungan frekuensi kemunculan kata (TF), pembobotan kata (TF-IDF) dan pembagian data (*10-fold cross validation*). Struktur data yang didapatkan ada pada Tabel 5.

Tabel 5 menunjukkan bahwa data pelaporan masyarakat yang telah berbentuk (*document term matrix),* dapat dilakukan perhitungan jumlah kata yang selanjutnya akan menjadi jumlah variabel dari data pelaporan masyarakat. Struktur data pelaporan masyarakat memiliki jumlah kata sebanyak 39 kata yang mewakili isi dari pelaporan masyarakat. Setelah terbentuk struktur data yang diinginkan, dilakukan perhitungan frekuensi kemunculan kata pada pelaporan masyarakat Surabaya tahun 2020.

Perhitungan frekuensi kemunculan kata (TF) yang tertinggi dan pembobotan kata (TF-IDF) pada pelaporan masyarakat Surabaya dapat ditunjukkan pada Tabel 6 dan Tabel 7.

Tabel 6 menunjukkan bahwa frekuensi kemunculan kata yang lebih besar dari 250 kemunculan kata, sehingga didapatkan 10 kata tertinggi dari data pelaporan masyarakat Surabaya tahun 2020 yakni “jalan”, “rumah”, “surat”, “kartu”, ”ktp”, “covid”, “keluarga”, “daftar”, “admin” dan “pemkot” dengan frekuensi keseluruhan kemunculan kata sebanyak 8899 kata yang telah muncul. Berikut pembobotan kata (TF-IDF) dengan dihitung menggunakan persamaan (2)



Gambar 3. *Bar Chart* Pembagian Data.

Tabel 7.
Pembobotan Kata

Komentar	Hasil Pembobotan Kata
1	7,7359
2	7,7429
3	7,0866
4	7,4857
5	8,1620
6	8,5193
7	8,0564
8	8,0308
9	8,7336
10	9,3320
⋮	⋮
3737	8,4967

dan (3) dari data pelaporan masyarakat Surabaya tahun 2020 dapat ditunjukkan pada Tabel 7.

Tabel 7 menunjukkan bahwa hasil pembobotan kata (TF-IDF) dari data pelaporan masyarakat Surabaya tahun 2020 yang berguna untuk merepresentasikan nilai numerik dokumen sehingga kemudian dapat dihitung kedekatan antar dokumen dari pelaporan masyarakat.

Langkah pertama dalam mengklasifikasikan data pelaporan masyarakat Surabaya tahun 2020 adalah melatih model menggunakan data *training*. Data *training* yang telah dilakukan *text pre-processing* digunakan untuk melatih model menggunakan *software* Rstudio. Model yang telah dilatih dengan data *training* kemudian digunakan untuk mengklasifikasi data *testing* ke dalam tiga kelas kategori sentimen yaitu positif, negatif dan netral. Pembagian data *training* dan data *testing* berdasarkan metode 10-*fold cross validation* dengan proporsi pembagian yaitu 20% : 80%. Berikut *bar chart* guna mengetahui pembagian data dengan frekuensi antar kelas ketegori dapat dilihat pada Gambar 3.

Gambar 3 menunjukkan bahwa pembagian data *training* dan data *testing* yang cenderung *imbalance* antara kelas kategori sentimen negatif, posistif maupun netral. Keadaan *imbalance* pada kelas kategori data tersebut akan berpengaruh pada perhitungan ketepatan klasifikasi. Ukuran ketepatan klasifikasi akurasi tidak cukup sesuai untuk data *imbalance* sehingga akan dilakukan pengukuran ketepatan klasifikasi dengan *G-mean* dan AUC.

Klasifikasi dengan metode *Naïve Bayes Classifier* menghasilkan probabilitas yang digunakan untuk menentukan apakah pelaporan masyarakat masuk ke dalam kelas kategori sentimen positif, negatif atau netral. Probabilitas tersebut diperoleh dengan persamaan (9). Berikut hasil nilai probabilitas klasifikasi NBC ditunjukkan pada Tabel 8.

Tabel 8 menunjukkan bahwa nilai probabilitas pada data pelaporan masyarakat Surabaya tahun 2020 pada setiap kelas kategori, didapatkan bahwa memiliki pelaporan atau komentar dengan sifat sentimen yang negatif, karena dari

Tabel 8.
Probabilitas Klasifikasi NBC

Data	Kelas Kategori	Probabilitas
Pelaporan Masyarakat	Negatif	0,5602
	Netral	0,1622
	Positif	0,2776

Tabel 9.
Confusion Matrix Data Pelaporan Masyarakat

Data	Kelas Aktual	Kelas Prediksi		
		Negatif	Netral	Positif
Training	Negatif	456	69	145
	Netral	995	372	516
	Positif	224	44	169
Testing	Negatif	122	15	47
	Netral	240	93	128
	Positif	57	13	32

Tabel 10.
Hasil Ketetapan Klasifikasi NBC

Data	*G-Mean*	AUC
Training	0,5397	0,5586
Testing	0,5314	0,5512

nilai probabilitas klasifikasi yang tertinggi dengan menunjukkan sentimen negatif sebesar 0,5602. Dimana nilai probabilitas tinggi menunjukkan banyak kelas kategori tersebut yang ada pada pelaporan masyarakat Surabaya.

Dari probabilitas tersebut didapat kelas kategori prediksi dari setiap pelaporan masyarakat. Langkah selanjutnya adalah mengukur ketepatan klasifikasi model dari setiap proporsi pembagian data. Pengukuran ketepatan klasifikasi dilakukan dengan membentuk tabel *confusion matrix* berdasarkan hasil prediksi. Berikut tabel *confusion matrix* kelas kategori aktual dan prediksi pada pelaporan masyarakat Surabaya dapat ditunjukkan pada Tabel 9.

Tabel 9 menunjukkan bahwa pada data *training* dengan kelas kategori negatif yang diprediksi dengan benar sebesar 456 dari jumlah 670, sedangkan kelas kategori netral yang diprediksi dengan benar sebesar 372 dari jumlah 1883 dan pada kelas kategori positif yang diprediksi dengan benar sebesar 169 dari jumlah 437. pada data *testing* dengan kelas kategori negatif yang diprediksi dengan benar sebesar 122 dari jumlah 184, sedangkan kelas kategori netral yang diprediksi dengan benar sebesar 240 dari jumlah 461 dan pada kelas kategori positif yang diprediksi dengan benar sebesar 32 dari jumlah 102. Berikut merupakan hasil pengukuran ketepatan klasifikasi dari setiap proporsi pembagian data menggunakan algoritma *Naïve Bayes Classifier* dapat ditunjukkan pada Tabel 10.

Tabel 10 menunjukkan bahwa ukuran ketepatan klasifikasi data pelaporan masyarakat karena data *imbalance* maka ukuran ketetapan klasifikasi yang digunakan adalah *G-Mean* dan AUC. Didapatkan bahwa hasil ketetapan klasifikasi pada data *training* dengan nilai *G-Mean* sebesar 53,97% dan nilai AUC sebesar 55,86%, sedangkan pada data *testing* dengan nilai *G-Mean* sebesar 53,14% dan nilai AUC sebesar 55,12%. Sehingga klasifikasi data pelaporan masyarakat Surabaya tahun 2020 pada data *training* maupun data *testing* menghasilkan ketepatan klasifikasi yang cukup tinggi atau cukup baik.

V. KESIMPULAN/RINGKASAN

Kesimpulan yang diperoleh dari hasil analisis dan pembahasan adalah analisis sentimen dan pengelompokan

kelas kategori dari data pelaporan masyarakat Surabaya tahun 2020, didapatkan bersifat negatif 56,03%, bersifat positif 27,75% dan bersifat netral 16,22% dengan tingkat klasifikasi yang cukup tinggi karena nilai nilai sensitifitas dan *specificity* yang diwakili oleh *G-Mean* dan AUC mencapai 53,14% dan 55,12% pada data *testing*.

UCAPAN TERIMA KASIH

Penulis mengucapkan terima kasih kepada Dinas Komunikasi dan Informatika Kota Surabaya yang telah memberikan kesempatan dan membantu dalam kelancaran pengerjaan makalah ini.

DAFTAR PUSTAKA

[1] A. Yusuf and T. Priambadha, "Support vector machines yang didukung K-Means clustering dalam klasifikasi dokumen," *J. Ilm. Teknol. Inf.*, vol. 11, no. 1, pp. 15–18, 2013, doi: 10.12962/j24068535.v11i1.a15.

[2] P. Nomleni, "Sentiment Analysis Menggunakan Support Vector Machine (SVM)," Institut Teknologi Sepuluh Nopember, 2015,

[3] E. . Pamungkas and D. G. . Putri, "An experimental study of lexicon-based sentiment analysis on Bahasa Indonesia," *Int. Annu. Eng. Semin.*, vol. 6, no. 1, pp. 28–31, 2016, doi: 10.1109/INAES.2016.7821901.

[4] J. Han, M. Kamber, and J. Pei, *Data mining: concepts and techniques*, 3rd ed., vol. 5, no. 4. Waltham, USA: Morgan Kaufmann Publisher imprint of Elsevier, 2011, ISBN: 978-0-12-381479-1.

[5] W. Wibowo, N. P. Sari, R. N. Wilantari, and S. A. Rahman, "Association rule mining method for the identification of internet use," *J. Phys. Conf. Ser.*, vol. 1874, no. 1, 2020, doi: 10.1088/1742-6596/1874/1/012009.

[6] E. B. Jayanti, "Pengelompokan dan Klasifikasi Laporan Masyarakat di Situs Media Center Kota Surabaya Menggunkan Metode K-Means Clustering dan Support Vector Machine," Institut Teknologi Sepuluh Nopember, 2017.

[7] D. Y. Praptiwi, "Analisis Sentimen Online Review Pengguna E-Commerce Menggunakan Metode Support Vector Machine dan Maximum Entropy (Studi Kasus: Review Bukalapak pada Google Play)," Universitas Islam Indonesia Yogyakarta, 2018.

[8] E. Nugroho, "Perancangan Sistem Deteksi Plagiarisme Dokumen Teks dengan Menggunakan Algoritma Rabin-Karp," Universitas Brawijaya, 2011, http://repository.ub.ac.id/id/eprint/152634.

[9] L. Agusta, "Perbandingan algoritma stemming Porter dengan algoritma Nazief & Adriani untuk stemming dokumen teks Bahasa Indonesia," *Konf. Nas. Sist. dan Inform.*, vol. 2009, no. 1, pp. 196–201, 2009, ISSN: KNS&I09-036.

[10] S. Robertson, "Understanding inverse document frequency: on theoretical arguments for IDF," *J. Doc.*, vol. 60, no. 5, pp. 503–520, 2004, doi: https://doi.org/10.1108/00220410410560582.

[11] C. Megawati, "Analisis Aspirasi dan Pengaduan di Situs Lapor dengan Menggunakan Text Mining," Universitas Indonesia, 2015, http://lib.ui.ac.id/detail?id=20411108&lokasi=lokal.

[12] E. Gokgoz and A. Subasi, "Comparison of decision tree algorithms for EMG signal classification using DWT," *Biomed. Signal Process. Control*, vol. 18, no. 1, pp. 138–144, 2015, doi: https://doi.org/10.1016/j.bspc.2014.12.005.

[13] B. Santosa, *Data Mining Teknik Pemanfaatan Data Untuk Keperluan Bisnis*. Yogyakarta: Graha Ilmu, 2007, ISBN: 978-979-756-224-3.

[14] Falahah and D. D. Adriadi Nur, "Pengembangan aplikasi sentiment analysis menggunakan metode Naive Bayes (Studi kasus sentiment analysis dari media twitter)," *Semin. Nas. Sist. Inf. Indones.*, vol. 1, no. 1, pp. 335–340, 2015.

[15] A. F. Rachimawan and B. S. Utama, "Ads filtering mengunakan jaringan syaraf tiruan perceptron, naïve bayes classifier, dan regresi logistik," *J. Sains dan Seni ITS*, vol. 5, no. 1, p. D-83-D-89, 2016, ISSN: 2337-3520.

[16] A. Indriani, "Klasifikasi data forum dengan menggunakan Metode Naïve Bayes Classifier," *Semin. Nas. Apl. Teknol. Inf.*, vol. 1, no. 1, 2014, ISSN: 1907-5022.

[17] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Inf. Process. Manag.*, vol. 45, no. 4, pp. 427–437, 2009, doi: https://doi.org/10.1016/j.ipm.2009.03.002.

[18] M. Bekkar, H. Djemaa, and T. Alitouche, "Evaluation measures for models assessment over Imbalanced Data Sets," *J. Inf. Eng. Appl.*, vol. 3, no. 1, pp. 27–38, 2013, ISSN: 2225-0506.