e-ISSN 2962-6129
Seminar Nasional Mahasiswa Ilmu Komputer dan Aplikasinya (SENAMIKA)
Jakarta-Indonesia, 20 Agustus 2022

Analisis Sentimen Terhadap Layanan Transjakarta Pada Media Sosial *Instagram* Menggunakan *Naïve Bayes* dan Seleksi Fitur *Information Gain*

Ivtytah Ein [1], Iin Ernawati [2], Yuni Widiastiwi[3]
S1 Informatika / Fakultas Ilmu Komputer
Program Studi Informatika, Universitas Pembangunan Nasional Veteran Jakarta
Jl. RS. Fatmawati Raya, Pd. Labu, Kec. Cilandak, Kota Depok, Jawa Barat 12450
ivtythein@upnvj.ac.id[1], iin_ernawati@yahoo.com[2], widiastiwi@yahoo.com[3]

Abstrak. PT. Transportasi Jakarta (Transjakarta) menanggapi kebijakan Pemerintah Provinsi DKI Jakarta mengenai pembatasan pada moda transportasi. Pada tanggal 22 Oktober 2021, Transjakarta mengumumkan layanan akan kembali normal. Walaupun layanan kembali beroperasional dengan normal masih banyak pengguna transjakarta yang menuangkan kritik dan opini yang berkaitan dengan layanan pada akun *instagram* transjakarta. Penelitian ini bertujuan untuk membangun model klasifikasi sentimen dengan menggunakan metode *naïve bayes* dan seleksi fitur *information gain* terhadap opini masyarakat terkait pelayanan transjakarta di media sosial *instagram*. Data komentar dibagi menjadi kelas positif dan negatif selanjutnya data tersebut akan dilakukan *preprocessing*, pembobotan TF-IDF, seleksi fitur, dan pembagian data sebesar 70% data latih dan 30% data uji. Hasil evaluasi untuk model klasifikasi *naïve bayes* memperoleh akurasi sebesar 81,42%, *recall* sebesar 69,64%, *precision* sebesar 63,93%, dan *specificity* sebesar 85,71%. Sedangkan hasil model klasifikasi *naïve bayes* dengan *information gain* memperoleh akurasi sebesar 86,66%, *recall* sebesar 71,42%, *precision* sebesar 76,92%, dan *specificity* sebesar 92,20%.

Kata Kunci: Transjakarta, *Instagram*, *Naïve Bayes*, *Information Gain*.

1 Pendahuluan

Pemerintah Provinsi DKI Jakarta membuat kebijakan baru mengenai pembatasan pada moda transportasi umum salah satunya pada transjakarta yang bertujuan untuk mencegah penyebaran Covid-19[1]. PT. Transportasi Jakarta menanggapi kebijakan tersebut dengan membatasi rute pelayanan, jam operasional, kapasitas angkut, dan jumlah armada bus [2]. Per tanggal 22 Oktober 2022 layanan transjakarta akan beroperasi dengan normal, walaupun layanan telah kembali normal masih banyak pengguna transjakarta yang merasa kesulitan dan mengalami beberapa kendala terhadap layanan yang diberikan dikarenakannya keridaksesuaian jam operasional, antrean yang panjang, pengguna tidak memenuhi protokol kesehan, dan keluhan lainnya. Pengguna transjakarta tersebut mulai menuangkan kritik dan sarannya melalui *instagram* dengan menuliskan komentar pada unggahan di akun transjakarta.

Analisis sentimen dapat memberikan informasi mengenai gambaran sentimen positif dan negatif masyarakat mengenai suatu topik atau permasalahan tertentu. Analisis sentimen dilakukan dengan membangun model klasifikasi dengan menggunakan metode tertentu yang mana dalam prosesnya menggunakan data dalam bentuk teks yang memiliki fokus pada opini ataupun pendapat yang mengandung sentimen positif dan negatif yang nantinya akan dilakukan *preprocessing* untuk memastikan bahwa data telah siap untuk mengurangi terjadinya permasalahan pada data. Setelah ini dilakukan pembobotan pada data dan dilakukan proses klasifikasi. Pada penelitian ini penulis menggunakan metode *naïve bayes* untuk proses klasifikasi. Selain itu penulis juga membuat model klasifikasi dengan metode *Naïve Bayes* dan seleksi fitur. Seleksi fitur yang digunakan pada penelitian yaitu *Information Gain* untuk mendeteksi tingkat kepentingan sebuah atribut dalam suatu kategori.

Dengan menggunakan metode *Naïve Bayes* dalam proses pengklasifikasian dan *Information Gain* untuk seleksi fitur diharapkan dapat memberikan perbandingan mengenai performa model klasifikasi yang baik sehingga dapat memberikan suatu informasi mengenai analisis sentimen terhadap layanan transjakarta yang di dapat melalui komentar di media sosial instagram pada akun transjakarta.

2 Landasan Teori

2.1 *Fleiss Kappa*

Fleiss Kappa adalah metode umum untuk menentukan antar penilai reliabilitas yang merupakan generalisasi dari statistik *pi Scott*. *Fleiss Kappa* memiliki kekuatan untuk menilai konsensus antar penilai yang diantaranya lebih dari dua penilai [3].

$$K = \frac{P_0 - P_e}{1 - P_e} \quad (1)$$

Keterangan :
K : Koefisien dari nilai kesepakatan dimana 0 untuk persetujuan secara kebetulan, 1 untuk persetujuan total
P_0 : Proporsi frekuensi pengamatan
P_e : Peluang kesepakatan antar pengamat

Untuk mendapatkan nilai P_0 dan P_e dapat dilakukan dengan menggunakan persamaan sebagai berikut :

$$\overline{P_0} = \frac{1}{N}\left(\sum_{i=1}^{N} P_i\right) \quad (2)$$

$$\overline{P_e} = \sum_{j=1}^{k} {P_j}^2 \quad (3)$$

Untuk mendapatkan nilai P_i dan P_j dapat dilakukan dengan menggunakan persamaan sebagai berikut :

$$P_i = \frac{1}{n(n-1)}\left(\sum_{j=1}^{k} n_{ij}^2 - n_{ij}\right) \quad (4)$$

$$P_j = \frac{n.j}{Nn} = \frac{1}{Nn}\sum_{i=1}^{N} n_{ij} \quad (5)$$

Keterangan :
$n.j = \sum_{i=1}^{N} n_{ij}$: Total jumlah label untuk kategori
N : Jumlah data
n : Jumlah annotator

Adapun tabel skala *kappa value* yang dimuat dalam sebagai berikut :

Tabel 1. Tabel skala *kappa value*.

K	*Analysis*
<0	*Poor*
0.00-0.20	*Slight*
0.21-0.40	*Fair*
0.41-0.60	*Moderate*
0.61-0.80	*Substantial*
0.81-1.0	*Almost Perfect*

Berdasarkan tabel diatas untuk dapat melanjutkan ke tahapan berikutnya hasil perhitungan kappa berada pada skala 0.41 - 0.60 atau bisa dikatakan kesepakatan tersebut "*moderate*" atau sedang.

2.2 *Preprocessing*

Preprocessing merupakan tahapan yang dilakukan sebelum melakukan tahapan *data mining*. Tahapan ini bertujuan untuk memastikan bahwa data siap untuk dianalisis. Pada tahapan *preprocessing* dilakukan guna mengurangi permasalahan seperti terjadinya jumlah populasi yang besar, banyaknya anomali data, dan

sebagainya [4]. *Preprocessing* merupakan tahapan untuk membersihkan data dan melakukan konversi teks agar memiliki standar yang sesuai dengan kebutuhan [5].

a. ***Case Folding***, adalah proses untuk mengubah bentuk huruf kapital dengan menyetarakan menjadi huruf kecil [6].
b. ***Data Cleaning***, adalah proses untuk melakukan pembersihan pada tanda baca seperti koma (,), titik (.) ataupun tanda baca lainnya dihilangkan untuk mengurangi adanya *noise* pada data [6].
c. ***Normalization***, adalah proses untuk menormalisasikan bahasa. Pada proses ini apabila terdapat kata yang tidak baku akan dinormalisasikan kedalam bentuk kata baku yang berdasarkan kaidah dari Kamus Besar Bahasa Indonesia (KBBI) [6].
d. ***Stopword Removal***, adalah proses untuk menghilangkan kata umum yang tidak memiliki arti penting apabila digunakan, hal ini dilakukan agar dapat mengurangi banyaknya data yang tidak perlu [6].
e. ***Stemming***, adalah proses untuk mencari kata dasar atau stem yang terdapat pada kata yang dihasilkan pada proses *stopword removal*. *Stemming* sendiri memiliki aturan dengan pendekatan kamus [6].
f. ***Tokenizing***, adalah proses untuk memangkas dokumen menjadi bagian-bagian kecil seperti bab, sub-bab, paragraf, kalimat, dan kata (token) [6].

2.3 Pembobotan TF-IDF

Term Frequency–Inverse Document Frequency (TF–IDF) adalah metode yang dilakukan untuk melakukan pembobotan pada suatu kata. Pembobotan TF–IDF bertujuan untuk melihat seberapa banyak jumlah kemunculan kata pada suatu dokumen, yang mana hal ini berpengaruh pada nilai kontribusi [7]. Apabila TF memiliki nilai yang tinggi maka nilai TF–IDF akan naik yang bertanda bahwa kata tersebut penting. Jika nilai DF tinggi maka nilai TF–IDF akan rendah.

$$TF\ IDF = TF\ \times IDF \tag{6}$$

$$TF\ IDF = TF\ \times \log\frac{|D|}{DF} \tag{7}$$

Keterangan dari rumus TF–IDF sebagai berikut
TF : *Term frequency*, jumlah kata yang muncul
DF : *Document frequency*, jumlah dokumen dimana kata tersebut muncul
D : Jumlah total dokumen

2.4 *Naïve Bayes*

Naïve Bayes merupakan salah satu algoritma klasifikasi probabilistik sederhana berdasarkan teorema *Bayes*. Cara kerja dari *Naïve Bayes* sendiri adalah untuk memprediksi probabilitas di masa depan berdasarkan dengan pengalaman yang sebelumnya. *Naïve Bayes* memungkinkan tiap atributnya memiliki kontribusi yang sama terhadap keputusan akhir dan proses komputasi akan jauh lebih efisien jika disandingkan dengan algoritma pengklasifikasian teks lain. *Naïve Bayes* memiliki dua proses penting yaitu proses pelatihan dan proses pengujian. Model dari hasil proses pelatihan berisi sekumpulan konstanta untuk setiap data latih, yang mana model ini akan digunakan untuk menghitung keakuratan model tersebut. *Naïve Bayes* menjumlahkan frekuensi dan kombinasi nilai dari dataset untuk dilakukan perhitungan probabilitas [8]. Dalam penelitian ini menggunakan algoritma *Multinomial Naïve Bayes* dikarenakan memiliki tingkat akurasi yang tinggi, sederhana, dan efektif karena menggunakan ide dasar peluang gabungan kata-kata dan kategori untuk memprediksi peluang kategori pada suatu dokumen [9].

$$C_{MAP} = \underset{c \in V}{\arg max}\, P(p) \prod_{t=i}^{|V|} P(W_i|p) \tag{8}$$

Keterangan :
$P(p)$: Peluang kemunculan dokumen yang ada pada kelas p
$P(W_i|p)$: Peluang kemunculan W_i pada kelas p

Adapun perhitungan untuk probabilitas pada tiap kelas yang dideskripsikan sebagai berikut :

$$P(p) = \frac{dok\ p}{dokumen} \quad (9)$$

Keterangan :
$P(p)$: Peluang kemunculan dokumen pada tiap kelas p
$dok\ p$: Jumlah dokumen untuk tiap kelas p
$dokumen$: Jumlah keseluruhan dokumen

Untuk mendapatkan nilai probabilitas suatu kata dalam suatu kelas menggunakan persamaan berikut :

$$P(W_i|p) = \frac{count(w_{ij,p})}{|p| + |V|} \quad (10)$$

Keterangan :
$P(W_i|p)$: Peluang kemunculan kata W_i pada kelas p
$count(w_{ij,p})$: Jumlah kemunculan kata W_i pada kelas p
$|p|$: Total keseluruhan kata pada kelas p
$|V|$: Total keseluruhan kata (term)

Nilai peluang pada kata yang tidak terjadi akan dibeli nilai 0 (nol), tentunya hal ini tidak diinginkan sehingga membutuhkan *laplace smoothing* dengan menggunakan persamaan sebagai berikut :

$$P(W_i|p) = \frac{count(w_{ij,p}) + 1}{|p| + |V| * 1} = \frac{count(w_{ij,p}) + 1}{|p| + |V|} \quad (11)$$

2.5 *Information Gain*

Information Gain atau IG adalah salah satu dari metode seleksi fitur. Seleksi fitur ini mendeteksi fitur yang banyak memuat informasi yang berdasarkan dengan kelas tertentu dengan melakukan perhitungan pada nilai *entropy* untuk mengukur ketidakpastian kelas dengan menggunakan probabilitas kejadian atau atribut tertentu [10]. Adapun perhitungan untuk mendapatkan nilai *entropy* dengan menggunakan persamaan berikut :

$$Ent\ (X) = \sum_{i=1}^{c} -P_i\ log_2 P_i \quad (12)$$

Keterangan :
c : Jumlah nilai yang ada pada kelas klasifikasi
P_i : Jumlah sampel untuk kelas i

Untuk mendapatkan nilai *entropy* pada tiap fitur dilakukan dengan menggunakan persamaan berikut :

$$Ent_V\ (X) = -\sum_{j=1}^{z} \frac{|X_j|}{|X|}\ Ent\ (X_j) \quad (13)$$

Keterangan :
V : Atribut
$|X|$: Jumlah untuk keseluruhan sampel data
$|X_j|$: Jumlah sampel untuk nilai *j*
z : Nilai yang memungkinkan untuk atribut *V*
$Ent\ (X_j)$: *Entropy* untuk tiap nilai *j*

Untuk mendapatkan nilai *information gain* dilakukan dengan menggunakan persamaan berikut :

$$IG\ (V) = Ent\ (X) - |Ent_V\ (X)| \quad (14)$$

2.6 *Confusion Matrix*

Hasil evaluasi metode klasifikasi dituangkan dalam *Confusion Matrix*, hal ini untuk mendeskripsikan hasil evaluasi dari klasifikasi. Adapun metode yang dilakukan dari terminologi *confusion matrix* seperti akurasi, *recall*, *precision*, dan *specificity* [11].

Tabel 2. Tabel *confusion matrix*.

	Nilai yang sebenarnya

		Positive	*Negative*
Prediksi	*Positive*	*True Positive (TP)*	*False Positive (FP)*
	Negative	*False Negative (FN)*	*True Negative (TN)*

a. Akurasi, merupakan hasil perhitungan semua nilai prediksi yang benar dibagi dengan keseluruhan data.

$$Akurasi = \frac{TP + TN}{TP + TN + FN + FP} \quad (15)$$

b. *Recall*, merupakan hasil perhitungan dari jumlah prediksi positif yang benar dibagi dengan jumlah keseluruhan kelas yang positif.

$$Recall = \frac{TP}{TP + FN} \quad (16)$$

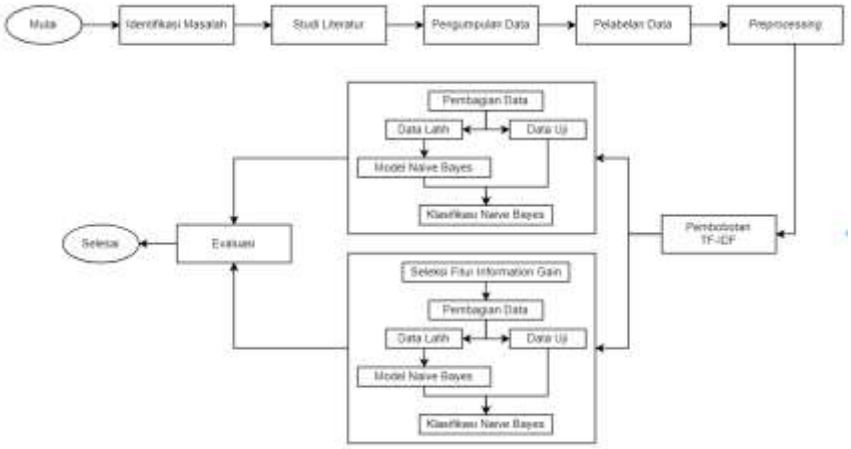
c. *Precision*, dihitung dari jumlah keseluruhan nilai prediksi positif yang benar dibagi dengan jumlah keseluruhan prediksi kelas yang benar.

$$Precision = \frac{TP}{TP + FP} \quad (17)$$

d. *Specificity*, dihitung dari keseluruhan jumlah prediksi salah yang benar dibagi dengan keseluruhan jumlah kelas yang salah.

$$Specificity = \frac{TP}{TP + FP} \quad (18)$$

3 Metodologi Penelitian



Gambar 1. merupakan kerangka pikir pada penelitian ini dimana terdapat beberapa langkah-langkah penelitian yang dijelaskan sebagai berikut :

1. **Identifikasi Masalah**, tahapan untuk mencari suatu permasalahan yang terjadi dilingkungan sekitar sehingga teridentifikasi permasalahan dalam penelitian yang dilakukan.
2. **Studi Literatur**, merupakan tahapan untuk mencari jurnal, buku, prosiding, ataupun literatur lainnya yang berkaitan dengan topik penelitian.
3. **Pengumpulan Data**, tahapan ini dilakukan dengan menggunakan *scraping* dan mengumpulkan data komentar di akun *instagram* transjakarta pada unggahan yang membahas mengenai informatu rute dan layanan.
4. **Pelabelan Data**, tahapan ini menggunakan opini yang menggambarkan kesenangan, kepuasan, dan kegembiraan sebagai sentimen positif. Sedangkan opini yang menggambarkan kekecewaan dan ketidakpuasan sebagai sentimen negatif. Pelabelan data dilakukan secara manual oleh 3 *annotator*. Kesepakatan hasil pelabelan menggunakan metode *fleiss kappa* dengan menggunakan persamaan rumus (1), (2), dan (3) dan hasil pengamatan berpacu pada tabel skala *kappa value* dimana hasil perhitungan dikatakan baik pada skala sedang atau *moderate*.
5. ***Preprocessing***, pada tahapan ini terdapat beberapa langkah sebagai berikut :

a. *Case folding*, pada tahapan ini mengubah huruf kapital (*uppercase*) menjadi huruf kecil (*lowercase*). Adapun contohnya seperti kata "Dream" menjadi "dream".
b. *Data cleaning*, pada tahapan ini melakukan pembersihan pada data yang terdapat tanda baca, angka, emoji, tagar, dan karakter lain. Adapun contoh pada penelitian ini melakukan penghapusan pada tanda baca titik dua (:), koma atas(‘), dan kurung tutup ()).
c. *Normalization,* pada tahapan ini peneliti membuat kamus khusus yang memuat daftar kata yang akan dinormalisasikan. Kamus pada tahapan normalisasi memuat sebanyak 984 daftar kata. Adapun contohnya seperti kata "tj" menjadi "transjakarta".
d. *Stopword Removal*, pada tahapan ini penghapusan pada kata stopword dilakukan degan menggunakan bantuan *library* sastrawi dan kamus modifikasi oleh Oswin RH sebanyak 758 kata. Adapun contohnya seperti kata "menjadi", "akhirnya", "ada".
e. *Stemming,* pada tahapan ini melakukan penghapusan pada kata yang memiliki imbuhan sehingga hanya menyisakan kata dasar dengan menggunakan *library* sastrawi. Adapun contohnya seperti kata "pelayanan" menjadi "layan".
f. *Tokenizing,* pada tahapan ini akan dilakukan pemisahan kalimat menjadi potongan kata-kata tunggal.

6. **Pembobotan TF-IDF,** tahapan ini akan dilakukan pembobotan pada kata dengan melakukan perhitungan berdasarkan pada persamaan (6) dan (7).
7. **Seleksi Fitur *Information Gain***, pada tahapan ini terdapat beberapa langkah sebagai berikut :
    a. Pemisahan fitur sesuai dengan label, pada tahapan ini memisahkan fitur sesuai dengan labelnya. Setiap fitur yang dipisahkan berisi variasi record, jumlah kemunculan setiap variasi record pada label positif dan negatif, serta total variasi record maupun label.
    b. Perhitungan nilai *entropy record* dan total, pada tahapan ini akan dilakukan perhitungan nilai *entropy record* dan total dengan menggunakan persamaan (12).
    c. Perhitungan nilai *entropy* fitur, pada tahapan ini akan dilakukan perhitungan nilai *entropy* fitur dengan menggunakan persamaan (13).
    d. Perhitungan nilai *information gain,* pada tahapan ini akan dilakukan perhitungan nilai *information gain* dengan menggunakan persamaan (14).
    e. Pengurutan fitur, pada tahapan ini dilakukan pengurutan fitur berdasarkan dengan nilai *information gain* dari yang paling tinggi hingga yang paling rendah.
    f. Pengambilan fitur, pada tahapan ini dilakukan pengambilan fitur berdasarkan dengan batas parameter yang dimasukkan.
8. **Pembagian Data**, pembagian data dilakukan dengan membagi data sebesar 70% untuk data latih dan 30% untuk data uji. Pada tahapan ini juga dilakukan *random oversampling* dengan menggunakan metode SMOTE untuk menyamaratakan kelas minor dengan kelas mayor.
9. **Klasifikasi**,pada tahapan ini dilakukan klasifikasi terhadap 2 model yaitu klasifikasi model *naïve bayes* dan klasifikasi model *naïve bayes* dan *information gain*.
10. **Evaluasi**, pada tahapan ini melakukan evaluasi dengan menggunakan *confusion matrix* dengan parameter yang digunakan yaitu akurasi, *recall*, *precision*, dan *specificity* menggunakan persamaan (15), (16), (17), dan (18). Evaluasi dilakukan untuk menganalisis model klasifikasi yang sebelumnya telah dibuat.

4 Hasil dan Pembahasan

Data yang digunakan pada penelitian ini didapatkan dari data komentar pada unggahan akun *instagram* transjakarta pada rentang waktu bulan Januari 2020 – Februari 2020 (sebelum adanya pandemi) sebanyak 550 data dan pada bulan Oktober 2021 – Februari 2022 (saat pandemi) data yang didapatkan sebanyak 2.351 data. Setelah data terkumpul dilakukan penghapusan pada data yang tidak mengandung sentimen sehingga didapatkan sebanyak 139 data sebelum adanya pandemi dan 700 data saat pandemi. Data tersebut akan digunakan pada tahapan pelabelan data yang dilakukan secara manual oleh 3 *annotator* dengan kategori sentimen positif dan sentimen negatif. Hasil pelabelan data didapatkan berdasarkan nilai mayoritas dari ketiga annotator. Dari tahapan pelabelan data pada rentang waktu sebelum adanya pandemi didapatkan sebanyak 10 data pada kategori sentimen positif dan 129 data pada kategori sentimen negatif . Hasil pelabelan data pada rentang waktu saat pandemi didapatkan sebanyak 177 data pada kategori sentimen positif dan 523 data pada kategori sentimen negatif. Data pada penelitian ini berfokus pada data dengan rentang waktu pada bulan Oktober 2021 – Februari 2022. Untuk mendapatkan hasil kesepakatan dari proses pelabelan data dilakukannya perhitungan dengan menggunakan persamaan rumus (1), (2), (3) pada 700 data komentar. Adapun perhitungan dan hasil dari perhitungan itu didapatkan sebagai berikut :

$$K = \frac{P_0 - P_e}{1 - P_e} = \frac{0.974286 - 0.625478}{1 - 0.625478} = 0.931341$$

Hasil kappa yang didapat dari 700 data komentar sebesar 0.931341 dimana masuk pada hasil *almost perfect* berdasarkan skala kappa value pada Tabel 1 yang mana memiliki arti baik dan konsisten. Setelah melakukan pelabelan data, masuk ketahapan selanjutnya yaitu *preprocessing*. Berikut merupakan hasil pada tahapan *preprocessing* pada satu data komentar yang sebelum dan sesudah dilakukannnya *preprocessing* :

Tabel 3. Hasil *Preprocessing*

Sebelum	Sesudah
mimpi nyata transjakarta rute sunter langsung harmoni	[‘mimpi’, ‘nyata’, ‘transjakarta’, ‘rute’, ‘sunter’, ‘langsung’, ‘harmoni’]
sayang bus transjakarta operasi pekan	[‘sayang’, ‘bus’, ‘transjakarta’, ‘operasi’, ‘pekan’]
layan transjakarta buruk banyak citra	[‘layan’, ‘transjakarta’, ‘buruk’, ‘banyak’, ‘citra’]
transjakarta jaga jarak covid ampun parah	[‘transjakarta’, ‘jaga’, ‘jarak’, ‘covid’, ‘ampun’, ‘parah’]

Setelah melalui tahapan *preprocessing* selanjutnya dilakukan dengan melakukan pembobotan kata menggunakan TF-IDF. Pada tahapan ini akan dilakukan dengan 700 dokumen/komentar. Pada Tabel 4 akan memuat beberapa contoh fitur pada tahapan TF-IDF.

Tabel 4. Pembobotan TF-IDF

Fitur	TF				DF	IDF	TF-IDF			
	D1	D2	D3	D4			D1	D2	D3	D4
mimpi	1	0	0	0	0.602	0.602	0.602	0	0	0
transjakarta	1	1	1	1	0	0	0	0	0	0

Pada tahapan pembobotan kata menghasilkan 1.489 kata/*term* yang diberi bobot yang mana kata/term tersebut akan dijadikan fitur dalam proses klasifikasi. Untuk model klasifikasi sentimen *naïve bayes* dan *information gain* dilanjutkan dengan melakukan tahapan seleksi fitur dengan metode *information gain*.

Tabel 5. Hasil Seleksi Fitur *Information Gain*

Fitur		Positif	Negatif	Total	Entropi	Entropi Fitur	Gain
mimpi	0	0	3	3	0	0	0.811278
	0.602	1	0	1	0		
transjakarta	0	1	3	4	0.811278	0.811278	0
Total		1	3	4	0.811278		

Pada Tabel 5 memuat beberapa contoh fitur pada tahapan seleksi fitur. Pada tahapan seleksi fitur *information gain* ini menggunakan 700 data komentar dengan fitur sebanyak 1.489 yang didapatkan dari tahapan TF-IDF. Pada tahapan seleksi fitur menggunakan parameter batas dimana N>0.001 dan mendapatkan 597 fitur. Setelah tahapan pembobotan kata dan seleksi fitur selesai dilakukan, maka akan dilanjutkan dengan melakukan tahapan pembagian data. Data latih dibagi menjadi 70% dan data uji dibagi menjadi 30% pada 700 data komentar yang akan dimuat pada Tabel 6 dan Tabel 7 sebagai berikut :

Tabel 6. Pembagian Data Model Klasifikasi *Naïve Bayes*

Pembagian Data	Model Klasifikasi *Naïve Bayes*		Jumlah	Jumlah Fitur
	Positif	Negatif		
Data Latih	121	369	490	1.489
Data Uji	56	154	210	
Total	177	523	700	

Tabel 7. Pembagian Data Model Klasifikasi *Naïve Bayes* dan *Information Gain*

Pembagian Data	Model Klasifikasi *Naïve Bayes*	Jumlah	Jumlah Fitur

	Positif	Negatif		
Data Latih	121	369	490	597
Data Uji	56	154	210	
Total	177	523	700	

Jika dilihat pada Tabel 6 dan Tabel 7, terjadi ketidakseimbangan data antara kelas positif dan kelas negatif, untuk itu dilakukannya oversampling dengan menerapkan metode SMOTE untuk menyamaratakan jumlah data pada data latih yang ada pada kelas yang minoritas agar sesuai dengan kelas mayoritas yang dimuat dalam Tabel 8.

Tabel 8. Penerapan metode SMOTE

	Positif	Negatif
Data Latih	369	369

Setelah melakukan pembagian data kemudian masuk pada tahapan selanjutnya yaitu melakukan klasifikasi. Proses klasifikasi pada penelitian ini menggunakan *Multinomial Naïve Bayes*. Hasil dari proses klasifikasi akan dimuat dalam Tabel 9 dan Tabel 10 berupa *confusion matrix.*

Tabel 9. *Confusion Matrix* Model Klasifikasi *Naïve Bayes*

		Nilai yang sebenarnya	
		Positive	Negatif
Prediksi	Positive	39	22
	Negative	17	132

Tabel 10. *Confusion Matrix* Model Klasifikasi *Naïve Bayes* dan *Information Gain*

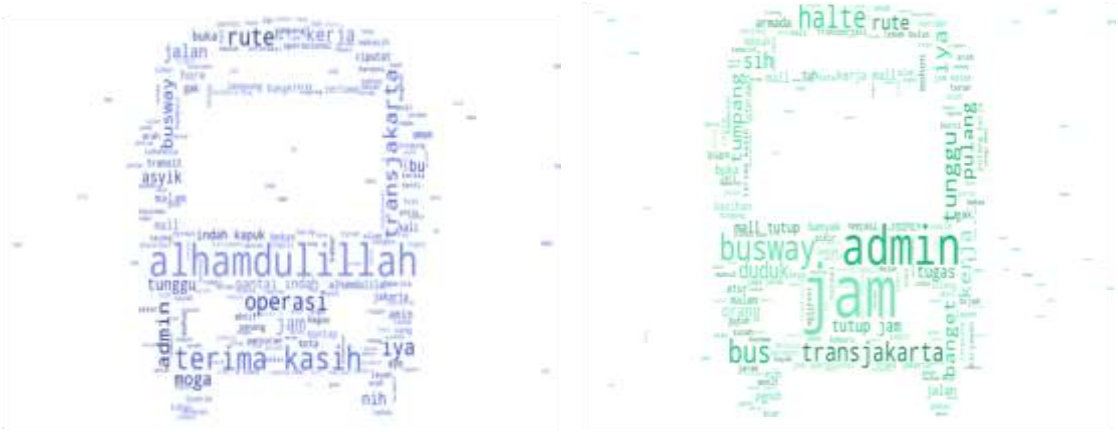
		Nilai yang sebenarnya	
		Positive	Negatif
Prediksi	Positive	40	12
	Negative	16	142

Berdasarkan dengan hasil confusion matrix akan dilakukan evaluasi untuk mengukur performa model yang telah dibangun dengan menggunakan parameter akurasi, *recall*, *precision*, dan *specificity* yang akan dimuat pada Tabel 11.

Tabel 11. Hasil Evaluasi Perbandingan Model Klasifikasi

	Model Klasifikasi *Naïve Bayes*	Model Klasifikasi *Naïve Bayes* dan *Information Gain*
Akurasi	0.8142	0.8666
Recall	0.6964	0.7142
Precision	0.6393	0.7692
Specificity	0.8571	0.9220

Setelah mendapatkan hasil evaluasi kemudian melakukan visualisasi yang dimuat dalam bentuk wordcloud pada sentimen positif dan sentimen negatif.



(a) (b)

Gambar 2. *Wordcloud* sentimen (a) positif dan (b) negatif. merupakan alur penelitian yang telah dilakukan dimana terdapat beberapa langkah-langkah penelitian. *Wordcloud* menunjukan kata yang dominan yang digunakan dalam data komentar terhadap layanan transjakarta di media sosial instagram. Jika dilihat pada bagian (a) terdapat kata "alhamdulillah", "terima kasih", "asyik" yang digunakan dalam sentimen positif. Sedangkan pada bagian (b) terdapat kata "tunggu", "tutup", "jam" yang digunakan dalam sentimen negatif.

5 Kesimpulan dan Saran

Pada penelitian ini menggunakan 700 data komentar yang terdiri dari 177 data dengan label sentimen positif dan 523 data dengan label sentimen negatif. Hasil evaluasi performa model klasifikasi *Naïve Bayes* dengan jumlah fitur yang digunakan sebanyak 1.489 kata memperoleh akurasi sebesar 81,42%, *recall* sebesar 69,64%, *precision* sebesar 63,93% dan *specificity* sebesar 85,71%. Sedangkan hasil evaluasi performa model klasifikasi *Naïve Bayes* dan *Information Gain* dengan jumlah fitur sebanyak 597 kata memperoleh akurasi sebesar 86,66%, *recall* sebesar 71,42%, *precision* sebesar 76,92%, dan *specificity* sebesar 92,20%. Berdasarkan hasil tersebut, diketahui bahwa hasil klasifikasi *Naïve Bayes* dengan seleksi fitur menghasilkan performa yang lebih baik.

Penelitian selanjutnya diharapkan dapat mengumpulkan lebih banyak data mengenai layanan transjakarta dari berbagai media sosial dengan cakupan pembahasan selain informasi layanan dan rute dengan menggunakan metode klasifikasi dan seleksi fitur lainnya. Diharapkan juga dapat memuat lebih banyak daftar kamus normlisasi dan proses pelabelan dilakukan oleh bantuin ahli.

Referensi

[1] Yustinus, A. D. P. (2020). Mulai Senin, Pembatasan Transportasi Umum Di Jakarta Berlaku. URL : https://jakarta.bisnis.com/read/20200321/77/1216430/mulai-senin-pembatasan-transportasi-umum-di-jakarta-berlaku. Diakses pada 10 November 2021.

[2] Lestari, T. Y. (2020). Transjakarta Hanya Operasikan 13 Rute, Penumpang di Sejumlah Halte Mengular Panjang. URL : https://www.merdeka.com/peristiwa/transjakarta-hanya-operasikan-13-rute-penumpang-di-sejumlah-halte-mengular-panjang.html. Diakses pada 10 November 2021.

[3] Nichols, T. R., Wisner, P. M., Cripe, G., & Gulabchand, L. (2010). Putting the kappa statistic to use. Quality Assurance Journal, 13(3–4), 57–61. https://doi.org/10.1002/qaj.481

[4] Werdiningsih, I., Nuqoba, B. & M., 2020. Data Mining Menggunakan Android, Weka, dan SPSS. Surabaya: Airlangga University Press.

[5] Nugraha, F. A., Harani, N. H. & Habibi, R., 2020. Analisis Sentimen Terhadap Pembatasan Sosial Menggunakan Deep Learning. Bandung: Kretif Industri Nusantara.

[6] Luqyana, W. A., Cholissodin, I., & Perdana, R. S. (2018). Analisis Sentimen Cyberbullying Pada Komentar Instagram dengan Metode Klasifikasi Support Vector Machine. Jurnal Pengembangan Teknologi Informasi Dan Ilmu Komputer (J-PTIIK) Universitas Brawijaya, 2(11), 4704–4713

[7] Yutika, C. H., Adiwijaya, A., & Faraby, S. Al. (2021). Analisis Sentimen Berbasis Aspek pada Review Female Daily Menggunakan TF-IDF dan Naïve Bayes. Jurnal Media Informatika Budidarma, 5(2), 422–430. https://doi.org/10.30865/mib.v5i2.2845
[8] Irsyad, H., Farisi, A., & Pribadi, M. R. (2019). Klasifikasi Opini Masyarakat Terhadap Jasa ISP MyRepublic dengan Naïve Bayes. Jurnal Nasional Teknik Elektro Dan Teknologi Informasi (JNTETI), 8(1), 30. https://doi.org/10.22146/jnteti.v8i1.487
[9] Sabrani, A., Wedashwara W., I. G. W., & Bimantoro, F. (2020). Multinomial Naïve Bayes untuk Klasifikasi Artikel Online tentang Gempa di Indonesia. Jurnal Teknologi Informasi, Komputer, Dan Aplikasinya (JTIKA), 2(1), 89–100. https://doi.org/10.29303/jtika.v2i1.87
[10] Aini, S. H. A., Sari, Y. A., & Arwan, A. (2018). Seleksi Fitur Information Gain untuk Klasifikasi Penyakit Jantung Menggunakan Kombinasi Metode K-Nearest Neighbor dan Naïve Bayes. Jurnal Pengembangan Teknologi Informasi Dan Ilmu Komputer, 2(9), 2546–2554.
[11] Adinugroho, Sigit., & Yuita Arum Sari. (2018). Implementasi Data Mining Menggunakan WEKA. Malang : UB Press.