Analisis Sentimen Pengguna Twitter Terhadap Layanan Provider IndiHome Menggunakan Algoritma Naïve Bayes

Sentiment Analysis of IndiHome Service Provider Twitter Using Naïve Bayes Algorithm

Muhamad Enggar Aziz Hibbannuari*[1], Hartatik[2]
[1,2,3] Program Studi Informatika, Fakultas Ilmu Komputer, Universitas AMIKOM Yogyakarta
e-mail: [1]m.engga.azizr@mhsamikom.ac.id, [2]hartatik@amikom.ac.id
*Correspondence author email: *[2]hartatik@amikom.ac.id*

Abstrak

Indihome merupakan salah satu penyediaan layanan internet (ISP) yang ada di Indonesia yang jangkauan layanan sudah tercakup ke seluruh wilayah Indonesia. Sebuah penyedia layanan seperti Indihome pasti memiliki keluhan atau aduan tentang kepuasan atau kelayakan menggunakan Indihome, melalui data pengguna Twitter untuk mendapatkan akurasi nilai positif dan negatif terhadap layanan Indihome. Penelitian ini akan melakukan klasifikasi sentiment tweet terhadap layanan provider indihome dengan algoritma naiva bayes classifier. Selanjutnya dataset akan diklasifikasi kedalam class positif dan negative yang diuji melalui confusion matrix untuk mendapatkan nilai akurasi, recall dan precission. Adapun nilai akurasi yang dihasilkan cenderung cukup tinggi melalui dua kali pengujian.
Katakunci: *Sentiment Analisis, Pengguna Twitter, Naïve Bayes Classification, Layanan, Provider Indihome*

Abstrack

Indihome is one of the internet service providers (ISP) in Indonesia whose service coverage has covered all regions of Indonesia. A service provider like Indihome definitely has complaints or complaints about the satisfaction or feasibility of using Indihome, through Twitter user data to get accurate positive and negative values for Indihome services. This research will classify tweet sentiments for indihome service providers using the Naiva Bayes classifier algorithm. Furthermore, the dataset will be classified into positive and negative classes which are tested through a confusion matrix to get accuracy, recall and precision values. The resulting accuracy values tend to be quite high through two test
Keyword: *Sentiment Analysis, Twitter User, Naïve Bayes Classification, Service, Indohome Provider*

1. PENDAHULUAN

Indihome merupakan salah satu penyedia layanan internet (ISP) yang ada di Indonesia yang jangkauan layanan sudah tercakup ke seluruh wilayah Indonesia[1]. Sebuah penyedia layanan seperti Indihome pasti memiliki keluhan atau aduan tentang kepuasan atau kelayakan menggunakan Indihome, melalui data pengguna Twitter yang didapat, selanjutnya akan dilakukan analisis sentimen untuk mendapatkan akurasi nilai positif, negatif ataupun netral terhadap layanan Indihome. *Naïve Bayes Classifier* adalah algoritma pembelajaran mesin probabilistik yang dapat digunakan untuk klasifikasi[2]. Pada saat melakukan klasifikasi, algoritma akan melakukan pencarian nilai probabilitas yang tertinggi. Pada pemrosesan perhitungan Naïve Bayes hanya memerlukan tiga tahapan saja, dimulai dari perhitungan *prior, likelihood dan posterior*[3].

Menggunakan data *tweet* pada Twitter merupakan salah satu manfaat layanan yang disediakan untuk aduan atau keluhan terhadap Indihome. Dari data yang di ambil dari Twitter tersebut berupa ulasan keluhan atau aduan terhadap layanan untuk setiap orang memiliki perbedaan oleh karena itu peneliti melakukan penelitian ini melalui proses klasifikasi analisis sentimen pengguna Twitter menggunakan *Naive Bayes Classifier.* Pengambilan data *tweet* pada

History of article:
Received: mm, yyyy : Accepted: mm, yyyy

Twitter menggunakan script Bahasa Python kemudian data tersebut di klasifikasi menggunakan *Naive Bayes Classifier*. Prosesnya meliputi pengumpulan data *tweet* pada Twitter dengan beberapa kata kunci seperti, indihome, IndiHomeCare dan sebagainya. Selanjutnya akan dilakukan pengumpulan data *tweet* menggunakan *script* bahasa Python dengan *package snscrape*. Kemudian akan dihitung pembobotan TF-IDF. Yang selanjutnya akan dilakukan klasifikasi dan pengolahan dataset untuk memperoleh nilai akurasi menggunakan *Naive Bayes Classifier*. Berdasarkan latar belakang yang telah dikemukakan, maka permasalahan yang dapat dirumuskan adalah bagaimana algoritma *Naive Bayes Classifier* mampu mengklasifikasi sentimen *tweet* terhadap layanan provider Indihome. Adapun tujuan dari penelitian ini adalah mengetahui ketepatan akurasi klasifikasi algoritma *Naive Bayes Classifier* dengan metode *TF-IDF*.

Beberapa penelitian mengenai analisis sentiment pengguna pada twitter pernah dilakukan untuk mengetahui respon masyarakat terhadap pelayanan BMKG Nasional dengan akurasi sebesar 69,97%[4], *Naïve Bayes Classifier* juga pernah digabungkan dengan SVM untuk mengetahui dampak covid terhadap masyarakat [5], selain itu Naïve Bayes juga pernah digunakan untuk mengetahui respon masyarakat terhadap proses pembelajaran daring [6]. Dari sisi analisis pelayanan, Naïve Bayes juga pernah digunakan untuk mengetahui respon masyarakat terhadap pelayanan BPJS melalui 3 algoritma yaitu KNN, *Decision Tree* dan *Naïve Bayes*[7], pelayanan KRL commuterline dengan Bernoulli Naïve Bayes [8].

2. METODE PENELITIAN

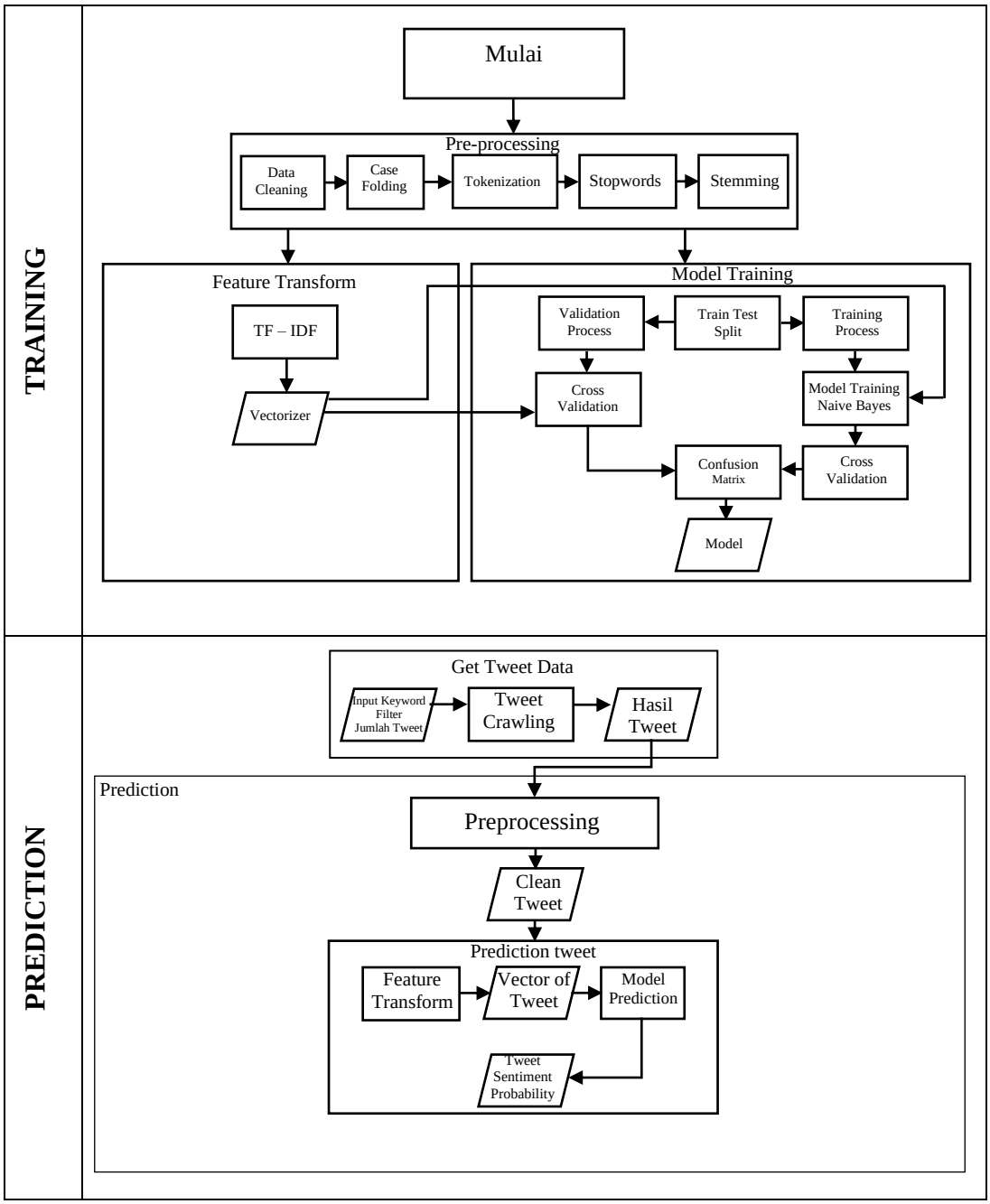
Obyek Penelitian

Dalam penelitian ini obyek penelitian yang digunakan adalah penyedia layanan milik BUMN yaitu IndiHome. Dengan obyek tersebut, peneliti akan mengumpulkan dataset dari twitter mengenai berbagai macam tweet yang berkaitan dengan layanan IndiHome yang diberikan. Untuk mendapatkan data tersebut peneliti melakukan proses scrapping data pada twitter menggunakan Bahasa pemrograman phyton dengan library *snscrape*

Metode Pengumpulan Data

Dalam penelitian ini, metode pengumpulan data yang digunakan adalah:

a. Studi Pustaka
   Pada studi pustaka, peneliti akan melakukan pencarian baik dari jurnal dan prosiding[9] terkait dengan topik serupa yang berkaitan dengan *Naïve Bayes Classification, Text Mining* serta Analisis Sistem[10].
b. Interview
   Pada proses ini, peneliti melakukan wawancara dengan seorang ahli yang pernah melakukan penelitian mengenai sentiment analisis baik dengan Algoritma *Naïve Bayes* maupun yang lain.
c. Alur penelitian
   Sedangkan untuk alur penelitian, peneliti melakukan sesuai dengan gambar 1. Dimana berdasarkan gambar tersebut penelitian akan dibagi menjadi 2 tahapan yaitu training dan testing[11].
   Pada tahapan data training akan dimulai dengan tahapan pre-processing (data *cleaning – folding – tokenization – stopword* – dan *stemming*) [12] yang kemudian dilanjutkan dengan *feature transform* (melalui algoritma TF-IF yang kemudian dilakukan tahapan *vectorizer*) kemudian dilakukan *training* data. Sedangkan pada tahapan *prediction* akan diawali dengan tahapan mendapatkan data tweet melalui penyaringan jumlah *tweet – crawling* data – hasil tweet. Selanjutnya akan dilakukan proses *prediction* yang diawali dengan *preprocessing* data kemudian *cleen tweet* kemudian *prediction* hasil tweet



Gambar 1. Alur penelitian yang dilakukan

Bahan Penelitian

Untuk bahan penelitian, dataset didapatkan menggunakan Twitter API dengan menggunakan proses *scrapping* data menggunakan Bahasa phyton melalui katakunci IndiHome ataupun IndiHomeCare. Tabel 1 adalah contoh dari dataset yang digunakan.

Tabel 1. Sample dataset penelitian

Index	Data Tweet
1	@koniciwaciwa Ko indihome gue fine fine aja dah? \ud83d\ude2d
2	teknisi Indihome cepet banget dah anyink. gua laporan masalah kabel LAN doang aja mas-masnya dateng dalam 5 menit wkwk

3	Gak terasa udah hampir 5 thn lbh memakai INDIHOME \ud83e\udd70 , Januari besok 6 thn https://t.co/YfqGGNl4Nz
4	@ndawwuu__ @IndiHome gangguan kah bang? disini malah dari jam 3 pagi gak kelar kelar.
5	Indihome ni knp buset merah mulu lampu indikator nya dari tadi\ud83d\ude2d
…	….

3. HASIL DAN PEMBAHASAN

Pada tahap awal, dataset akan didapatkan setelah melalui proses *scraping*. Pada bagian ini dataset akan diolah dengan cara dibersihkan seperti *mention* / *username* dan *hastag*. Proses *textminig* akan dilakukan setelah selesai membersihkan kata yang tidak diperlukan. Setelah dibersihkan maka akan dilakukan case folding yang akan merubah dataset menjadi huruf kecil semua. Tabel 2 merupakan hasil case folding ari data cleaning.

Tabel 2. Hasil case folding

Data	Label
ko indihome gue fine fine aja dah	Positive
teknisi indihome cepet banget dah anyink gua laporan masalah kabel lan doang aja mas-masnya dateng dalam 5 menit wkwk	Positive
gak terasa udah hampir 5 thn lbh memakai indihome januari besok 6 thn	Positive
gangguan kah bang disini malah dari jam 3 pagi gak kelar kelar	Negative
….	….
check paket langganan indihome kalian guys	Positive

Selanjutnya akan dilakukan *tokenization* untuk mendapatkan kata dari sebuah kalimat melalui pemisahan kalimat menjadi perkata dan juga menghapus data tanda baca. Langkah selanjutnya melakukan *stopwords* untuk menghapus kata pada kalimat yang tidak diperlukan. Hasil dari *tokenization* dan *stopword* terangkum pada Tabel 3.

Tabel 3. Hasil tokenization dan stopword

Data Token	Data Stopwords
['ko', 'indihome', 'gue', 'fine', 'fine', 'aja', 'dah']	['ko', 'indihome', 'gue', 'fine', 'fine', 'aja', 'dah']
['teknisi', 'indihome', 'cepet', 'banget', 'dah', 'anyink', 'gua', 'laporan', 'masalah', 'kabel', 'lan', 'doang', 'aja', 'mas-masnya', 'dateng', 'dalam', '5', 'menit', 'wkwk']	['teknisi', 'indihome', 'cepet', 'banget', 'dah', 'anyink', 'gua', 'laporan', 'masalah', 'kabel', 'lan', 'doang', 'aja', 'mas-masnya', 'dateng', '5', 'menit', 'wkwk']
['gak', 'terasa', 'udah', 'hampir', '5', 'thn', 'lbh', 'memakai', 'indihome', 'januari', 'besok', '6', 'thn']	['gak', 'terasa', 'udah', 'hampir', '5', 'thn', 'lbh', 'memakai', 'indihome', 'januari', 'besok', '6', 'thn']
['gangguan', 'kah', 'bang', 'disini', 'malah', 'dari', 'jam', '3', 'pagi', 'gak', 'kelar', 'kelar']	['gangguan', 'bang', 'disini', 'malah', 'jam', '3', 'pagi', 'gak', 'kelar', 'kelar']

Kemudian akan dilakukan *stemming* untuk mengembalikan kata ke dalam kata dasar. Pada penelitian ini proses *stemming* menggunakan library pyhton bernama sastrawi. Selanjutnya pada proses feature transform akan dilakukan dengan menggunakan algoritma TF-IDF. Adapun hasil dari TF-IDF dari dokumen yang digunakan terdapat pada tabel 4.

Tabel 4. Hasil TF-IDF

Term	Term Frequency										Inverse Document Frequency									
	T1	T2	T3	T4	T5	T6	T7	T8	T9	DF	IDF	T1	T2	T3	T4	T5	T6	T7	T8	T9
indihome	1	1	1	0	1	1	1	2	1	8	0,051	0,05	0,051	0,051	0	0,051	0,051	0,051	0,102	0,051

gue	1	0	0	0	0	1	0	0	0	2	0,653	0,65	0	0	0	0	0,653	0	0	0
masalah	0	1	0	0	0	0	0	0	0	1	0,954	0	0,954	0	0	0	0	0	0	0
dah	1	1	0	0	0	0	0	0	0	2	0,653	0,65	0,653	0	0	0	0	0	0	0
…	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..
season	0	0	0	0	0	0	0	1	0	0,00	0,00	0	0	0	0	0	0	0	0,954	0

Selanjutnya akan dilakukan klasifikasi melalui pembagian data *training* dan data *testing*. Pada penelitian ini peneliti akan menggunakan perbandingan 50 : 50 untuk data *training* dan data *testing* yang akan dilakukan. Proses klasifikasi yang digunakan adalah Algoritma *Naïve Bayes* dengan mencari nilai probabilitas tertinggi dengan menggunakan probabilitas *likelihood*. Untuk mencari probabilitas *likehood* positif dihitung dengan cara:

$$pelayanan = \frac{0{,}602059991 + 0}{5{,}719569918} = 0{,}1052631579$$

$$internet = \frac{0 + 0{,}301029996}{5{,}719569918} = 0{,}05263157895$$

$$banget = \frac{0 + 0{,}301029996}{5{,}719569918} = 0{,}05263157895$$

Dengan cara yang sama akan dihitung juga probabilitas *likehood* negative. Selanjutnya akan dihitung prior data positif dan negativenya. Berdasarkan hasil perhitungan didapatkan nilai prior positif dan negative memiliki nilai yang sama yaitu 0,5.
Langkah selanjutnya yaitu mencari nilai posterior. Hasil dari perhitungan ini adalah:

$$pelayanan = \frac{0{,}1052631579 \times 0{,}5}{1} = 0{,}05263157895$$

$$internet = \frac{0{,}05263157895 \times 0{,}5}{1} = 0,\ 02631578947$$

$$banget = \frac{0{,}05263157895 \times 0{,}5}{1} = 0{,}02631578947$$

Selanjutnya akan dilakukan prediksi terhadap frekuensi tiap kata terhadap dataset baru untuk mengetahui score pada tiap dokumen.

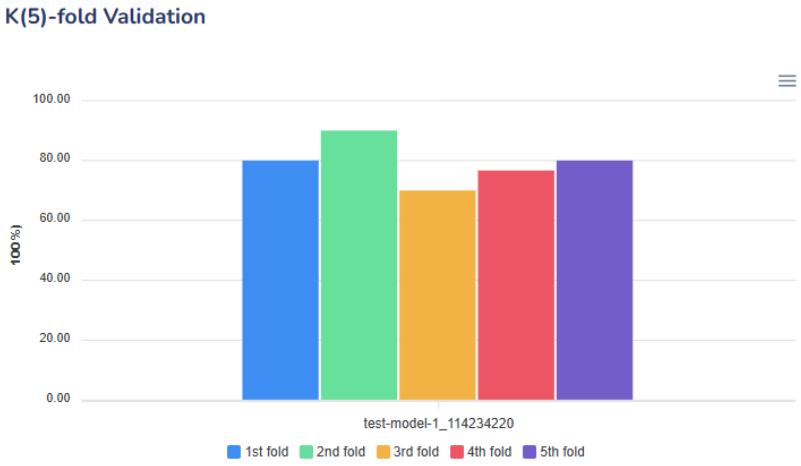
Hasil Pengujian Sistem

Pada pengujian sistem, peneliti menggunakan *k-fold cross validation* dengan nilai k=5. Adapun hasil dari pengujian ini adalah:

a. Pada tahapan kfold pertama yang peneliti uji hasil dari pengujian menggunakan 5 kfold cross validation belum menggunakan confusion matrix maka memperoleh data akurasi sebanyak 80% dari total data uji, yaitu 120 data yang sesuai dengan label untuk prediksi tweet dan 180 data prediksi tidak sesuai dengan label untuk prediksi tweet.
b. Pada tahapan kfold kedua yang peneliti uji hasil dari pengujian menggunakan 5 kfold cross validation belum menggunakan confusion matrix maka memperoleh data akurasi sebanyak 90% dari total data uji, yaitu 135 data yang sesuai dengan label untuk prediksi tweet dan 165 data prediksi tidak sesuai dengan label untuk prediksi tweet
c. Tahapan kfold ketiga yang peneliti uji hasil dari pengujian menggunakan 5 kfold cross validation belum menggunakan confusion matrix maka memperoleh data akurasi sebanyak 70% dari total data uji, yaitu 105 data yang sesuai dengan label untuk prediksi tweet dan 195 data prediksi tidak sesuai dengan label untuk prediksi tweet
d. Tahapan kfold keempat yang peneliti uji hasil dari pengujian menggunakan 5 kfold cross validation belum menggunakan confusion matrix maka memperoleh data akurasi sebanyak 76,67% dari total data uji, yaitu 115 data yang sesuai dengan label untuk prediksi tweet dan 185 data prediksi tidak sesuai dengan label untuk prediksi tweet

e. Tahapan kfold keempat yang peneliti uji hasil dari pengujian menggunakan 5 kfold cross validation belum menggunakan confusion matrix maka memperoleh data akurasi sebanyak 80% dari total data uji, yaitu 120 data yang sesuai dengan label untuk prediksi tweet dan 180 data prediksi tidak sesuai dengan label untuk prediksi tweet.

Rangkuman mengenai hasil pengujian tersebut terdapat pada gambar 2.



Gambar 2. Hasil pengujian 5 fold

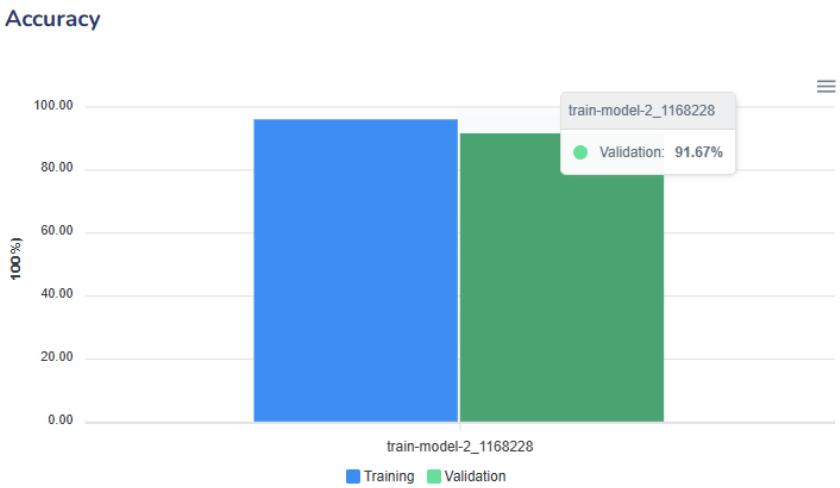
Berdasarkan gambar 2 dapat disimpulkan bahwa *kfolds ke-2* memiliki akurasi tertinggi pada *kfold cross validation* dengan akurasi sebesar 90% dan akurasi terendah pada *kfolds ke-3* dengan akurasi sebesar 70% serta rata-rata nilai akurasi nya adalah 79,34%. Pengujian kedua dilakukan dengan menggunakan confusion matrix dimana pengujian dilakukan dengan membagi 180 data training dan 120 data testing. Berdasarkan hasil perhitungan confusion matrix pada data training didapatkan:

a. *False Negative* (FN), jumlah nilai negative salah 173.
b. False Positive (FP), jumlah negative benar 7.
c. True Negative (TN), jumlah positif salah 0.
d. True Positive (TP), jumlah positif benar 0.

Sedangkan nilai untuk data testing pada perhitungan confusion matrixnya adalah:

a. *False Negative* (FN), jumlah nilai negative salah 110.
b. False Positive (FP), jumlah negative benar 10.
c. True Negative (TN), jumlah positif salah 0.
d. True Positive (TP), jumlah positif benar 0

Rangkuman mengenai hasil pengujian tersebut tergambar pada gambar 3. Berdasarkan gambar 3 diperoleh nilai akurasi sebesar 91,67%. Untuk precision, recall dan f1 score tidak memiliki diagram batang, dikarenakan kurangnya dataset, untuk itu perlu dicoba lagi dengan dataset baru yang memiliki nilai sentiment lebih baik. Dari tahapan pembanding dataset yang peneliti uji pada dataset 1, nilai akurasi 81,33% seperti terlihat pada gambar 4.4, sedangkan Ketika diuji dengan dataset baru, akurasi mengalami kenaikan menjadi 91,67%.



Gambar 3. Hasil uji confussion matrix

4. KESIMPULAN

Pada penelitian ini dilakukan pengujian dataset menggunakan Naïve Bayes dan pembobotan TF-IDF dengan cara melakukan proses *mining data* pada aplikasi Twitter, kemudian akan memasuki *pre-processing* guna data dapat diolah dan digunakan, dataset akan dilabeli dengan positif dengan negatif. Untuk selanjutnya dilakukan proses training dan prediksi. Dari tahap penelitian ini nanti akan disajikan nilai akhir berupa *accuracy, recall, precision* dan *f1 score*. Dari setiap algoritma *k-fold(5) cross validation*. Dari pengujian Naïve Bayes didapatkan nilai 81,33% dengan pembobotan TF-IDF dan *recall 74,29%, precision 83,87% dan f1 score 78,79%*. Sedangkan pengujian kedua dengan dataset baru, memiliki hasil yang berbeda yaitu untuk nilai *accuracy 91,67%*. Pada pengujian kedua ini nilai *accuracy* lebih tinggi, tetapi nilai recall, precision, dan f1 score 0, dikarenakan label pada dataset tidak seimbang, karena dominan negatif. Untuk itu diperlukan pengujian kembali. Dari hasil penelitian ini diharapkan calon pelanggan dapat lebih bijak sebelum memutuskan dalam menentukan provider yang akan di gunakan. Walaupun pengujian terhadap testing memberikan hasil memuaskan, untuk perlu diperhatikan bahwa presisi atas suatu klasifikasi diperlukan evaluasi lagi terhadap prediksi yang telah didapat.

DAFTAR PUSTAKA

[1] I. K. Suarjana and N. W. S. Suprapti, "Pengaruh Persepsi Harga, Pengetahuan Produk, Dan Citra Perusahaan Terhadap Niat Beli Layanan Multi Servis Merek Indihome," *E-Jurnal Manaj. Univ. Udayana*, vol. 7, no. 4, 2018, doi: 10.24843/EJMUNUD.2018.v7.i04.p08 ISSN.

[2] A. Nugroho and Y. Religia, "Analisis Optimasi Algoritma Klasifikasi Naive Bayes menggunakan Genetic Algorithm dan Bagging," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 5, no. 3, pp. 504–510, 2021, doi: 10.29207/resti.v5i3.3067.

[3] A. B. Putra Negara, H. Muhardi, and I. M. Putri, "Analisis Sentimen Maskapai Penerbangan Menggunakan Metode Naive Bayes Dan Seleksi Fitur Information Gain," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 7, no. 3, pp. 599–606, 2020, doi: 10.25126/jtiik.202071947.

[4] D. Darwis, N. Siskawati, and Z. Abidin, "Penerapan Algoritma Naive Bayes Untuk Analisis Sentimen Review Data Twitter BMKG Nasional," *J. Tekno Kompak*, vol. 15,

no. 1, p. 131, 2021, doi: 10.33365/jtk.v15i1.744.
[5] C. F. Hasri and D. Alita, "Penerapan Metode Naïve Bayes Classifier Dan Support Vector Machine Pada Analisis Sentimen Terhadap Dampak Virus Corona Di Twitter," *J. Inform. dan Rekayasa Perangkat Lunak*, vol. 3, no. 2, pp. 145–160, 2022, doi: 10.33365/jatika.v3i2.2026.
[6] S. Samsir, A. Ambiyar, U. Verawardina, Fi. Edi, and R. Watrianthos, "Analisis Sentimen Pembelajaran Daring Pada Twitter di Masa Pandemi COVID-19 Menggunakan Metode Naïve Bayes," *J. Media Inform. Budidarma*, vol. 5, no. 1, p. 149, 2021, doi: 10.30865/mib.v5i1.2604.
[7] R. Puspita and A. Widodo, "Perbandingan Metode KNN, Decision Tree, dan Naïve Bayes Terhadap Analisis Sentimen Pengguna Layanan BPJS," *J. Inform. Univ. Pamulang*, vol. 5, no. 4, p. 646, 2021, doi: 10.32493/informatika.v5i4.7622.
[8] M. Saraswati and D. Riminarsih, "Analisis Sentimen Terhadap Pelayanan Krl Commuterline Berdasarkan Data Twitter Menggunakan Algortima Bernoulli Naive Bayes," *J. Ilm. Inform. Komput.*, vol. 25, no. 3, pp. 225–238, 2020, doi: 10.35760/ik.2020.v25i3.3256.
[9] C. Hasanudin, S. Subyantoro, I. Zulaeha, and R. Pristiwati, "Strategi Menyusun Bahan Ajar Inovatif Berbasis Mobile Learning untuk Pembelajaran Mata Kuliah Keterampilan Menulis di Abad 21," *Pros. Semin. Nas. Pascasarj.*, pp. 343–347, 2021, [Online]. Available: http://pps.unnes.ac.id/prodi/prosiding-pascasarjana-unnes/
[10] M. Muhibun, A. Firmansyah, M. Fatchan, and I. Afrianto, "Sistem Informasi Data Stok Pallet Pada CV. Selang Surya Kencana," *J. Autom. Comput. Inf. Syst.*, vol. 1, no. 1, pp. 1–7, 2021, doi: 10.47134/jacis.v1i1.1.
[11] H. Hartatik, S. D. Nurhayati, and W. Widayani, "Sistem Rekomendasi Wisata Kuliner di Yogyakarta dengan Metode Item-Based Collaborative Filtering," *J. Autom. Comput. Inf. Syst.*, vol. 1, no. 2, pp. 55–63, 2021, doi: 10.47134/jacis.v1i2.8.
[12] F. S. Lestari, H. Harliana, M. M. Huda, and T. Prabowo, "Sentiment Analysis of iPusnas Application Reviews on Google Play Using Support Vector Machine," *Proc. Int. Semin. Business, Educ. Sci.*, vol. 1, no. August, pp. 178–188, 2022, doi: 10.29407/int.v1i1.2656.