Progresif: Jurnal Ilmiah Komputer
Jl. Ahmad Yani, K.M. 33,5 - Kampus STMIK Banjarbaru
Loktabat – Banjarbaru (Tlp. 0511 4782881), e-mail: puslit.stmikbjb@gmail.com
e-ISSN: 2685-0877
p-ISSN: 0216-3284

Algoritme Multinomial *Naïve Bayes* Pada Aplikasi *Chatbot* Layanan Informasi Berbasis Teks

Asep Muhidin[1*], Muhtajuddin Danny[2], Elkin Rilvani[3]
Teknik Informatika, Universitas Pelita Bangsa, Bekasi, Indonesia
*e-mail *Corresponding Author:* asep.muhidin@pelitabangsa.ac.id

Abstract

The existence of Customer Service (CS) officers whose number is not balanced with the number of students served, has resulted in a decrease in the level of satisfaction with campus services. This study aims to create a chatbot application that can help with CS tasks. This text-based chatbot application was created using a Natural Language Processing (NLP) approach and a Machine Learning algorithm. NLP is used to process a text question from the user, while the MultinomialNB machine learning algorithm is used to find the appropriate data. If found, the system will answer the question based on the label obtained from the machine learning model. The dataset used as Chatbot training data is in the form of data on questions that are often asked by students in the CS section, and 120 questions in the questionnaire which are divided into 10 labels or classes. The test is carried out using 60 conversations that are different from the dataset but have the same purpose. From 60 chatbot conversations, 50 conversations answered correctly and 10 conversations were wrong. The test results show good results, namely having a modeling accuracy of 98% and 84% test data.

Keyword : *Chatbot; Natural Language Processing; Machine Learning; Supervised Learning; Multinomial Naïve Bayes.*

Abstrak

Keberadaan petugas *Customer Service (CS)* yang jumlahnya tidak berimbang dengan jumlah mahasiswa yang dilayani, mengakibatkan tingkat kepuasan pelayanan kampus menjadi berkurang. Penelitian ini bertujuan untuk membuat aplikasi *chatbot* yang bisa membantu tugas CS. Aplikasi *chatbot* berbasis teks ini dibuat dengan pendekatan *Natural Language Processing* (NLP) dan algoritme *Machine Learning*. NLP digunakan untuk memproses sebuah pertanyaan teks dari pengguna, sedangkan algoritma *machine learning MultinomialNB* digunakan untuk mencari data yang sesuai. Jika didapatkankan maka sistem akan menjawab pertanyaan tersebut berdasarkan label yang didapatkan dari model *machine learning*. Dataset yang digunakan sebagai data latih *Chatbot* berupa data pertanyaan yang sering ditanyakan mahasiswa di bagian CS, dan 120 pertanyaan dalam kuisioner yang dibagi kedalam 10 label atau kelas. Pengujian dilakukan dengan menggunakan 60 percakapan yang berbeda dengan dataset tetapi mempunyai maksud yang sama. Dari 60 percakapan *Chatbot* berhasil menjawab dengan tepat sebanyak 50 percakapan dan jawaban salah sebanyak 10 percakapan. Hasil pengujian menunjukkan hasil yang baik yaitu mempunyai akurasi pemodelan 98% dan data test 84%.

Kata kunci: *Chatbot; Natural Language Processing; Machine Learning; Supervised Learning; Multinomial Naïve Bayes.*

1. Pendahuluan

Kemajuan teknologi informasi membuat banyak industri menerapkan otomatisasi pada pekerjaan, begitu juga dalam industri pendidikan. Salah satu teknologi otomatisasi yang digunakan adalah *chatbot,* yang merupakan suatu program kecerdasan buatan yang berbentuk simulasi percakapan interaktif antara mesin dengan manusia melalui teks, suara dan visual atau gambar. Dalam mengenali dan memberikan respon layaknya seperti percakapan manusia. *Chatbot* sangat bergantung dengan data pengetahuan yang sudah

dibuat atau masukan dari pengembang sistem. Untuk meningkatkan akurasi sistem dapat menggunakan cara pemberian label pada setiap pola kalimat pada dataset [1]. Penggunaan *chatbot* dapat membantu pegawai maupun mahasiswa dalam hal layanan informasi secara cepat tanpa terkendala dengan waktu jam kerja.

Kampus Universitas Pelita Bangsa merupakan kampus dengan perkembangan jumlah mahasiswa yang meningkat tiap tahunnya. Jumlah mahasiswa ini tidak sebanding dengan jumlah pelayanan pada bagian informasi (*Customer service*). Sebagian besar, layanan akademik dalam penyampaian informasi oleh *customer service* tidak memiliki waktu banyak dan tidak melayani pengunjung selama 24 jam, hal ini karena keterbatasan waktu kerja yang telah ditetapkan dan hanya berpaku selama jam kantor berlangsung. Hal tersebut mengakibatkan tingkat kepuasan pelayanan terhadap mahasiswa berkurang. Solusi yang ditawarkan yaitu memanfaatkan teknologi kecerdasan buatan salah satunya adalah aplikasi *chatbot.* Dengan aplikasi *chatbot* ini memungkinkan mahasiswa dapat dilayani dengan mesin. Pertanyaan yang diajukkan oleh mahasiswa akan dijawab secara otomatis oleh *chatbot* berdasarkan pengetahuan yang diberikan kepada mesin.

Pada penelitian sebelumnya yang berjudul Algoritma *Artificial Neural Network* pada *Text-based Chatbot Frequently Asked Question* (FAQ) Web Kuliah Universitas Nasional oleh Feri Mustakim, hasil akurasi yang dihasilkan menggunakan metode ANN rata-rata sebesar 72.193% [2].

Fokus penelitian kami adalah bagaimana mengembangkan sebuah *Chatbot* dengan algoritme *Multinomial Naive Bayes*. Penelitian ini akan menggunakan dataset pertanyaan yang sering ditanyakan oleh mahasiswa dan diubah dalam bentuk file JSON.

2. Tinjauan Pustaka

Penelitian yang telah dilakukan mengenai pemanfaatan algoritma *natural language processing* dan *Machine Learning* untuk membuat aplikasi *chatbot* sudah banyak dilakukan sebelumnya. Diantaranya yaitu Implementasi *Natural Language Processing* Dalam Pembuatan *Chatbot* Pada Program Information Technology Universitas Surabaya yang dilakukan oleh Vincentius Riandaru Prasetyo [3]. Pada penelitian ini, validasi sistem dilakukan dengan dua metode yaitu *cross validation* dan *user validation*. Berdasarkan validasi dengan metode cross validation didapatkan akurasi sebesar 83,33%. *User validation* dilakukan dengan cara meminta 10 user untuk melakukan uji coba sistem dan didapatkan akurasi sebesar 76%.

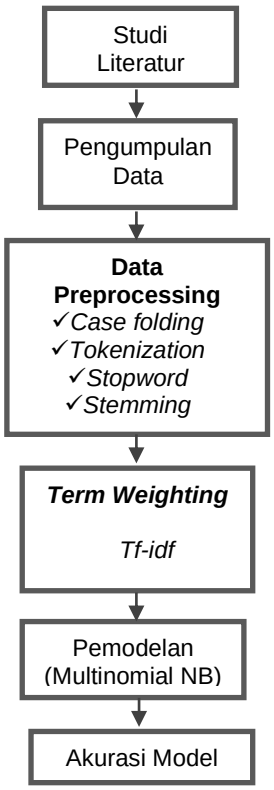
Penelitian dengan topik yang sama juga pernah dilakukan oleh Kristian Adi Nugraha dan Danny Sebastian dengan judul *Chatbot* Layanan Akademik Menggunakan *K-Nearest Neighborhood* [4]. *Chatbot* diimplementasikan untuk mnenggantikan solusi FAQ (*Frequently Asked Question*). FAQ dengan pertanyaan dalam jumlah besar seringkali membuat pengguna kesulitan saat mencari daftar pertanyaan yang sesuai dengan pertanyaan yang akan ditanyakan sehingga solusi yang ditawarkan adalah aplikasi *chatbot.* Algoritme yang digunakan adalah KNN dengan nilai awal K=3 mengghasilkan akurasi 53.48%.

Rena Cahya Hutama dengan judul penelitian aplikasi chatbot berbasis teks menggunakan algoritma naive bayes classifier FAQ grabads [5], meneliti penggunaan aplikasi *chatbot* untuk menangani pertanyaan-pertanyaan umum para pegawai baru. Penelitian ini menggunakan algoritme *Naive Bayes classifier* dengan split ratio sebesar 0,8 dan total 60 pertanyaan maka dihasilkan nilai akurasi sebesar 93,33 % dan nilai kesalahan sebesar 6,66 %.

Pada prinsipnya semua penelitian [3], [4] dan [5] memiliki kesamaan dengan penelitian yang kami lakukan, yaitu kesamaan memprediksi kelas jawaban dari chatbot dan tahapan *preprocessing*. *State of art* pada penelitian kami terletak pada model yang digunakan adalah *Naive Bayes Multinomial*, dengan subjektifitas yang digunakan *multi class,* juga penambahan tahap *preprocessing Stemming.*

3. Metodologi

Prosedur Penelitian dilakukan seperti pada Gambar 1.



Gambar 1 Alur Kerangka Penelitian

1) Studi Literatur

Studi literatur yang dilakukan penulis adalah mambaca sumber-sumber tertulis seperti jurnal ilmiah dari penelitian terdahulu dan referensi lainnya yang berguna sebagai dasar acuan melakukan penelitian yang dilakukan sekarang.

2) Pengumpulan Data

Dalam penelitian ini, data yang digunakan adalah kumpulan data dari pertanyaan-pertanyaan yang sering ditanyakan oleh mahasiswa kepada petugas *customer service* (CS).

Bentuk data yang didapatkan adalah data pertanyaan dan jawaban. Data yang didapatkan sebanyak 120 data. Sampel data disajikan pada Tabel 1.

Tabel 1 Contoh data

No	Pertanyaan	Jawaban
1	Fakultasnya apa saja?	Fakultas Teknik,Fakultas Ekonomi dan Bisnis, Fakultas Ilmu Pendidikan dan Humaniora dan Fakultas Agama Islam
2	Apa saja prodinya?	Banyak kak, tergantung fakultasnya, Bisa dilihat di official website https://www.pelitabangsa.ac.id/'
3	Berapa biaya kuliahnya?	untuk biaya satu semester sekitar 450 ribu per bulan
4	Fasilitas kampus apa saja?	Kami punya fasilitas 2 gedung bertingkat 3 dan 6 dengan eskalator dan lift, kantin, perpustakaan dan lainnya
5	Bisa sambil kerja?	Kita ada kelas reguler dan karyawan
6	Alamatnya dimana	Jl. Raya Inspeksi Kalimalang, Tegal Danas Cikarang Pusat , Kab. Bekasi Cikarang, Bekasi

Tahap berikutnya adalah analisis data untuk setiap pertanyaan untuk dikategorikan sebagai label/kelas yang kita inginkan, seperti pada Tabel 2. Selain itu, ditambahkan data-data percakapan umum.

Tabel 2 Contoh data dengan label

No	Pertanyaan	Jawaban	Label
1	Fakultasnya apa saja?	Fakultas Teknik,Fakultas Ekonomi dan Bisnis, Fakultas Ilmu Pendidikan dan Humaniora dan Fakultas Agama Islam	Fakultas
2	Apa saja prodinya?	Banyak kak, tergantung fakultasnya, Bisa dilihat di official website https://www.pelitabangsa.ac.id/'	Prodi
3	Berapa biaya kuliahnya?	untuk biaya satu semester sekitar 450 ribu per bulan	Biaya
4	Fasilitas kampus apa saja?	Kami punya fasilitas 2 gedung bertingkat 3 dan 6 dengan eskalator dan lift, kantin, perpustakaan dan lainnya	Fasilitas
5	Bisa sambil kerja?	Kita ada kelas reguler dan karyawan	Kelas
6	Alamatnya dimana	Jl. Raya Inspeksi Kalimalang, Tegal Danas Cikarang Pusat , Kab. Bekasi Cikarang, Bekasi	Alamat
7	nama kamu siapa?	Namaku Sinta, salam kenal kak!	nama
8	Ok makasih	Sampai jumpa lagi yaa	bye
9	Selamat Pagi	Halo!	salam
10	bayar berapa satu semester	untuk biaya satu semester sekitar 450 ribu per bulan	Biaya

Data pada Tabel 2 dijadikan satu file JSON dengan struktur seperti pada Tabel 3.

Tabel 3 Struktur Data dataset

No	Atribut	Keterangan
1	Labels	Label atau kelas dari kelompok pertanyaan
2	patterns	Daftar pertanyaan yang mungkin ditanyakan
3	responses	Daftar jawaban pertanyaan

3) *Preprocessing*

Preprocessing adalah proses pengolahan text sebelum digunakan oleh algoritma *machine learning*. Proses *preprocessing* pengolahan teks menggunakan teknik dari *Natural Language Processing* (NLP) atau disebut dengan pengolahan bahasa alami manusia. Teknik ini menerjemahkkan Bahasa manusia menjadi bahasa yang dimengerti oleh komputer. Teknik yang digunakan meliputi:

a) *Case Folding*
Mengubah semua data yang digunakan sebagai masukan (*corpus* atau *dataset*) menjadi huruf besar atau kecil. Ini akan menghindari kesalahan penafsiran kata yang salah jika dieja dengan huruf besar atau kecil.

b) *Tokenization*
Proses pemisahan teks menjadi token(kata) berdasarkan spasi.

c) *Stopword*
Menghilangkan kata-kata yang tidak diperlukan atau tidak mempunyai makna berdasarkan stop lists yang didefiniskan.

d) *Stemming*
Merupakan proses dalam menemukan kata dasar dari suatu kata. Stemming sendiri berfungsi untuk menghilangkan variasi-variasi morfologi yang melekat pada sebuah kata dengan cara menghilangkan imbuhan-imbuhan pada kata tersebut, sehingga nantinya di dapat suatu kata yang benar sesuai struktur morfologi bahasa Indonesia yang benar.

4) *Term Weighting*

Term weighting atau pembobotan term dilakukan berdasarkan hubungan antara kata dan dokumen serta frekuensi kemunculannya. Metode pembobotan yang digunakan adalah TF-IDF (*Term Frequency-Inverse Document Frequency*). *Term Frequency* menghitung jumlah kata yang muncul pada dokumen yang ada. *Inverse Document Frequency* menganggap istilah yang jarang muncul dalam dokumen lebih penting daripada yang sering muncul. Oleh karena itu, semakin sedikit kata yang muncul dalam dokumen, semakin penting kata tersebut dan semakin tinggi nilai IDF [6]. Pembobotan term dengan TF-IDF dihitung berdasarkan Persamaan 1.

$$tf.idf = tf\ x\ idf \qquad (1)$$

Di mana tf adalah *term frequency*, yaitu kemunculan suatu term/token pada dokumen tertentu. Sedangkan idf adalah *inverse document frequency*, yaitu nilai log basis 10 dari jumlah N dokumen dibagi nilai df. Variabel df adalah frekuensi dokumen yang merupakan jumlah dokumen yang memiliki term tertentu [7]. Oleh karena itu, IDF dapat dihitung berdasarkan Persamaan 2.

$$idf = log\ N/df \qquad (2)$$

5) Pemodelan

Algoritma *Multinomial Naive Bayes* merupakan salah satu metode pembelajaran probabilistik didasarkan pada teorema Bayes yang digunakan dalam *Natural Language Processing* (NLP). Algoritma ini bekerja pada konsep *term frequency* yang berarti berapa kali kata tersebut muncul dalam sebuah dokumen. Model ini menjelaskan dua fakta yaitu apakah kata tersebut muncul dalam sebuah dokumen atau tidak serta frekuensinya kemunculan dalam dokumen. *Multinomial Naïve Bayes* dapat diformulasikan Persamaan 3. [8]

$$\mathrm{P(p|n)} \propto \mathrm{P(p)} \prod_{1 \le k \le nd} \mathrm{P(tk\,|p)} \qquad (3)$$

dimana *P(tk|p)* : probabilitas munculnya dokumen text (tk), *n* adalah jumlah dokumen dan *p* adalah polaritas. Kemudian untuk menghitung polaritasnya atau dokumen yang mempunyai kemiripan dirumuskan sebagai berikut, persamaan 4

$$\mathrm{P(tk \mid p)} = \frac{count(tk|p)+1}{count(tp)+|V|} \qquad (4)$$

Dimana *(tk | p)* adalah jumlah *tk* muncul di dokumen text yang memiliki polaritas *p* dan jumlah *(tp)* berarti jumlah token yang ada di artikel berita dengan polaritas *p*.

6) Akurasi Model

Dalam melakukan pengujian terhadap data latih, perlu dilakukan evaluasi terhadap hasil prediksi. Dalam makalah ini, penliti menggunakan *Confusion Matriks* untuk menguji hasil testing dengan memperhatikan akurasi, presisi, dan recall. Untuk menghitung akurasi pada confusion matriks menggunakan persamaan 5 [7].

Accuracy = TP + TN / TP + TN + FP + FN (5)

TP: true positive, FN: false negative, dan FP: false positive dan TN: true negative

4. Hasil dan Pembahasan

1) Pengumpulan Data

Dataset yang digunakan pada penelitian didapatkan dari hasil interview dan pengisian *google form* dengan subjek petugas *customer service* dan bagian akademik. Data tersebut dianalisa untuk setiap pertanyaan untuk dikategorikan sebagai label/kelas yang diinginkan dan disimpan dalam file bertipe JSON.

```
{"data": [
   { "label": "salam",
     "patterns": ["Hai", "Hi", "Halo", "Apa Kabar", "Selamat Pagi", "Selamat Siang", "Selamat Malam", "Salam"],
     "responses": ["Halo!", "Hai", "Halo, ada yang bisa saya bantu?", "Halo selamat datang", "Hai Kawan"]
   },
  {"label": "nama",
```

"patterns": ["nama kamu siapa?", "lu siapa?", "siapa sih lo?", "lu sape ?", "nama lo sape dah ?", "nama?"],
"responses": ["Namaku Sinta, salam kenal kak!", "Halo, aku CS UPB!", "Kenalin, aku Sinta!"]
}
]}

Dari dataset tersebut yang digunakan untuk proses modeling adalah *labels* dan *patterns*. Data *responses* digunakan untuk menjawab terhadap pertanyaan yang sesuai dengan label hasil dari model dan diambil secara acak.

2) *Preprocessing*

Tahapan *preprocessing* yang dilakukan terhadap dataset adalah *Case Folding, Tokenization, Stopword dan Stemming.* Pada proses ini data *chat* "*Semoga harimu menyenangkan*" akan melewati beberapa proses *text preprocessing* yaitu *case folding* merubah semua teks menjadi hurup kecilnya menjadi "s*emoga harimu menyenangkan"*. Proses tokenisasi, yaitu memilah setiap kata berdasarkan spasi menjadi "s*emoga-harimu-menyenangkan*". Proses *stopword* yaitu menghilangkan kata yang tidak penting menjadi "s*emoga-harimu-menyenangkan"* (tidak ada kata yang dihilangkan). Dan proses *stemming* yaitu merubah menjadi kata dasar, menjadi "*moga-hari-senang*". Algoritma *stemming* yang peneliti gunakan adalah librari *Sastrawi Stemmer.* Hasil akhir dari hasil *preprocessing* digabungkan lagi menjadi kalimat, disimpan dalam kolom *text_prep*.

Tabel 4 Sample dataset sebelum dan sesudah *preprocessing*

	chat	label	text_prep
12	Dah	bye	dah
14	Semoga harimu menyenangkan	bye	moga hari senang
53	ada kelas karyawan?	kelas	ada kelas karyawan
24	ini siapa?	nama	ini siapa
54	kelas?	kelas	kelas
35	fakultas?	fakultas	fakultas
22	nama lo sape dah ?	nama	nama lo sape dah
27	alamat kampus?	alamat	alamat kampus
7	Salam	salam	salam
44	berapa biaya kuliahnya?	biaya	berapa biaya kuliah

3) *Bag of words (BoW)*

```
array(['ada', 'alamat', 'apa', 'bayar', 'berapa', 'biaya', 'bisa', 'bye',
       'daah', 'dadah', 'daftar', 'dah', 'di', 'fakultas', 'fasilitas',
       'hai', 'halo', 'hari', 'hi', 'info', 'jumpa', 'kabar', 'kak',
       'kampus', 'kamu', 'karyawan', 'kelas', 'kerja', 'kuliah', 'lagi',
       'lo', 'lu', 'makasih', 'malam', 'mana', 'mau', 'min', 'moga',
       'nama', 'nih', 'ok', 'pagi', 'ping', 'prodi', 'prodinya', 'saja',
       'salam', 'sambil', 'sampai', 'sape', 'satu', 'selamat', 'semester',
       'senang', 'siang', 'siapa', 'sih', 'syarat', 'tinggal', 'untuk'],
      dtype=object)
```

Gambar 2. Fitur pada dataset / BoW (Bag of Words)

Bag of words adalah kumpulan kata-kata unik yang didapatkan dari semua data *text_prep.* Hasil *Bag of words* menjadi fitur bagi setiap data chat. Contoh

Tabel 5. Format dataset

Chat/fitur	apa	alamat	berapa	biaya	kuliah	salam	bayar	....
apa syarat pendaftarannya?								
apa saja prodinya?								
biaya kuliahnya berapa?								
ada kelas karyawan?								
ada fakultas apa?								

4) *Term Weighting* (Pembobotan kata)

Metode pembobotan yang digunakan adalah TF-IDF (*Term Frequency - Inverse Document Frequency*).

Bobot Fitur "*ada*" pada data "*ada fakultas apa*" adalah 2.660 dengan perhitungan sebagai berikut :

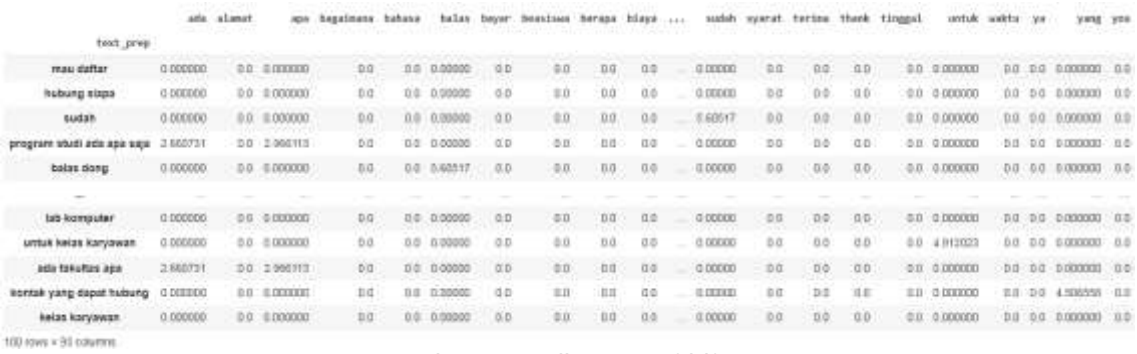
$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right) + 1$$

Pada *library sklearn feature_extraction.text TfidfVectorizer* perhitungan bobot, ditambah 1.

$Tf_{i,j}$ adalah kemunculan kata "*ada*" pada data "*ada fakultas apa*" yaitu sebanyak 1. Sedangkan

Log(N/df_i) adalah log dari jumlah data dibagi kemunculan data fitur "ada" pada keseluruhan dataset sebanyak 19 . Sehingga :

$W_{i,j}$ = 1 x log(100 /19)+1 = 2.660



Gambar 3. Hasil proses tf-idf

5) Pemodelan dan Evaluasi

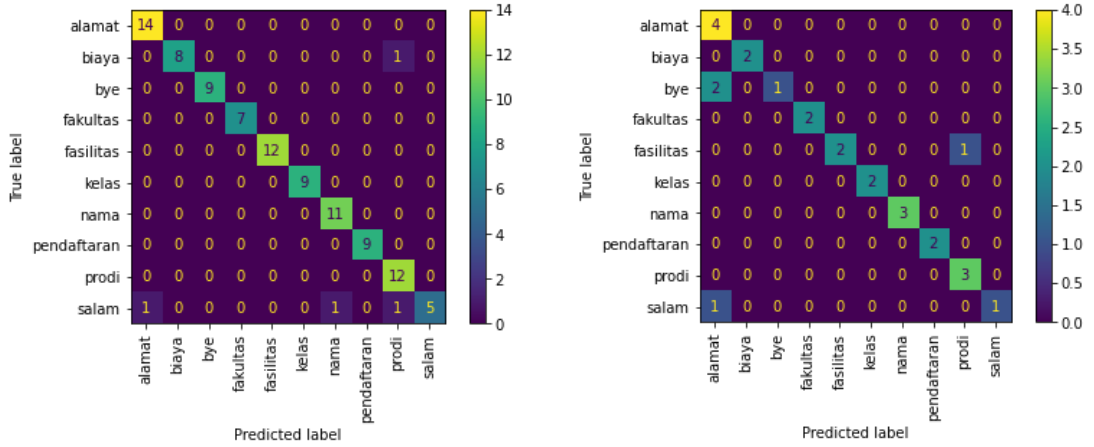
Dataset dibagi menjadi dataset latih dan dataset test denga pembagian 80% dataset latih dan 20% dataset test. Penerapan model *stratify* untuk menjamin pembagian data setiap kelas secara proposional.

Tabel 6 *stratify* Kelas

No	Kelas	Data latih	Data Test
1	alamat	14	4
2	biaya	9	2
3	bye	9	3
4	fasilitas	12	3
5	fakultas	7	2
6	salam	8	2
7	kelas	9	2
8	nama	11	3
9	pendaftaran	9	2
10	prodi	12	3
	T O T A L	100	26

Proses training model menggunakan *library sklearn Naive bayes multinomial* dengan jumlah data yang dilatih sebanyak 100 data. Setelah proses latih selesai dilakukan *testing* model dengan jumalh data yang ditest sebanyak 26 data. Adapun hasil akurasi model dengan menggunakan data *train* adalah 96% dan *test* adalah 84%.

Nilai akurasi untuk masing-masing didapat dari hasil evaluasi model menggunakan *Confusion Matrix.* Gambar 4 adalah hasil evaluasi model menggunakan *Confusion Matrix* untuk data training dan *test.*



Gambar 4 *Confusion Matrix data training dan test*

train_acc =TP/TP+N; TP: True positif. N :Jumlah data
train_acc =14+8+9+7+12+9+11+9+12+5/100
= 96/100=0.96 = 96%

Evaluasi model pada data training hanya gagal 1 pada data kelas “salam” tetapi dikategorikan sebagai kelas “alamat”, sisanya berhasil semua.

test_acc =TP/N; TP: True positif. N :Jumlah data
test_acc =4+2+1+2+2+2+3+2+3+1/26
= 22/26=0.84= 84%

Evaluasi model data *testing*, pada kelas “*bye”* model hanya bisa menjawab benar 1 dan salah 2 dengan menjawab kategori “alamat”. Pada kelas “fasilitas” model berhasil menjawab benar 2 dan salah 1 dengan menjawab kelas “prodi”. Sedangkan pada kelas “salam” model tidak berhasil menjawab benar. Untuk kelas lainnya model berhasil menjawab benar semua. Dengan hasil akurasi model pada penelitian ini sebesar 96%, menguatkan hasil penelitian sebelumnya[5] yaitu 93.33%.

5. Simpulan

Akurasi yang dihasilkan algoritma *Naive Bayes Multinomial* pada data teks menghasilkan akurasi 96% pada data *train* dan 84% pada data *test*. Dengan melihat perbandingan hasil tersebut, hasil dari model Multinomial Naive Bayes adalah mendekati *overfitting*. Hal ini disebabkan data yang kurang variatif dan data latih terlalu sedikit. Hasil model lebih baik jika dibandingkan algoritma *clustering* dan lebih baik juga dibandingkan dengan algoritma *naive bayes classifier* pada kasus yang sama. Dengan hasil akurasi yang dihasilkan, maka model dapat digunakan untuk membuat sebuah chatbot aplikasi web maupun mobile untuk menggantikan atau membantu peran *customer service* dalam melayani mahasiswa.

Model bisa lebih ditingkatkan lagi nilai akurasi dengan melatih banyak lagi data dan variatif data yag dilatih.Juga perlu dicoba ekstraksi teks lainnya seperti *CountVectorizer* dan variasi kata menggunakan *N-Gram*.

Daftar Referensi

[1] Wu, B., Wang, B. and Xue, H. "*Ranking responses oriented to conversational relevance in chat-bots*". In Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, pp. 652-662. Osaka, Japan, 2016

[2] F. Mustakim, Fauziah & N. Hayati. "Algoritma *Artificial Neural Network* pada *Text-based Chatbot Frequently Asked Question* (FAQ) Web Kuliah Universitas Nasional". *Jurnal JTIK (Jurnal Teknologi Informasi dan Komunikasi),* vol. 5, no. 4, pp. 438-446, 2021

[3] V.R. Prasetyo, N. Benarkah & V.J. Chrisintha. "Implementasi Natural Language Processing Dalam Pembuatan Chatbot Pada Program Information Technology Universitas Surabaya". *TEKNIKA*, vol. 10, no. 2, pp. 114-121, Juli 2021

[4] K.A. Nugraha, D. Sebastian. "Chatbot Layanan Akademik MenggunakanK-Nearest Neighborhood". *Jurnal Sains dan Informatika,* vol. 7, no. 1, pp. 11-19, Juni 2021

[5] R.C. Hutama, Fauziah &, R.T. Komalasari. "Aplikasi Chatbot Berbasis teks menggunakan Algoritma Naive Bayes Classifier Faq Grabads". *STRING (Satuan Tulisan Riset dan Inovasi Teknologi)*, Vol. 6 No. 1, pp. 90-97, Agustus 2021

[6] S. Qaiser, R. Ali, "Text Mining: Use of TF-IDF to Examine the Relevance of Words to Documents". *International Journal of Computer Applications*, Vol. 18 no. 1, pp. 25-29, July 2018

[7] H. Christian, M.P. Agus & D. Suhartono, "Single Document Automatic Text Summarization Using Term Frequency-inverse Document Frequency (TFIDF)*". ComTech: Computer, Mathematics and Engineering Applications*, Vol. 7, no. 4, pp. pp. 285-294, 2016

[8] G. Singh, B. Kumar, L. Gaur, and A. Tyagi, "Comparison between Multinomial and Bernoulli Naïve Bayes for Text Classification," *Int. Conf. Autom. Comput. Technol. Manag. ICACTM 2019*, pp. 593–596, 2019

[9] M.A. Hakim & S. Nurhayati, "Pembangunan Aplikasi Chatbot Midwify sebagai Media Pendukung Pembelajaran Ilmu Kebidanan Berbasis Android di Stikes Bhakti Kencana Bandung. Komputika", *Jurnal Sistem Komputer*, Vol. 8, no. 1, pp. 45-52, 2019

[10] E.A. Lisangan, "Natural Language Processing Dalam Memproses Informasi Akademik Mahasiswa Universitas Atma Jaya Makassar". *Jurnal TEMATIKA*, Vol. 1, no. 1, pp. pp. 1-9, 2021

[11] A.B. Rianto, E.P. Mutiara, Wibowo, and P. I. Santosa, "Improving the accuracy of text classification using stemming method, a case of non-formal Indonesian conversation," *Journal of Big Data*, vol. 8, no.1, pp. 1-16, 2021

[12] Yuyun, N. Hidayah, S. Sahibu "Algoritma Multinomial Naïve Bayes Untuk Klasifikasi Sentimen Pemerintah Terhadap Penanganan Covid-19 Menggunakan Data Twitter", *Jurnal Resti*, V ol. 5 No. 4, pp 820 - 826 , 2021

[13] Xu. Shuo, Li. Yan & Z. Wang," Bayesian Multinomial Naïve Bayes Classifier to Text Classification", *International Conference on Multimedia and Ubiquitous Engineering*, pp 347–352, 2017

[14] M. Abbas, K.A. Memon, A.A. Jamali, S. Memon, A. Ahmed." Multinomial Naive Bayes Classification Model for Sentiment Analysis", *IJCSNS International Journal of Computer Science and Network Security*, VOL.19 No.3, pp. 62-71, March 2019

[15] L. Dey, S. Chakraborty, A. Biswas, B. Bose, and S. Tiwari, "Sentiment Analysis of Review Datasets Using Naïve Bayes' and K-NN Classifier," *Int. J. Inf. Eng. Electron. Bus*., vol. 8, no. 4, pp. 54–62, 2016.

[16] A.A. Farisi, Y. Sibaroni, S.A. Faraby. ”Sentiment analysis on hotel reviews using Multinomial Naïve Bayes classifier” *Journal of Physics: Conference Series*, Volume 1192, no. 1, p. 012024, 2019