Klasifikasi Data Percakapan dan Pencarian Pasangan Jawaban dengan Naive Bayes untuk Chatbot Youme

Muhammad Husain | 1301153626

BAB I

Abstrak

Chatbot saat ini sedang populer sebagai teman atau partner chatting. Sebagai partner chatting, kebanyakan Chatbot di design agar meniru bagaimana manusia berperilaku saat berbicara, maka dari itu Chatbot biasanya digunakan di sistem dialog untuk memenuhi berbagai macam tujuan seperti layanan pelanggan, pengambilan informasi dan tanya jawab. Untuk memenuhi semua tujuan ini, perlu dilakukan beberapa proses, salah satu diantaranya adalah proses klasifikasi pertanyaan yang diberikan oleh pengguna terhadap pertanyaan latih yang sudah di klasterisasi sebelumnya. Dalam klasifikasi text, terdapat banyak sekali algoritma yang sering digunakan, diantaranya adalah Support Vector Machine, Hidden Markov Model, Naive Bayes, Neural Networks dan k-Nearest Neighbour. Dari studi literatur yang sudah dilakukan, algoritma Naive Bayes sangatlah cocok digunakan untuk data yang sangat banyak, dikarenakan waktu latih yang dibutuhkannya relatif kecil. Selain itu Naive Bayes juga baik digunakan untuk klasfikasi text dikarenakan setiap di kata di perlakukan unik sebagai fitur, sehingga kata-kata yang ada sangat banyak ragamnya akan membuat banyak fitur.

Latar Belakang

Chatbot saat ini sedang populer sebagai teman atau partner chatting. Sebagai rekan chatting, kebanyakan Chatbot di di desain agar dapat meniru bagaimana manusia berperilaku saat berbicara, agar pengguna merasa sedang berbicara dengan manusia lainnya. Maka dari itu chatbot biasanya digunakan di sistem dialog untuk memenuhi berbagai macam tujuan seperti layanan pelanggan, pengambilan informasi dan tanya jawab.

Chatbot dapat dimanfaatkan untuk melakukan pekerjaan layanan pelanggan, karena chatbot dapat memberikan jawaban yang konsisten, cepat dan tidak memiliki emosi yang dapat mempengaruhi jawaban yang diberikan kepada pelanggan. Sehingga chatbot dapat digunakan untuk meningkatkan kepuasan pelanggan.

Selain itu tingkah laku pelanggan yang dulunya lebih sering mengunjungi kantor penyedia layanan untuk konsultasi, komplain maupun mencari informasi sekarang sudah beralih dengan melakukan hal tersebut melalui media sosial, sehingga dengan bantuan chatbot akan sangat membantu admin media sosial untuk menangani hal tersebut.

Proses pembangunan chatbot memiliki proses pokok yang harus dilakukan yaitu proses pengumpulan data latih, praproses, pengelompokkan data atau klasterisasi dan klasifikasi untuk memilih jawaban yang cocok dengan pertanyaan yang ditanyakan.

Pada proposal tugas akhir ini akan dibahas secara khusus mengenai klasifikasi data percakapan terhadap klaster klaster yang sudah ada. Klasifikasi diperlukan agar pertanyaan yang masuk dapat diberikan jawaban yang sesuai secara akurat

Perumusan Masalah

Pencarian pasangan jawaban yang cocok untuk sebuah pertanyaan yang tidak diketahui kelas atau topiknya.

Batasan Masalah

1. Jawaban yang akan digunakan untuk menjawab pertanyaan pengguna diambil secara mentah dari data latih.

Tujuan

Membangun sebuah model yang dapat digunakan untuk klasifikasi dialog dari pengguna ke topik topik yang sudah ada, lalu di pasangkan dengan jawaban yang ada di topik tersebut, sehingga pertanyaan dari user dapat terjawab dengan baik.

Hipotesis

Pertanyaan dari user dapat di klasifikasi dengan akurat sehingga membuat tanya jawab antara bot dan user menjadi natural/menyambung

Rencana Kegiatan

Rencana kegiatan yang akan dilakukan adalah sebagai berikut:

1. Studi Literatur

   Studi literatur adalah tahap pencarian informasi tentang topik yang diangkat, dengan mencari paper, jurnal dan buku referensi yang berkaitan dengan text mining, klasifikasi, Naive Bayes, dan sumber lain yang berhubungan dengan masalah tugas akhir.
2. Analisis dan perancangan sistem

   Sebelum implementasi, dibutuhkan perancangan model sistem yang akan dibuat
3. Implementasi sistem

   Implementasi dilakukan dengan akuisisi data yang sudah ada dan sudah di proses, lalu pengambilan data validasi untuk di klasifikasi dan kemudian dicarikan pasangan jawaban yang tepat. Hasil validasi ini dapat digunakan untuk mengukur akurasi algoritma terhadap data latih.
4. Analisa hasil implementasi

   Analisa dilakukan setelah implementasi selesai. Sistem di uji coba dengan mengklasifikasikan data test yang tidak diketahui topiknya. Setelah hasil pengujian selesai, hasil pengujian akan digunakan untuk analisis dan evaluasi sistem
5. Penulisan laporan

   Pada tahap akhir, semua pekerjaan dan kegiatan tugas akhir akan di dokumentasikan berupa laporan.

Jadwal Kegiatan

No	Kegiatan	Bulan ke-					
		1	2	3	4	5	6
1	Studi Literatur						
2	Pengumpulan Data						
3	Analisis dan Perancangan Sistem						
4	Implementasi Sistem						
5	Analisa Hasil Implementasi						
6	Penulisan Laporan						

BAB II

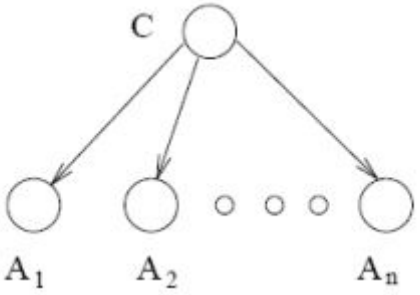
Kajian Pustaka

2.1 Tabel Himpunan Fitur

Tabel Himpunan Fitur merupakan tabel yang dimana atribute atau kolom nya terdiri dari indeks atau id dari pertanyaan, setiap kata yang mungkin dari seluruh data latih, dan kelas dari pertanyaan. Lalu baris atau nilai nya didapat dari jumlah kata tersebut muncul di pertanyaan. Tabel Himpunan Fitur ini nantinya akan digunakan untuk menghitung probabilitas kemunculan kata di sebuah topik.

2.2 Teorema Naive Bayes

Naive Bayes atau dapat di definisikan sebagai Teorema Bayes dengan sebuah asumsi independensi kondisional bahwa semua variabel pada suatu kategori adalah independen secara kondisional satu sama lain.



Gambar 1 menunjukkan jaringan Naive Bayes, dimana C adalah parent node dan A1, …, An adalah parent child dari C. Secara A1, …, An independen secara kondisional dengan satu sama lain. Tujuan utama kita adalah untuk komputasi probabilitas bahwa hipotesis C benar diberikan kondisi bahwa A1, …, An telah di observasi yang dimana adalah *P(C|A1, …, An).* Dengan joint probability, P(C), P(A1|C), …, P(An|C) dapat didapatkan dengan mudah dari data input.

Berdasarkan asumsi Naive Bayes independensi berkondisi, kita memiliki P(Ai|C, Aj) = P(Ai|C), dengan begitu joint probability dapat ditulis ulang sebagai

$$P(C, A_1, \cdots, A_n) = P(C)\prod_{i=1}^{n} P(A_i \mid C)$$

Dan dari formula di atas kita dapat

$$P(C \mid A_1, \cdots, A_n) = \frac{1}{Z} P(C)\prod_{i=1}^{n} P(A_i \mid C)$$

dimana Z = P(Cj | A1, A2, ... , An), dalam klasifikasi text menggunakan Naive Bayes, untuk mengklasifikasikan sebuah text ke kategori tertentu, kita memilih Cj yang menghasilkan nilai probabilitas terbesar dari P(Cj | A1, A2, …, An). Nilai probabilitas terbesar ini disebut “most probable target value” yang dinotasikan sebagai Vmap

$$v_{MAP} = \arg\max_{C_j \in C} P(C_j \mid A_1, A_2, \cdots A_n)$$

sehingga menjadi

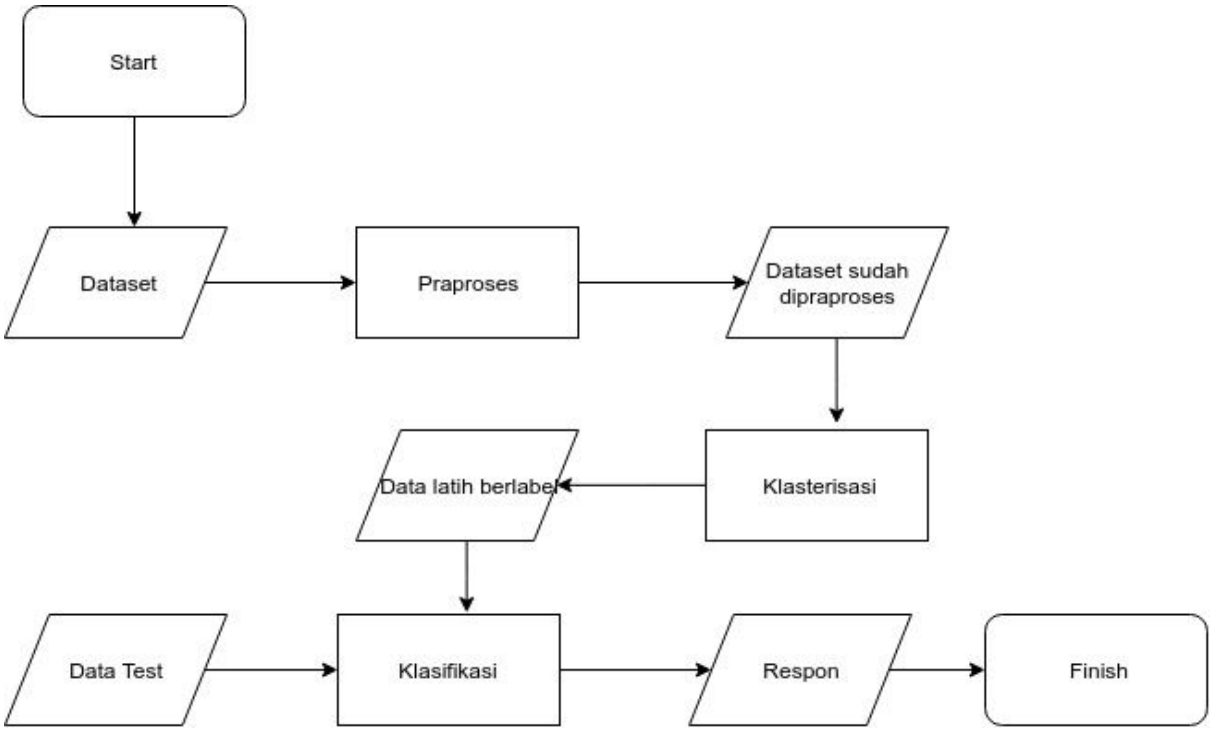
$$v_{NB} = \arg\max_{C_j \in C} P(C_j)\prod_{i} P(A_i \mid C_j)$$

2.3 Additive Smoothing

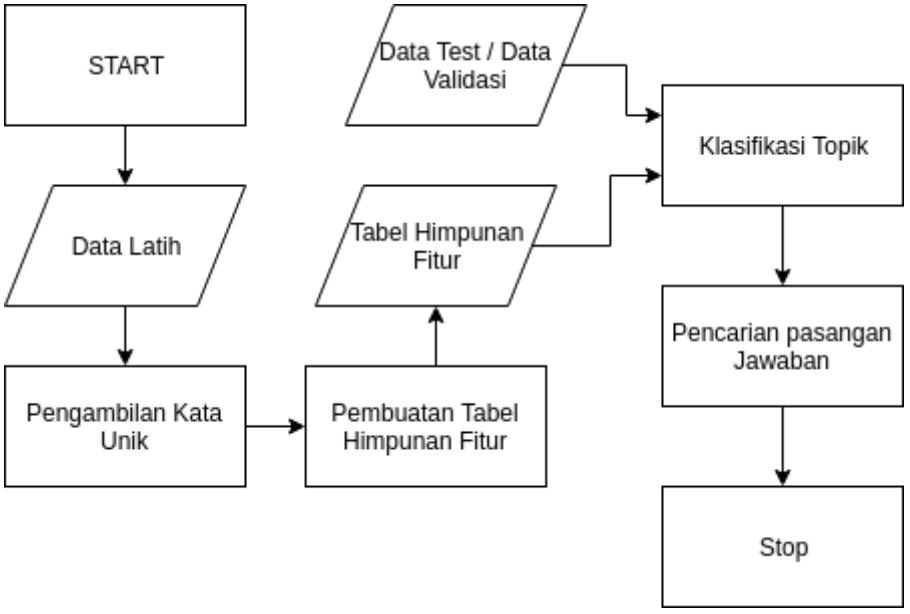
Additive Smoothing (biasa juga disebut sebagai Laplace Smoothing) adalah tehnik smoothing dengan menaikkan semua nilai sejumlah dengan sebuah parameter γ > 0. Penggunaan Additive Smoothing ini digunakan untuk menangani apabila nantinya di data test terdapat sebuah kata yang belum di temui sebelumnya di data latih. Apabila tidak menggunakan Additive Smoothing, hasil komputasi akan menjadi nol, dan akan mengganggu hasil klasifikasi.

BAB III
Metodologi dan Desain sistem

3.1 Flowchart Keseluruhan Sistem



3.2 Flowchart Sistem bagian Klasifikasi



3.3 Penjelasan Flowchart Keseluruhan sistem

Praproses menerima data mentah sebagai inputan yang berupa teks tidak terstruktur, lalu data tersebut dilakukan pembersihan data, penyaringan, penerjemahan, dan penghapusan sehingga menghasilkan data yang bersih, efisien, dan berisi konten yang diharapkan sehingga mudah dan maksimal saat dilakukan klasterisasi dan klasifikasi.

Pada proses klasterisasi, dataset yang sudah dipraproses dilakukan klasterisasi untuk mengelompokkan setiap dataset berdasarkan pertanyaan, pengelompokkan dataset sesuai dengan tingkat kemiripannya sesuai dengan perhitungan Euclidean distance, yang dikelompokkan dengan algoritma k-Means [1], untuk menentukan k yang optimal, setiap hasil klasterisasi divalidasi menggunakan silhouette coefficient. Setelah dataset dikelompokkan maka dataset akan siap digunakan untuk proses klasifikasi.
Proses Klasifikasi akan menerima inputan pertanyaan baru yang tidak diketahui labelnya, tugas besar dari proses klasifikasi adalah memberikan label kepada pertanyaan baru tersebut. Dengan Naive Bayes, akan didapat probabilitas dari data test terhadap setiap pertanyaan di setiap kelas. Nilai probabilitas yang telah didapat akan digunakan untuk menghitung rata-rata probabilitas dari setiap kelas dan pencarian nilai maksimal probabilitas dari setiap kelas. Kemudian nilai rata-rata setiap kelas akan dicari yang tertinggi untuk dijadikan kelas dari pertanyaan baru tersebut. Setelah pertanyaan memiliki label, jawaban dari pertanyaan yang memiliki nilai maksimal dari label yang terpilih akan digunakan sebagai respons kepada pengguna chatbot.

3.4 Penjelasan Flowchart bagian Klasifikasi Sistem

3.4.1 Data Latih

Data Latih yang akan diterima berupa file json dengan format sebagai berikut :

==========
Format Json disini
Inti dari format ada di pertanyaan dan jawabannya
serta kelas / hasil cluster dari Yogi
==========

3.4.2 Pengambilan Kata Unik

Kata unik dari seluruh data latih akan diambil, dijadikan dictionary atau corpus yang kemudian akan digunakan untuk iterasi pencarian frekuensi kemunculan kata di tabel himpunan fitur.

3.4.3 Pembuatan Tabel Himpunan Fitur

Tabel Himpunan Fitur adalah sebuah tabel dimana semua kata yang mungkin dan muncul di data input menjadi atribut di tabel, dan nilainya adalah jumlah kemunculan kata tersebut di dalam data input secara keseluruhan. Selain itu tabel himpunan fitur juga akan memiliki atribut index atau id dari pertanyaan dan kelas(hasil cluster) dari pertanyaan tersebut.

==========
Gambar Sample Data yang sudah dijadikan tabel
==========

3.4.4 Data Test

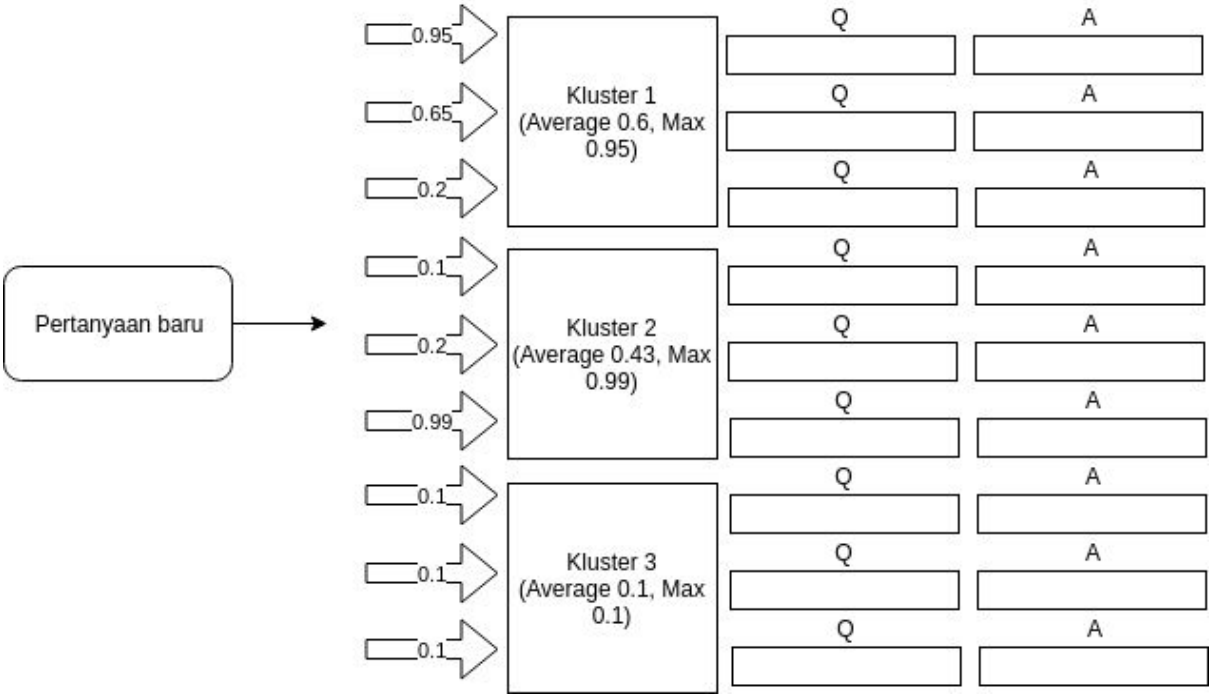
Data test yang digunakan adalah data test yang sebelumnya sudah di preproses layaknya data latih dan memiliki format yang sama.

3.4.5 Klasifikasi berdasarkan Topik

Klasifikasi dilakukan dengan Naive Bayes, dengan formula sebagai berikut

$$v_{NB} = \arg\max_{C_j \in C} P(C_j) \prod_i P(A_i \mid C_j)$$

dimana C adalah setiap pertanyaan yang ada, lalu A adalah setiap kata dalam data test. Dibawah terdapat ilustrasi proses Klasifikasi yang akan dilakukan



Pertanyaan baru atau data test yang masuk, akan di komputasi dengan rumus Naive Bayes terhadap pertanyaan di setiap cluster, yang kemudian hasilnya akan digunakan untuk komputasi rata-rata di kelasnya masing masing. Selain itu, akan disimpan juga nilai maksimal dari setiap kelas. Setelah didapatkan rata-rata dari setiap kelas, akan dipilih kelas dengan rata-rata tertinggi sebagai topik dari pertanyaan baru tersebut.

3.4.6 Pencarian Pasangan Jawaban

Pasangan jawaban dari pertanyaan atau data test ditentukan dengan menggunakan nilai maksimum dari topik atau kelas yang telah terpilih sebagai topik dari pertanyaan. Penggunaan rata-rata dan nilai maksimal ini dilakukan untuk menghindari pertanyaan yang memiliki kata-kata yang kebanyakan sama tetapi memiliki topik yang berbeda, contohnya adalah “Berapa biaya JPG” dan “Berapa biaya JPU” memiliki 2 dari 3 kata yang mirip yang sebenarnya berbeda topik.

Referensi

file:///media/husen/MEDIA/Download/10.1109@GrC.2007.40.pdf
https://sci-hub.tw/https://dl.acm.org/citation.cfm?id=2934737