Chatbot untuk Website Utama UK Petra dengan Hidden Markov Model dan k-Nearest Neighbor untuk Generate Jawaban

Kevin Koesoemo
Program Studi Informatika
Fakultas Teknologi Industri
Universitas Kristen Petra
Jl. Siwalankerto 121-131
Surabaya 60236
Telp. (031)-2983455
yk.terrier@gmail.com

Alexander Setiawan
Program Studi Informatika
Fakultas Teknologi Industri
Universitas Kristen Petra
Jl. Siwalankerto 121-131
Surabaya 60236
Telp. (031)-2983455
alexander@petra.ac.id

Indar Sugiarto
Program Studi Teknik Elektro
Fakultas Teknologi Industri
Universitas Kristen Petra
Jl. Siwalankerto 121-131
Surabaya 60236
Telp. (031)-2983455
indi@petra.ac.id

ABSTRAK

Universitas Kristen Petra (UK Petra) memiliki berbagai media layanan informasi seputar jurusan universitas dan pendaftaran mahasiswa baru, seperti media sosial dan *WhatsApp*. Namun media ini masih bergantung pada operator manusia yang terbatas pada jumlah dan waktu. Dengan adanya *chatbot*, informasi seputar UK Petra dapat diketahui kapan saja. Penelitian *chatbot* oleh S. C. P & Afrianto membutuhkan metode pencocokkan pertanyaan dengan dataset. Penelitian ini menggunakan dua metode yaitu kNN (*k-Nearest Neighbor*) dan HMM (*Hidden Markov Model*) untuk menyelesaikan masalah tersebut. Pada penelitian ini chatbot akan mencoba menggabungkan kedua metode ini serta membandingkannya, dicoba apakah dapat menghasilkan jawaban yang dapat dipahami serta dapat menjawab pertanyaan dengan berbagai tingkat kesulitan.

Metode kNN digunakan sebagai klasifikasi pertanyaan yang diberikan ke *chatbot* kira-kira mendekati dengan pertanyaan pada *knowledge base chatbot*. Metode HMM dipakai untuk merangkai kata-kata jawaban dari *knowledge base* yang terpilih. Hasil *chatbot* akan diuji dari segi validitas jawaban oleh dua responden (Humas dan Admisi kampus) beserta lama waktu yang dibutuhkan untuk menghasilkan sebuah jawaban.

Hasil *chatbot* dengan metode kNN memiliki akurasi jawaban hasil uji sebesar 64,44% (45 pertanyaan), dengan waktu rata-rata 0,08 detik. Sedangkan hasil *chatbot* dengan penggabungan kNN-HMM menghasilkan jawaban yang random dan tidak beraturan, dengan waktu sistem rata-rata 0,12 detik disebabkan HMM merupakan metode berbasis probabilitas.

Kata Kunci: *chatbot, Hidden Markov Model* (HMM), *k-Nearest Neighbor* (kNN), *machine learning, Natural Language Processing* (NLP)

ABSTRACT

Petra Christian University has various services for general information about university majors and student admissions, such as social media and WhatsApp. However, these services still limited by number and working time of operators as human. Therefore, with this chatbot, information about PCU can be found anytime. Chatbot Study by S. C. P & Afrianto needs method to match chatbot question with the dataset. This thesis uses two methods, namely kNN (k-Nearest Neighbor) and HMM (Hidden Markov Model) to solve these problem. In this chatbot, it will try to combine and compare these two methods, and see if it can produces answers that can be understood and in accordance with various difficulty questions given.

The kNN is used as a classification for questions given to chatbot which approximately match with questions on the chatbot's knowledge base. HMM is used to assemble answer words from the selected knowledge base. Chatbot's answers will be tested in terms of validity of the answers by two respondents (Public Relation and Admission staff) also the length of time it takes to produce answers.

The results of the chatbot with kNN has an accuracy of 64.44% (45 questions), with average system runtime of 0.08 seconds. While the results of chatbot with kNN-HMM produces random and irregular answers, with average system runtime of 0.12 seconds, cause by HMM which is a probability based method.

Keywords: *chatbot, Hidden Markov Model* (HMM), *k-Nearest Neighbor* (kNN), *machine learning, Natural Language Processing* (NLP)

1. PENDAHULUAN

Universitas Kristen Petra (UK Petra) menyediakan beberapa layanan informasi seputar kampus bagi khalayak umum berupa situs resmi (*website*), media sosial, dan layanan *chatting* dengan operator melalui *platform WhatsApp* untuk orang tua, maupun mahasiswa dan calon mahasiswa baru dalam mencari informasi seputar UK Petra. Layanan *chatting* disini pada umumnya dilakukan oleh dua individu yaitu antara orang yang bertanya dan operator itu sendiri. Namun, layanan operator sebagai salah satu sumber informasi saja masih dirasa kurang efektif bagi calon mahasiswa ataupun orang tua yang ingin mendapatkan informasi dengan cepat. Apalagi ditambah dengan operator yang hanya dapat melayani pertanyaan dalam jam kerja saja, serta jumlah operator yang sedikit dibandingkan jumlah orang yang bertanya. Hal ini akan menjadi kendala utama bagi layanan *chatting* ini.

Chatbot merupakan merupakan suatu sistem yang dapat membalas pesan yang dikirim oleh pengguna. *Chatbot* berasal atas dua kata, yaitu *chat* dan *bot* [10]. *Chatbot* menjadi cikal bakal berbagai aplikasi berbasis *messenger* seperti *Google, Facebook*, dan *WhatsApp* [1]. Selain itu, *chatbot* dapat membantu meningkatkan responsivitas dan ketersediaan, serta

mengurangi ketergantungan kekuatan manusia di dunia otomatisasi saat ini.

Chatbot memiliki kelebihan dimana pengguna tidak harus memantau aktivitas pesan untuk melakukan interaksi dan kekurangan *chatbot* yaitu terbatasnya topik atau pola pembahasan. Sedangkan kelemahannya adapun seperti respon yang dihasilkan merupakan pencocokan *pattern* pada *knowledge base* [7]. Selain itu, kelemahan lainnya yaitu masih kurangnya pertanyaan-pertanyaan yang belum dimasukkan ke dalam sistem serta belum dapat menganalisis apabila masukan yang diberikan pengguna terjadi kesalahan penulisan [15].

Penelitian ini juga berusaha untuk mengatasi kelemahan-kelemahan yang ada pada *chatbot* yang sudah ada. Pada hasil penelitian *chatbot* Informasi Objek Wisata Bandung [20], terdapat kekurangan dalam hal klasifikasi/pengkategorian lokasi-lokasi wisata, sehingga membuat daftar wisata menjadi tidak rapi. Masalah pengkategorian tersebut, diharapkan dapat diatasi dengan menggunakan metode kNN. Algoritma *K-Nearest Neighbor* adalah salah satu algoritma yang bisa dimanfaatkan untuk implementasi pengklasifikasiannya [16].

Menurut survei Conversational Agents/*Chatbot*s and Design Techniques [18], algoritma *Hidden Markov Model* merupakan metode yang baik dalam membuat *chatbot*, karena dapat memberikan umpan balik/jawaban yang koheren dengan pertanyaan yang diberikan. Namun, metode ini kurang baik dalam menangani pertanyaan yang panjang/kompleks.

Chatbot berbasis *Markov Chain* [23] memiliki tingkat akurasi 65% dari total 20 pertanyaan yang diajukan. Namun, pada penelitian *chatbot* ini masih adanya kesalahan dalam menjawab pertanyaan yang berkaitan tentang topik tertentu (*student admission*), dan penelitian yang dilakukan Siswadi dan Tarigan akan menjadi standar/acuan penelitian yang akan dilakukan ini. Setidaknya ada lima komponen dalam *Hidden Markov Model* untuk penandaan POS/*Part of Speech* [17] Penelitian ini menggunakan sistem IPOSTAgger, penanda POS untuk bahasa Indonesia yang menerapkan metode *Hidden Markov Model*.

Pada penelitian ini *chatbot* berbasis *Natural Language Processing* dan algoritma *k-Nearest Neighbor* akan digunakan sebagai pelayanan informasi bagi UK Petra, dan diharapkan dapat membantu seluruh proses kerja dari layanan informasi bagi calon mahasiswa yang akan mendaftar di UK Petra menjadi efisien dan efektif, serta juga membantu pihak Humas UK Petra dalam hal pelayanan pemberian informasi. Selain itu, penelitian ini juga berusaha untuk menyelesaikan masalah pada penelitian yang dilakukan S. C. P & Afrianto yang membutuhkan metode pencocokkan pada chatbot sebagai *research gap*. Dengan adanya metode kNN, masalah pencocokkan (dalam kasus ini antara pertanyaan *chatbot* dengan *dataset knowledge base*) dapat terselesaikan.

2. DASAR TEORI

2.1 Kecerdasan Buatan *(Artificial Intelligence)*

Kecerdasan buatan/*Artificial Intelligence* merupakan salah satu cabang ilmu komputer/*computer science* yang menggeluti dalam pengembangan mesin/komputer yang sebagaimana dapat melakukan tugas/hal-hal yang dapat dilakukan oleh manusia/yang membutuhkan kecerdasan manusia (*human intelligence*). Kecerdasan buatan juga dianggap sebagai sebuah kegiatan mengkonstruksi artefak yang cerdas/intelijen [9]. Menurut [3], kecerdasan buatan adalah cabang ilmu komputer yang membahas tentang penangkapan, pemodelan, dan penyimpanan kecerdasan manusia ke dalam sebuah teknologi informasi yang nantinya dapat dimanfaatkan untuk pengambilan keputusan.

Kecerdasan buatan berevolusi dengan pengembangan komputer bahkan mengandalkan pengembangan komputasi [5]. Cara dan proses pemikiran manusia yang dikembangkan oleh psikolog dan disambut baik oleh para ahli komputasi menghasilkan ilmu Kecerdasan Buatan. Hal ini berlanjut dengan perkembangan ilmu kognitif yang mendorong pengembangan Kecerdasan Buatan hingga Kecerdasan Berpikir Kognitif, jalur baru menuju ilmu Kecerdasan Buatan yang dapat meniru sebagian kemampuan kognitif manusia, bahkan jika tidak seluruh kemampuan kognitif.

2.2 *Natural Language Processing* (NLP)

Natural Language Processing (NLP) adalah bidang interdisipliner yang mempelajari dan mengembangkan algoritma dan sistem yang memungkinkan komputer untuk memahami dan melakukan tugas-tugas yang melibatkan bahasa manusia [6]. NLP juga merupakan salah satu cabang dari ilmu kecerdasan buatan/AI yang mengeksplorasi bagaimana komputer/mesin dapat dimanfaatkan untuk memahami dan memanipulasi bahasa natural manusia, baik dalam bentuk teks ataupun suara, untuk melakukan banyak hal yang berguna [11]. NLP juga dapat disebut sebagai komputasi linguistik, pidato komputer dan pemrosesan bahasa atau manusia teknologi bahasa.
Natural Language Processing (NLP) merupakan salah satu cabang ilmu AI yang berfokus pada pengolahan bahasa natural [24]. NLP berkaitan dengan analisis bahasa manusia baik dalam bentuk tertulis maupun lisan dan dapat mengekstrak perintah atau informasi yang berguna. Saat ini sudah terdapat banyak pemanfaatan NLP dalam berbagai bidang, seperti mesin penerjemah, aplikasi perangkum/*summarization*, *information retrieval*, *speech recognition*, sistem pakar/*expert system*, dan masih banyak lagi.

2.3 *Text Preprocessing*

Sebelum *dataset* dapat digunakan, diperlukan adanya *text preprocessing*/pengolahan data awal sehingga bentuknya lebih seragam dan rapi. Adapun terdapat empat tahap dalam text preprocessing, yaitu *tokenization, stop-word removal, lowercase conversion*, dan *stemming* [2]. *Tokenizing* merupakan tahap pertama dimana suatu kalimat/kata dipisah untuk dijadikan potongan/disebut *token*, yang nantinya dapat dianalisa. *Lowercase conversion* juga dapat dilakukan dalam tahap awal ini, yang hanya bertujuan untuk mengubah semua token menjadi huruf kecil/*lowercase*.

Tahap selanjutnya yaitu menghapus *stopword*/kata yang sering dipakai dalam suatu tatanan bahasa (misalkan kata sambung seperti “yang, di, ke, dan, untuk, atau” serta tanda baca yang umum dipakai seperti titik (.), koma (,), garis miring (/) dan tanda tanya (?). Tahap akhir dari *text preprocessing* ini yaitu *stemming*/merubah kata yang berimbuhan menjadi kata dasar, misalkan kata “memakan”, “dimakan”, “makanlah” memiliki kata dasar “makan”.

2.4 *Natural Language Toolkit* (NLTK)

Natural Language Processing (NLP) sangat aktif bidang penelitian, yang menarik banyak peneliti setiap tahunnya. Hal ini

memberikan kemungkinan untuk mempelajari bahasa manusia untuk diterapkan, bukan hanya secara teoritis, dan untuk mencoba menyelesaikan beberapa tugas-tugas yang berkaitan dengan bahasa manusia [12].

Toolkit Bahasa Alami/*Natural Language Toolkit* (NLTK) memungkinkan setiap pengembang, bahkan pemula, untuk berkenalan dengan NLP dengan mudah tanpa menghabiskan terlalu banyak waktu untuk belajar atau mengumpulkan sumber daya. NLTK merupakan sebuah program bersifat *open source* yang memiliki modul serta tutorial yang mudah untuk digunakan dalam pengembangan *software* yang berkaitan dengan pengolahan bahasa [14]. NLTK digunakan untuk membangun program perangkat lunak yang bekerja dengan bahasa manusia untuk diterapkan dalam pemrosesan bahasa alami [13]. NLTK juga mencakup representasi grafis dan kumpulan data sampel yang menjelaskan ide-ide yang terkait dengan tanggung jawab pemrosesan bahasa alami yang difasilitasi NLTK.

2.5 *K-Nearest Neighbor (kNN)*

Algoritma *K-Nearest Neighbor (kNN)* adalah algoritma yang digunakan untuk melakukan klasifikasi terhadap suatu objek, berdasarkan *k* buah data latih (*training*) yang jaraknya paling dekat dengan objek tersebut. Algoritma *K-Nearest Neighbor* merupakan metode klasifikasi yang mengelompokan data baru berdasarkan jarak data baru itu ke beberapa data/tetangga (*neighborhood*) terdekat [22]. Syarat nilai *k* adalah tidak boleh lebih besar dari jumlah data *training*, dan nilai *k* harus ganjil dan lebih dari satu. Dekat atau jauhnya jarak data latih yang paling dekat dengan objek yang akan diklasifikasi dapat dihitung dengan menggunakan metode *Eucledian Distance*. Persamaan 1 berikut menunjukkan rumus perhitungan untuk mencari jarak dengan d adalah jarak dan p adalah dimensi data:

$$di = \sqrt{\sum_{i=1}^{p}(x_{2i} - x_{1i})^2} \quad (1)$$

dengan:

x1	= sampel data
x2	= data uji
i	= variabel data
d	= jarak
p	= dimensi data

Ada cara lain untuk mengukur jarak antara dua data selain dengan metode Eucledian, yaitu dengan menggunakan *Cosine Similarity* [19]. Metode *Cosine Similarity* menghitung jarak/keterkaitan antara dua buah data berdasarkan kedekatan sudut, dengan cara mencari hasil *dot product* dari dua data tersebut. Semakin dekat sudut antara dua data tersebut, maka kemungkinan data tersebut mirip/hampir sama, begitu juga sebaliknya. Persamaan 2 berikut menunjukkan rumus *Cosine Similarity:*

$$sim(A, B) = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|} \quad (2)$$

Prinsip kerja *k-Nearest Neighbor* (kNN) adalah mencari jarak terdekat antara data yang akan dievaluasi dengan k tetangga (*neighbor*) terdekatnya dalam data pelatihan [21]. Tujuan algoritma kNN adalah mengklasifikasikan obyek baru berdasarkan atribut dan *training sample.* Diberikan titik *query*, akan ditemukan sejumlah *k* obyek atau (titik training) yang paling dekat dengan titik *query*. Klasifikasi menggunakan voting terbanyak diantara klasifikasi dari *k* obyek.

2.6 *Hidden Markov Model* (HMM)

Hidden Markov Model (HMM) merupakan metode *Machine Learning* yang menggunakan model probalitas berurutan untuk menyelesaikan masalah dan data latih yang digunakan membutuhkan POS-Tagging untuk setiap kata, dan data latih di modelkan kedalam bentuk model *hidden markov* [4]. Metode HMM terbukti sangat fleksibel dalam pemodelan statistik dalam penggambaran berbagai jenis data [8]. Untuk saat ini penerapan metode HMM sudah banyak digunakan dalam ilmu *science* dan *engineering*.

Dalam *Hidden Markov Model* ada pemodelan umum yang dipakai dalam pemodelan permasalahan, Persamaan 3 berikut merupakan permodelan umum yang digunakan dalam *Hidden Markov Model.*

$$\lambda = (A, B, \pi) \quad (3)$$

Pada Persamaan 3 diatas, terdapat symbol lamda (λ) sebagai model Markov, A sebagai probabilitas transisi, B adalah probabilitas emisi, dan symbol phi (π) merupakan probalitas keadaan awal. Dalam penulisan pemodelan sebuah permasahan dengan *Hidden Markov Model* menggunakan lima tuple yaitu:

1. *Observed state* (O) Pada *observed state* di buat dengan simbol O = O1, O2, O3, ..., On *observed state* yaitu pemodelan permasalahan yang dapat diamati.
2. *Hidden state* (Q) *Hidden State* merupakan *state* yang tersembunyi dan tidak dapat diamati, di simbolkan dengan Q = Q1, Q2, Q3, …Qn.
3. Matrik Peluang Transisi (A) Peluang transisi merupakan peluang perpindahan dari state i menuju ke state j. Di simbolkan dengan A = a01, a02, an1, … anm ; aij , banyaknya jumlah matrik peluang transisi yaitu Q x Q.
4. Matrik Peluang Emisi (B) Peluang emisi merupakan peluang perpindahan state i dengan syarat waktu Ot (*Observed State*). Di simbolkan dengan B = bi(Ot) banyaknya jumlah matrik peluang emisi yaitu Q x O.
5. Peluang Keadaan Awal (π) Peluang awal di simbolkan dengan π

3. ANALISIS DAN DESAIN

3.1 Analisis Data

Data yang digunakan pada penelitian ini berupa list/daftar pertanyaan beserta jawaban, berkaitan dengan informasi jurusan di Universitas Kristen Petra/UK Petra (informasi singkat, visi misi, sejarah, kurikulum, profil lulusan, fasilitas, dosen) dan informasi admisi/penerimaan mahasiswa baru (jalur pendaftaran, biaya studi, tes khusus, beasiswa, dll). Kumpulan pertanyaan dan jawaban ini diambil dari dua sumber untuk menambah variasi dari dataset. Mengingat semakin banyak dataset yang didapatkan, maka pengetahuan/*knowledge* yang dimiliki chatbot juga semakin beragam. Jumlah data pertanyaan dan jawaban yang terkumpul (per Mei 2021) sebanyak 1122 (seribu seratus dua puluh dua) buah.

Sumber pertama diambil dari kumpulan teks percakapan/chat dari aplikasi *WhatsApp* yang dimiliki oleh dua pihak, yang pertama dari tim Humas UK Petra (HP: +62 812-3406-7323, per Mei 2021 dipegang oleh Ibu Lady Stefani selaku Admin Humas) dan yang kedua dari tim Admisi UK Petra (HP: +62 812-1776-5265, per Mei 2021 dipegang oleh tim Admisi yang dikepalai oleh Bapak Pratjoyo Kushandoko). Kumpulan teks percakapan diambil dalam rentang waktu antara Juli 2020 hingga

pertengahan Mei 2021 melalui platform WhatApp Web. Sumber kedua diambil dari mengambil informasi singkat mengenai tiap jurusan di UK Petra langsung dari situs web petra.ac.id serta sitemap dari tiap-tiap jurusan (bagian FAQ/diambil setiap halaman situs web secara manual). Sedangkan informasi admisi diambil di situs web admission.petra.ac.id (per Mei 2021).

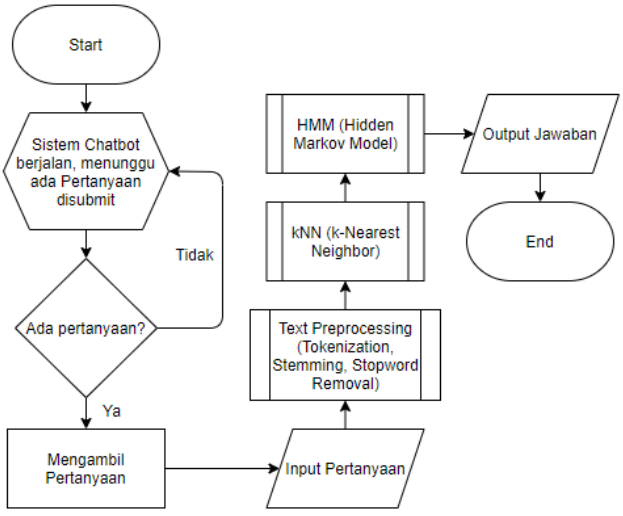
Dari kedua sumber ini, data kemudian dikumpulkan dalam format Excel (.xlsx), dimana kolom pertama berisi daftar pertanyaan yang ada, kemudian diikuti jawaban di kolom berikutnya. Pada kolom ketiga, berisi jawaban kedua yang berfungsi untuk menambah variasi pilihan jawaban. Kolom keempat berisi intent/topik kunci dari masing-masing pertanyaan.

Pertanyaan	Jawaban	Jawaban2	Intent
apa ada japres?	Kami masih membuka pendaftaran Jalur Prestasi	Jalur prestasi masih dibuka kak. Link sele	jalur admisi
uang gedung/uang pangkal jurusan IBM berapa ya?	Jadwal admission 2021/2022 periode SEPTEMBER dapat dibaca di http://admission.petra.ac.id/index.	Untuk uang gedung jurusan IBM dapat di biaya	
uang semester apa sama dengan uang sks?	Info biaya studi admission 2021/2022 dapat dibaca di http://admission.petra.ac.id/index.php/site/page	Untuk biaya uang semester bisa dilihat di pembayaran	
cara pembayaran uang gedung?	Info biaya studi admission 2021/2022 dapat dibaca di http://admission.petra.ac.id/index.php/site/page	Cara pembayaran uang gedung dapat di pembayaran	
cara pembayaran uang semester, uang sks.	Info biaya studi admission 2021/2022 dapat dibaca di http://admission.petra.ac.id/index.php/site/page	Cara pembayaran uang semester, uang : pembayaran	
biaya pendaftaran	Cara pembelian PAC admision 2021/2022, dapat anda baca di http://admission.petra.ac.id/index.ph	Cara pembelian PAC dapat di baca http: biaya	
Informasi pendaftaran untuk Program Studi Bahasa	Pembelian PAC dapat mentransfer uang Rp. 350.000 ke rekening: CIMB NIAGA A.n. YPTK PETRA No	Informasi pendaftaran dapat mentransfer jalur admisi	
Jalur pendaftaran Suarna Tiongkhoa	Jadwal dan Alur admision 2021/2022 periode FEBRUARI dapat dibaca di http://admission.petra.ac.id	Jalur pendaftaran dapat dilihat di http://a jalur admisi	
ini cara login di sim petra gimna yaah?	Langkah Daftar (Jang/Delima9 setelah pembayaran USM pertama ke konfirmasi 1. Silakan login dulu di	Cara login di SIM Petra dapat dilakukan s admisi	
Untuk pembelian pac disini ya kak?	Pembelian PAC dapat mentransfer uang Rp. 350.000 ke rekening: CIMB NIAGA A.n. YPTK PETRA No	Untuk pembelian PAC dapat ditransfer Rp beli PAC	
Sama untuk cara pendaftarannya gimana ya	Jadwal dan Alur admision 2021/2022 periode FEBRUARI dapat dibaca di http://admission.petra.ac.id	Cara pendaftaran dibaca di http://admiss admisi	
Min batas pengumpulan ya kapan min,yang berkas s	http://admission.petra.ac.id/index.php/site/pagesId/53.di poin 6.	Batas pengumpulan dapat dilihat http://a admisi	
Kalau yang di cetak batas nya kapan?	http://admission.petra.ac.id/index.php/site/pagesId/53.di poin 6.	Batas pencetakan dapat dilihat http://ac admisi	
siang kak mau tanya ni kaka	Halo, Selamat Pagi	Siang ada yang bisa dibantu halo	
ini saya sudah kirimsa di ulip jurusan fashion design	Jika ingin pindah prodi, silakan baca terlebih dahulu syarat & ketentuannya, http://admission.petra.ac.	Jika ingin mengajukan perpindahan dapo pindah prodi	
visi,misi,tu ada dua pilihan pri saya yg satunya pilih finar	Jika merujuk ke aturan di atas, Valen harus daftar lagi Manajemen Perhotelan nya Beli PAC lagi, daftar	Bila ingin mengajukan perpindahan harus pindah prodi	
ini yang uspp dan uprs ini pembayarannya gimana yi	Iya itu nanti dibayarkan mulai 9 September 2021	Pembayaran USPP dan UPRS dibayarkan pembayaran	
uspp dan uprs pbayarannya tiap berapa bulan?	Setiap 1 semester sekali (6 bulan sekali)	Pembayaran USPP dan UPRS setiap 1 sx pembayaran	
berapa semester ya ?	Jika normal tepat waktu, 8 semester	Untuk kelulusan jika tepat waktu 8 semes waktu studi	
mau tanya ya kalau pembayarannya terlambat apa yan	Jika terlambat , ada denda. Dan jika terlalu lama hingga batas terentu bisa di ke ikut UTS atau UAS	Jika pembayaran terlambat akan dikenak pembayaran	
Selamat malam, saya mau bertanya untuk teman sa	Kalau mau apply beasiswa peduli maka harus mengulang proses pendaftarannya dari awal. Beli PAC l	Bila ingin apply beasiswa peduli harus me beasiswa	
sama apa perlu mendaftar ulang lagi? kalau sebelun	Informasi Beasiswa dapat dibaca di https://student.petra.ac.id/dana/beasiswa	Untuk melihat informasi Beasiswa dapat r beasiswa	
Chi ke, kalau sudah beli pac lagi beasiswa dari jalur	Iya benar	Iya benar masih bisa diklaim beasiswa	
Halo saya mau tanya, saya ada keterlambatan pemb	Selamat sore, Belum ada denda, silakan melakukan pembayaran terima kasih.	Untuk bulan januari masih belum ada der pembayaran	

Gambar 1. Bentuk Dataset

3.2 Analisis Sistem

Analisis sistem membahas bagaimana input data diproses sehingga dapat menghasilkan output yang sesuai. Sistem mencakup pengenalan pertanyaan yang diajukan, pengolahan dataset, dan memberikan jawaban yang sesuai/mendekati dengan pertanyaan yang ada. Gambar 2 menggambarkan flowchart keseluruhan alur sistem *Chatbot*.



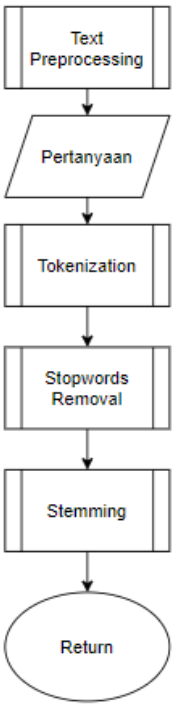
Gambar 2. *Flowchart* Sistem Keseluruhan *Chatbot*

Flowchart pada Gambar 2, sistem akan berjalan di background sambil menunggu menerima adanya inputan berupa teks pertanyaan. Jika ada pertanyaan yang diajukan, maka pertanyaan tersebut akan menjadi input sistem. Pertanyaan akan melewati proses preprocessing terlebih dahulu (*tokenizing*, *stemming stopwords removal*). Kemudian diterapkan metode kNN (k *Nearest Neighbor*) untuk mengklasifikasi jenis pertanyaan yang diajukan mendekati pertanyaan mana dalam dataset, kemudian dilanjutkan dengan metode HMM (*Hidden Markov Model*) untuk merangkai jawaban dari dua macam varian jawaban yang ada berdasarkan dataset yang ada.

3.2.1 Text Preprocessing

Text Preprocessing dilakukan pada pertanyaan yang masuk ke *chatbot*. Hal ini bertujuan untuk mempersiapkan pertanyaan sebagai data input yang akan diterima sistem. Dengan menghilangkan elemen-elemen tidak penting dari data tersebut dan membuat maksud pertanyaan lebih mudah dipahami oleh *chatbot* itu sendiri, diharapkan *chatbot* dapat memberikan jawaban dengan hasil yang maksimal.

Proses *preprocessing* sendiri terdiri dari tiga tahap, meliputi tokenization (mengubah kalimat menjadi per kata, serta mengubah huruf kecil dan menghilangkan spasi yang berlebihan), *stemming* (mengembalikan kata berimbuhan menjadi kata dasar), dan *stopwords removal* (membuang elemen/kata-kata yang seringkali digunakan namun tidak memiliki makna, misal kata penghubung/tanda baca). Seluruh tahap proses *preprocessing* akan dilakukan oleh *library* NLTK dan Sastrawi. Alur text preprocessing secara umum dapat dilihat pada Gambar 3.



Gambar 3. *Flowchart* Proses *Text Preprocessing*

3.2.1.1 Tokenizing

Tokenizing adalah suatu proses untuk memotong dokumen menjadi pecahan kecil yang dapat berupa bab, sub-bab, paragraf, kalimat, dan kata (token). Pada konteks kali ini, pertanyaan yang diajukan ke *chatbot* akan diproses menjadi kata/token. Proses *tokenizing* juga akan mengubah huruf besar menjadi huruf kecil (*case folding*) serta menghilangkan *whitespace*/spasi berlebihan. *Tokenizing* akan memanfaatkan library NLTK.

3.2.1.2 Stopwords Removal

Stopwords removal merupakan bagian *preprocessing* selanjutnya yang mengeliminasi kata/tanda baca yang sering digunakan namun memiliki makna yang sedikit/tidak berhubungan dengan inti dari sebuah kalimat. Biasanya kata yang termasuk dalam kategori ini adalah kata sambung. Dengan menghapus kata-kata yang tidak bermakna/yang memiliki informasi rendah ini, kita dapat fokus dengan kata lain dari sebuah pertanyaan sehingga maksud dari pertanyaan dapat lebih mudah dipahami. *Tokenizing* akan memanfaatkan library Sastrawi.

3.2.1.3 Stemming

Stemming merupakan proses preprocessing dimana token-token yang memiliki kata berimbuhan diubah kembali menjadi kata dasar/pokok, sehingga mengurangi variasi dari dataset yang sebenarnya berasal dari kata dasar yang sama. Contoh dari proses *stemming* yaitu, misalkan kata "membayar", "pembayaran", "dibayar", "dibayarkan" sebenarnya berasal dari kata dasar yang sama yaitu "bayar". *Tokenizing* akan memanfaatkan library Sastrawi.

3.2.2 kNN (k-Nearest Neighbor)

Pada proses kNN data kumpulan pertanyaan dan jawaban yang dimasukkan akan di pilah sebagai data *training*, kemudian ditentukan jumlah klasifikasi (nilai K, misal K=3). Proses kNN akan mengunakan *library* NMSLib dan Tensorflow Hub. Saat sebuah pertanyaan diajukan ke *chatbot*, sistem akan melakukan klasifikasi terhadap pertanyaan tersebut, mendekati/mirip dengan jenis pertanyaan seperti apa yang ada di dataset.

3.2.3 HMM (Hidden Markov Model)

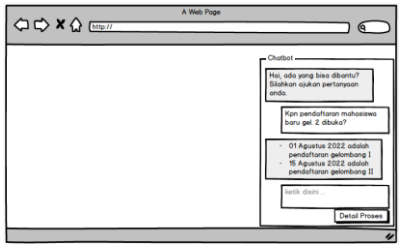
Proses *Hidden Markov Model*/HMM dilakukan dalam mengolah jawaban dari sebuah pertanyaan yang diajukan ke sistem *chatbot*. Proses HMM menggunakan library HMM dan NumPy. HMM dimanfaatkan untuk merangkai jawaban dari dua sumber berdasarkan dari dataset yang ada, untuk memberikan kalimat yang lebih bervariasi. Dataset yang berisi kumpulan jawaban akan menjadi data *train*, setelah sebelumnya dilakukan *preprocessing*.

3.3 Desain *User Interface* (UI)

Pada tahap ini dilakukan pembuatan tampilan/*desain interface* dari sistem yang akan dibuat. Sistem dibuat dalam bentuk *button*/tombol *pop-up* yang dapat diklik dalam sebuah situs web untuk memunculkan sebuah boks yang berisi dialog percakapan dengan sistem *chatbot*. Gambar 4 berikut merupakan tampilan *button* dipojok situs web yang dapat diklik.

Gambar 4. Tampilan *Button Chatbot*

Pada Gambar 5 berikut, merupakan penggambaran sekilas tampilan *chatbot* yang terdiri atas interaksi pertanyaan dari user yang menggunakan *chatbot* dan jawaban yang diberikan oleh sistem.



Gambar 5. Tampilan Sistem *Chatbot*

4. IMPLEMENTASI SISTEM

4.1 Konfigurasi Umum

Penerapan kNN dan HMM dalam pembuatan *chatbot* menggunakan IDE (*Integrated Development Environment*) Spyder 4 yang berbasis pada Python 3.8 versi 64-bit. Spesifikasi sekilas komputer yang digunakan dalam penelitian ini yaitu laptop Dell Inspiron 14R 5437 dengan rincian sebagai berikut:

- Windows 10 Home
- prosesor Intel Core i5-4200U 1.6GHz
- RAM 8GB
- GPU Nvidia Geforce GT 740M

4.2 Persiapan Dataset

Dataset berupa file Excel (.xlsx) yang berisi kumpulan pertanyaan dan jawaban harus diubah menjadi file .csv (*Comma separated values*) dulu. Hal ini bertujuan agar lebih mudah dalam mengekstrak data yang akan diproses sebagai dasar pengetahuan/*knowledge base* dari *chatbot* itu sendiri. Bentuk dataset .csv secara sekilas dapat dilihat pada Gambar 6. Setelah diubah dalam bentuk .csv, dataset masih perlu diubah lagi menjadi bentuk dictionary (dict) sesuai dengan format yang dapat dibaca oleh NMSLib, bentuknya seperti pada Gambar 7.

Pertanyaan	Jawaban	Jawaban2	Intent
apa ada japres?	Kami masih membuk	Jalur prestasi masih di	jalur admisi
uang gedung/uang	Jadwal admission 20	Untuk uang gedung jur	biaya
uang semester apa	Info biaya studi admi	Untuk biaya uang seme	pembayaran
cara pembayaran ua	Info biaya studi admi	Cara pembayaran uang	pembayaran
cara pembayaran ua	Info biaya studi admi	Cara pembayaran uang	pembayaran
biaya pendaftaran	Cara pembelian PAC	Cara pembelian PAC da	biaya
Informasi pendafta	Pembelian PAC dapa	Informasi pendaftaran	jalur admisi
Jalur pendaftaran S	Jadwal dan Alur admi	Jalur pendaftaran dapa	jalur admisi
ini cara login di sim	Langkah Daftar Ulang	Cara login di SIM Petra	admisi
Untuk pembelian p	Pembelian PAC dapa	Untuk pembelian PAC	beli PAC
Sama untuk cara pe	Jadwal dan Alur admi	Cara pendaftaran diba	admisi
Min batas pengump	http://admission.pet	Batas pengumpulan da	admisi
Kalau yang di cetak	http://admission.pet	Batas pencetakan dapa	admisi
siang kak mau tanya	Halo, Selamat Pagi.	Siang ada yang bisa di	halo

Gambar 6. Dataset dalam Bentuk .csv

```
"apa ada japres? ":"1",
"uang gedung/uang pangkal jurusan IBM berapa ya?":"2",
"uang semester apa sama dengan uang sks? ":"3",
"cara pembayaran uang gedung?":"4",
"cara pembayaran uang semester,uang sks.":"5",
"biaya pendaftaran ":"6",
"Informasi pendaftaran untuk Program Studi Bahasa Mandarin":"7",
"Jalur pendaftaran Sastra Tionghoa":"8",
"ini cara login di sim petra gimna yah?":"9",
"Untuk pembelian pac disini ya kak?":"10",
"Sama untuk cara pendaftarannya gimana ya":"11",
"Min batas pengumpulan ya kapan min,yang berkas upload":"12",
"Kalau yang di cetak batas nya kapan?":"13",
"siang kak mau tanya kakk":"14",
```

Gambar 7. Sekilas Dataset dalam Bentuk *Dictionary*

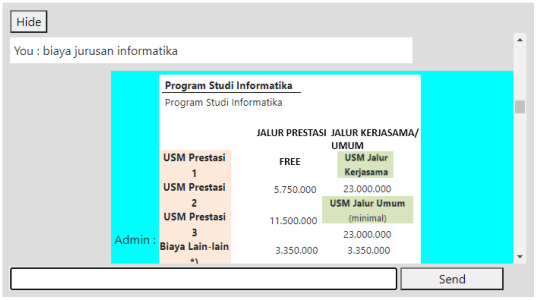
5. ANALISA DAN PENGUJIAN

5.1 Cara Pengujian

Pengujian dilakukan untuk mengetahui metode mana yang lebih baik diantara kNN dan HMM dalam memberikan jawaban dari pertanyaan yang diberikan pada *chatbot*. Waktu yang dibutuhkan dalam memberikan jawaban juga diperhitungkan sebagai pertimbangan.

Pertanyaan yang diberikan terdiri atas tiga bagian/data, masing-masing data terdiri atas 15 buah pertanyaan dengan berbagai topik (baik jurusan umum/pendaftaran mahasiswa baru). Pertanyaan akan diberikan berurutan pertama kali pada *chatbot* yang hanya menjalankan fungsi kNN. Kemudian pertanyaan yang sama akan diberikan pada chatbot dengan fungsi kNN dan HMM. Hasil jawaban dari kedua metode akan dibandingkan beserta dengan waktu prosesnya.

Program *Chatbot* juga diukur performanya dari segi waktu/berapa lama waktu yang dibutuhkan untuk menjawab pertanyaan. Waktu hasil pengukuran akan muncul pada halaman *console* Spyder seperti pada Gambar 8.



Gambar 8. Tampilan *Chatbot*

5.2 Hasil Pengujian

Tabel 1. Persentase Jawaban Benar dari Chatbot (kNN)

Data	Jawaban Benar (R1)	Persentase R1	Jawaban Benar (R2)	Persentase R2
Data 1	12	80%	14	93,33%
Data 2	12	80%	11	73,33%
Data 3	5	33,33%	4	26,67%
Rata-rata (%)		64,44%		64,44%

Dari Tabel 1 sistem *chatbot* memiliki tingkat akurasi dalam menjawab pertanyaan sebesar 64,44% berdasarkan hasil pengujian dari tiga data (Data 1, 2, dan 3). Masing-masing data terdiri atas 15 buah pertanyaan (total 45 pertanyaan), dengan jenis pertanyaan yang bervariasi, baik pertanyaan pendek maupun pertanyaan panjang. *Chatbot* dapat memberikan akurasi baik saat diberikan pertanyaan yang pendek dan simpel, dapat terlihat pada tingkat akurasi pada Data 1 dan Data 2 dimana jenis pertanyaan pendek hingga menengah. Namun saat diujikan dengan Data 3 yang memiliki pertanyaan panjang dan kompleks, chatbot mememberikan jawaban dengan akurasi yang lebih rendah, sehingga tingkat akurasi dapat berubah tergantung dari jenis pertanyaan yang diberikan.

Dari hasil pengujian dengan penambahan HMM, dapat dilihat bahwa jawaban yang dihasilkan/di-*generate* bisa dibilang kacau. Padahal kumpulan jawaban yang diberikan untuk dirangkai oleh HMM sudah dibatasi pada proses sebelumnya, yaitu ketika kNN mengklasifikasikan pertanyaan dan jawaban yang dianggap sesuai. Penyebab utama dari jawaban random yang dihasilkan oleh HMM yaitu karena HMM/*Hidden Markov Model* merupakan metode/model yang berbasis *probabilistic* atau berbasis probabilitas dalam generate/merangkai suatu sequence, dalam hal ini sequence berupa kalimat jawaban. Dengan kondisi begini, bisa saja HMM merangkai jawaban dengan mengambil kata-kata secara acak hanya dengan melihat probabilitas urutan sequence kalimat dari dataset jawaban yang ada.

Tabel 2. Persentase Jawaban Benar dari Chatbot (kNN)

Data	Rata-rata Waktu (detik)	Data	Rata-rata Waktu (detik)
Data 1 (kNN)	0.072093 detik	Data 1 (kNN-HMM)	0.11668 detik
Data 2 (kNN)	0.096953 detik	Data 2 (kNN-HMM)	0.165307 detik
Data 3 (kNN)	0.07036 detik	Data 3 (kNN-HMM)	0.095833 detik
Rata-rata (detik)	**0.079802 detik**	**Rata-rata (detik)**	**0.12594 detik**

Dari Tabel 2 dapat diketahui bahwa *chatbot* dapat memberikan jawaban/timbal balik dari input pertanyaan dalam waktu yang sangat cepat. Waktu sistem *chatbot* yang dihasilkan dalam pengujian ini berdasarkan pada spek komputer Intel Core i5-4200U 1.6GHz, dengan RAM 8GB. Perlu diperhatikan bahwa hasil waktu *chatbot* dapat berbeda sesuai dengan spek komputer yang dipakai.

Sistem *Chatbot* dengan metode kNN memiliki waktu hasil pengujian dengan rata-rata 0.08 detik. Sedangkan sistem *chatbot* dengan kNN-HMM menghasilkan waktu pengujian rata-rata 0.12 detik untuk memberikan jawaban. Baik kedua metode kNN dan kNN-HMM menghabiskan waktu tidak sampai 1 detik untuk memberikan *output* jawaban, walaupun hasil jawaban HMM random dan sedikit lebih lama dibandingkan kNN karena butuh waktu dalam pengolahan untuk merangkai jawaban dari *dataset* yang ada.

6. KESIMPULAN

Berdasarkan hasil pembuatan dan pengujian sistem, dapat disimpulkan bahwa:

- Pembuatan sistem *chatbot* sebagai salah satu layanan informasi UK Petra dapat dibuat dengan metode kNN (*k-Nearest Neighbor*) namun tidak disarankan dengan metode HMM (*Hidden Markov Model*), disebabkan HMM merupakan metode berbasis probabilitas dalam menghasilkan suatu *sequence*, dalam hal ini susunan kalimat. Sistem chatbot yang dibuat memiliki akurasi rata-rata 64% (kNN) dengan waktu sistem sebesar 0,08 detik (kNN) dan 0,12 detik (kNN-HMM).
- Metode kNN (*k-Nearest Neighbor*) memberikan hasil akurasi rata-rata sebesar 64,44% dari 45 buah pertanyaan

dengan berbagai macam kesulitan (pertanyaan pendek hingga panjang), dengan rata-rata waktu sistem menjawab sebesar 0,08 detik pada komputer penguji.

- Metode HMM (*Hidden Markov Model*) kurang cocok digunakan sebagai metode pembuatan *chatbot* karena memberikan susunan jawaban yang random, walaupun dari sisi efisiensi waktu hanya sedikit lebih lama dibandingkan kNN, yaitu sebesar 0,12 detik pada komputer penguji. HMM lebih cocok digunakan sebagai algoritma *speech recognition/speech-to-text.*
- *Chatbot* dapat menjawab berbagai macam pertanyaan mulai dari pertanyaan simpel (pendek) dengan akurasi 80% (berdasarkan hasil uji Data 1 dan Data 2) hingga pertanyaan rumit (panjang) dengan akurasi 33,33% (berdasarkan hasil uji Data 3). Semakin panjang pertanyaan maka akan berdampak pada tingkat akurasi *chatbot*. Hal ini juga bergantung dari banyaknya jumlah *dataset* sebagai *knowledge base* dari *chatbot* itu sendiri.

Saran untuk pengembangan kedepannnya adalah:

- Diperlukan penambahan dataset pertanyaan dan jawaban yang lebih spesifik dan beragam agar dapat menjawab pertanyaan yang lebih rumit, mengingat pertanyaan *chatbot* pasti bervariasi walaupun memiliki inti pertanyaan yang sama.
- Perlu adanya pengembangan sistem *Part of Speech* (POS) *Tagging* yang lebih baik lagi sehingga dapat merangkai jawaban layaknya bahasa natural, khususnya dalam Bahasa Indonesia.
- Dapat menggunakan metode *Machine Learning* sebagai alternatif HMM, misalnya *Long Short-Term Memory* (LSTM)/Transformer.

7. DAFTAR PUSTAKA

[1] A. Mondal, . M. Dey, D. Das, S. Nagpal and . K. Garda, "Chatbot: An automated conversation system for the educational domain," 2018.

[2] A. N. Rohman, E. Utami and S. Raharjo, "Deteksi Emosi Media Sosial Menggunakan Pendekatan Leksikon dan Natural Language Processing," 2019.

[3] A. O. Puspita Dewi, "Kecerdasan Buatan sebagai Konsep Baru pada Perpustakaan," 2020.

[4] A. Rohman, "MODEL ALGORITMA K-NEAREST NEIGHBOR (K-NN) UNTUK PREDIKSI KELULUSAN MAHASISWA," 2015.

[5] A. S. Ahmad, "Brain Inspired Cognitive Artificial Intelligence for Knowledge Extraction and Intelligent Instrumentation System," 2017.

[6] Chowdhury, G. G. 2003. *Natural language processing. Annual Review of Information Science and Technology*, 37, 51–89. https://doi.org/10.1002/aris.1440370103

[7] D. W. Sari, "Implementasi Natural Language Processing pada Chatbot Peribahasa Indonesia," 2018.

[8] E. Susatio, L. Novamizanti and M. A. Pratama, "SISTEM PENGENALAN WAJAH 3D DENGAN MENGGUNAKAN METODE GABOR WAVELET DAN HIDDEN MARKOV MODEL," 2020.

[9] Ginsberg, M. 2012. *Essentials of Artificial Intelligence*. Elsevier Science.

[10] I. N. S. Paliwahet, . I. M. Sukarsa and . I. K. G. Darma Putra, "Pencarian Informasi Wisata Daerah Bali menggunakan Teknologi Chatbot," 2017.

[11] Muljono, U. Afini and C. Supriyanto, "Morphology analysis for Hidden Markov Model based Indonesian part-of-speech tagger," 2017.

[12] Loper, E., & Bird, S. 2002. NLTK: The Natural Language Toolkit. https://doi.org/10.3115/1225403.1225421

[13] M. Elbes, A. Aldajah and O. Sadaqa, "P-Stemmer or NLTK Stemmer for Arabic Text Classification?," 2019.

[14] M. Lobur, A. Romanyuk and M. Romanyshyn, "Using NLTK for educational and scientific purposes," 2011.

[15] M. N. Alfareza, "PEMBANGUNAN CHATBOT MENGGUNAKAN NATURAL LANGUAGE PROCESSING DI JURUSAN TEKNIK INDUSTRI UNIVERSITAS ISLAM INDONESIA," 2020.

[16] M. Rivki and A. M. Bachtiar, "IMPLEMENTASI ALGORITMA K-NEAREST NEIGHBOR DALAM PENGKLASIFIKASIAN FOLLOWER TWITTER YANG MENGGUNAKAN BAHASA INDONESIA," 2017.

[17] Muljono, U. Afini and C. Supriyanto, "Morphology analysis for Hidden Markov Model based Indonesian part-of-speech tagger," 2017.

[18] Ramesh, K., Ravishankaran, S., Joshi, A., & Chandrasekaran, K. 2017. A survey of design techniques for conversational agents. *Communications in Computer and Information Science*, 750, 336–350. https://doi.org/10.1007/978-981-10-6544-6_31

[19] Rismawan, T., Irawan, A. W., Prabowo, W., & Kusumadewi, S. 2008. Sistem Pendukung Keputusan Berbasis Pocket Pc Sebagai Penentu Status Gizi Menggunakan Metode Knn (K-Nearest Neighbor). Teknoin, 13(2), 18–23. https://doi.org/10.20885/teknoin.vol13.iss2.art5

[20] S. C. P, E. N., & Afrianto, I. 2015. Rancang Bangun Aplikasi Chatbot Informasi Objek Wisata Kota Bandung Dengan Pendekatan Natural Language Processing. *Komputa : Jurnal Ilmiah Komputer Dan Informatika*, 4(1), 49–54. https://doi.org/10.34010/komputa.v4i1.2410

[21] Setiyoaji, A., Muflikhah, L., & Fauzi, M. A. 2017. Named Entity Recognition Menggunakan Hidden Markov Model dan Algoritma Viterbi pada Teks Tanaman Obat. *Jurnal Pengembangan Teknologi Informasi Dan Ilmu Komputer*, 1(12), 1858–1864. http://j-ptiik.ub.ac.id/index.php/j-ptiik/article/view/673

[22] Singh, R. H., Maurya, S., Tripathi, T., Narula, T., & Srivastav, G. 2020. Movie Recommendation System using Cosine Similarity and KNN. *International Journal of Engineering and Advanced Technology*, 9(5), 556–559. https://doi.org/10.35940/ijeat.e9666.069520

[23] Siswadi, A. A. P., & Tarigan, A. 2018. Ugleo: a Web Based Intelligence Chatbot for Student Admission Portal Using Megahal Style. *Jurnal Ilmiah Informatika Komputer*, 23(3), 175–191. https://doi.org/10.35760/ik.2018.v23i3.2373

[24] Uysal, A. K., & Gunal, S. 2014. The impact of preprocessing on text classification. *Information Processing and Management*, 50(1), 104–112. https://doi.org/10.1016/j.ipm.2013.08.006