JEPIN
(Jurnal Edukasi dan Penelitian Informatika)
ISSN(e): 2548-9364 / ISSN(p) : 2460-0741

Vol. 7
No. 1
April 2021

Klasifikasi Pertanyaan Bidang Akademik Berdasarkan 5W1H menggunakan *K-Nearest Neighbors*

Kristian Adi Nugraha[#1], Herlina[*2]

[#]*Informatika, Universitas Kristen Duta Wacana*
Jalan Dr. Wahidin Sudirohusodo Nomor 5-25 Yogyakarta 55224
[1]adinugraha@ti.ukdw.ac.id

[*] *Informatika, Universitas Atma Jaya Yogyakarta*
Jalan Babarsari Nomor 44 Yogyakarta 55281
[2]herlina@uajy.ac.id

***Abstrak*— Pertanyaan merupakan metode terbaik dan termudah untuk menggali sebuah informasi. Menurut aturan 5W1H, terdapat enam bentuk dasar pertanyaan yang dapat digunakan untuk memperoleh informasi, yaitu: *what, where, when, why, who, how*. Banyak jurnalis yang menggunakan metode ini, karena dapat diimplementasikan dengan cepat dan mudah untuk membangun sebuah pertanyaan. Untuk membuat sebuah sistem yang dapat memahami sebuah pertanyaan, misalnya seperti pada *chatbot*, terdapat metode khusus yang harus diterapkan untuk dapat membedakan keenam jenis pertanyaan yang ada. Penelitian ini mencoba untuk melakukan klasifikasi terhadap dokumen pertanyaan berdasarkan aturan 5W1H, dengan menggunakan tokenisasi dan stemming pada tahap pra-pemrosesan, kemudian *K-Nearest Neighbors* (*K-NN*) untuk mengklasifikasikan pertanyaan. Berdasarkan hasil pengujian, nilai akurasi tertinggi adalah 70.27% untuk k = 5.**

***Kata kunci*— *K-nearest Neighbors*, Klasifikasi, Pemrosesan Teks, Pertanyaan, 5w1h**

Abstract* — *Questioning are the best and simplest way to gather information. Based on 5W1H rule, there are six basic form of questions to collecting information: what, where, when, why, who, how. Journalist use this approach to find needed data, because it's fast & easy to forming a question using this method. To build a machine that can understand questions, such as chatbot, must have specific processing in order to differentiate six basic form of questions. This study tried to classify a questions based on 5W1H rule, by using tokenization and stemming method for pre-processing phase, then followed by K-Nearest Neighbors (K-NN) to classify questions. Based on test results, the highest K-NN accuracy percentage is 70.27% for k = 5.

Keywords*— *K-nearest Neighbors, Classification, Text Processing, Question, 5w1h

I. PENDAHULUAN

Pertanyaan merupakan salah satu cara dalam menggali sebuah informasi yang ditujukan kepada lawan bicara, baik dalam bentuk lisan maupun tertulis. Pada umumnya, sebuah pertanyaan dapat dikategorikan menggunakan aturan 5W1H (*What*, *Where*, *When*, *Why*, *Who*, *How*), yaitu pertanyaan untuk mencari informasi mengenai apa, di mana, kapan, mengapa, siapa, dan bagaimana dari sebuah topik yang hendak ditanyakan. Aturan tersebut merupakan teknik dasar yang dapat dilakukan seseorang untuk memperoleh informasi yang dibutuhkan secara lengkap.

Salah satu tantangan dalam bidang ilmu pengolahan bahasa natural, khususnya pada sistem penjawab pertanyaan seperti *chatbot*, adalah bagaimana cara mengidentifikasi pertanyaan dengan tepat, sehingga jawaban yang diberikan sesuai dengan harapan pengguna. Untuk menanyakan sebuah jawaban yang sama, pertanyaan yang diajukan dapat memiliki berbagai macam variasi bentuk susunan maupun pemilihan kata. Dua atau lebih pertanyaan tidak selalu memiliki kata tanya tertentu ketika menanyakan sebuah jawaban yang sama [1]. Misalnya pada pertanyaan untuk menanyakan waktu, sebuah kalimat tanya tidak selalu memiliki kata 'kapan' di dalamnya. Seperti pada kalimat "kapan acara akan dimulai?" dan kalimat "pada jam berapa acara akan dimulai?", di mana keduanya memiliki jawaban yang sama meskipun memiliki kata tanya yang berbeda. Selain itu, beberapa faktor lain seperti bahasa tidak baku (misalnya 'ngapain', 'gimana', 'gitu'), penggunaan singkatan (misalnya 'sy', 'bgmn', 'spy'), dan penggunaan bahasa *slang* (misalnya 'mager', 'julid', 'cabut') juga turut menambah variasi pertanyaan yang dapat diajukan untuk memperoleh sebuah jawaban yang sama.

Berdasarkan pemaparan di atas, penulis mencoba untuk menyelesaikan permasalah tersebut dengan melakukan penelitian terkait klasifikasi pertanyaan berdasarkan aturan

5W1H dengan menggunakan metode *K-Nearest Neighbors* (*K-NN*). Metode *K-Nearest Neighbors* dipilih dengan pertimbangan proses komputasi yang ringan untuk data dengan struktur yang tidak kompleks. Penelitian dilakukan pada lingkungan universitas, sehingga data-data pertanyaan yang diperoleh merupakan pertanyaan di bidang akademik terkait perkuliahan, jadwal kegiatan, administrasi kampus, dan sejenisnya. Data-data pertanyaan dikumpulkan dari mahasiswa dan wali studi mahasiswa, khususnya bagi mereka yang sebelumnya pernah mengajukan pertanyaan kepada pihak kampus. Data pertanyaan-pertanyaan tersebut akan dikategorikan secara *manual* ke dalam enam kelas berdasarkan aturan 5W1H untuk keperluan pelatihan maupun pengujian. Dari masing-masing kelas, diambil 10 pertanyaan secara acak untuk dijadikan data latih, kemudian sisanya akan dijadikan data pengujian. Penelitian ini diharapkan dapat memberikan kontribusi dalam bidang pengolahan bahasa natural, khususnya pada topik-topik terkait dengan *chatbot* dalam menjawab pertanyaan. Hasil dari penelitian ini dapat dijadikan acuan dalam melakukan pra-pemrosesan teks (*text pre-processing*) untuk mengetahui tipe pertanyaan yang diajukan berdasarkan aturan 5W1H. Apabila sebuah pertanyaan telah diketahui kategorinya terlebih dahulu, maka jumlah pilihan jawaban yang tersedia akan semakin kecil sehingga tingkat akurasi ketepatan jawaban akan semakin meningkat.

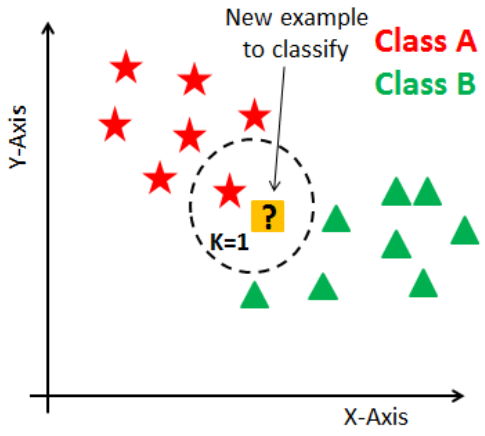
II. TINJAUAN PUSTAKA

Klasifikasi merupakan salah satu topik kecerdasan buatan (*artificial intelligence*) yang berfokus mengenai bagaimana cara mengelompokkan sekumpulan data ke kelompok-kelompok (kelas) tertentu [2, 3]. Proses pengklasifikasian data dibedakan ke dalam dua jenis, yaitu klasifikasi untuk basis data pengetahuan yang telah ditentukan sebelumnya (*supervised learning*) [4, 5] atau klasifikasi untuk basis data yang belum ditentukan (*unsupervised learning*). Klasifikasi diterapkan pada berbagai jenis objek, misalnya seperti klasifikasi data secara umum [6], klasifikasi citra [7, 8] dan klasifikasi teks [9, 10]. *K-Nearest Neighbor* merupakan salah satu metode kecerdasan buatan yang dapat digunakan untuk keperluan klasifikasi [11, 12]. Salah satu contoh implementasi *K-Nearest Neighbors* adalah pada penelitian mengenai klasifikasi dokumen tekstual, yaitu dokumen tertulis yang memiliki jumlah halaman cukup besar, diklasifikasikan dengan menggunakan metode *K-Nearest Neighbors* [13]. Karena jumlah dokumen tekstual yang disimpan semakin bertambah setiap tahun, maka diperlukan adanya pemrosesan otomatis agar dokumen-dokumen tersebut dapat terklasifikasi ke dalam kategori-kategori yang telah ditentukan sebelumnya. Penelitian lain mengenai pemrosesan teks untuk keperluan klasifikasi teks juga dilakukan dengan menggunakan *K-Nearest Neighbors* [14]. Serupa dengan penelitian [13], pada penelitian ini algoritma *K-Nearest Neighbor* digunakan untuk mengklasifikasikan dokumen ke dalam enam topik (kelas) yang telah ditentukan sebelumnya. Secara keseluruhan, algoritma *K-Nearest Neighbors* dapat menghasilkan luaran yang baik dengan proses komputasi yang tidak terlalu berat. Dengan demikian, *metode K-Nearest Neighbors* dapat diimplementasikan pada berbagai *platform*, baik *platform* dengan sumber daya yang besar seperti *server*, maupun *platform* dengan sumber daya terbatas seperti pada perangkat berbasis *Internet of Things* (*IoT*) atau perangkat *smartphone*.

Pada umumnya, data-data yang digunakan pada pemrosesan kecerdasan buatan bersifat mentah (*raw*), artinya di dalam data tersebut terdapat banyak informasi yang tidak dibutuhkan atau tidak berhubungan dengan tujuan dari pemrosesan, hal ini disebut juga dengan istilah derau (*noise*). Untuk meminimalkan informasi-informasi yang tidak terpakai tersebut, maka perlu dilakukan tahap pra-pemrosesan (*pre-processing*), yaitu satu atau lebih proses tambahan yang dilakukan sebelum proses utama dieksekusi [15]. Tujuan dari pra-pemrosesan adalah untuk mengurangi informasi-informasi di dalam data yang bersifat *noise*, sehingga data akan dapat diproses lebih efektif dan efisien. Selain itu, tahap pra-pemrosesan juga dapat dilakukan dengan tujuan untuk mengubah data dengan bentuk tidak beraturan menjadi data bentuk normal atau dikenal dengan istilah normaslisasi [16]. Pada topik-topik terkait pemrosesan teks, tahap pra-pemrosesan yang dilakukan umumnya adalah tokenisasi, *stopwording*, dan *stemming* (lematisasi). Namun pada penelitian ini akan diuji tahap pra-pemrosesan dengan dan tanpa menggunakan *stopwording*. Hal tersebut dikarenakan sebagian besar kata-kata tanya yang terdapat pada sebuah pertanyaan termasuk ke dalam *stopword*. Sehingga apabila tahap *stopwording* ikut dilakukan, maka terdapat kemungkinan bahwa kata-kata kunci yang menjadi penanda jenis pertanyaan tersebut akan terhapus dan berakibat pada pertanyaan yang terklasifikasi dengan tidak tepat.

A. *K-Nearest Neighbor*

K-Nearest Neighbor merupakan salah satu metode atau algoritma kecerdasan buatan (*artificial intelligence*) yang dapat digunakan untuk melakukan klasifikasi atau identifiaksi terhadap sebuah data uji berdasarkan basis data yang digunakan [17, 18]. *K-Nearest Neighbors* bekerja dengan cara mencari sejumlah (k) data dari basis data yang paling mirip dengan data yang diujikan seperti ilustrasi yang ditunjukkan pada gambar 1.



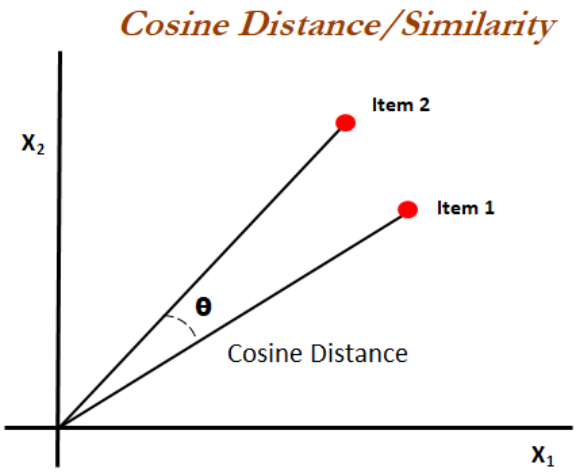
Gambar 1. Ilustrasi klasifikasi menggunakan metode *k-nearest neighbors* untuk nilai k = 1 [19]

Tingkat kemiripan ditentukan berdasarkan jarak dari dua buah data yang dihitung menggunakan persamaan matematika. Pada umumnya, perhitungan jarak dilakukan dengan menggunakan persamaan spasial seperti *euclidean distance*. Namun untuk data berbentuk teks atau dokumen seperti pada penelitian ini, digunakan persamaan *cosine similarity* untuk menghitung tingkat kemiripan dua buah kalimat yang sedang dibandingkan seperti ditunjukkan pada persamaan (1).

$$cos\_similarity(\mathbf{x},\mathbf{y}) = \frac{\sum_{i=1}^{n} x_i \times y_i}{\sqrt{\sum_{i=1}^{n} x_i^2} \times \sqrt{\sum_{i=1}^{n} y_i^2}} \quad (1)$$

Keterangan:
n = jumlah atribut
x = data uji
y = data target

Cosine similarity merupakan persamaan menghitung jarak berbasis vektor, sehingga dapat digunakan untuk data berupa teks seperti data pertanyaan [20]. Ilustrasi dari metode *cosine similarity* ditunjukkan pada gambar 2.

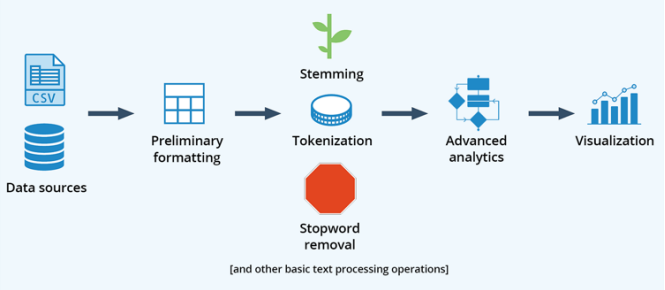


Gambar 2. Ilustrasi *cosine similarity* [21]

B. *Pra-pemrosesan Teks*

Pemrosesan teks atau *text processing* merupakan salah satu cabang dari bidang ilmu komputer yang membahas mengenai bagaimana memproses dokumen dalam bentuk teks secara otomatis untuk keperluan tertentu. Pemrosesan teks membutuhkan implementasi dari metode kecerdasan buatan agar dapat mengenali teks dan memprosesnya agar sesuai dengan harapan. Beberapa contoh implementasi dari pemrosesan teks adalah translasi kalimat dari satu bahasa ke bahasa lain [22], analisis sentimen [23], dan filterisasi spam [24].

Sebelum sebuah dokumen diproses menggunakan metode kecerdasan buatan, terdapat beberapa tahapan yang perlu dilakukan terlebih dahulu agar pemrosesan teks dapat berjalan optimal, tahap ini disebut sebagai pra-pemrosesan teks seperti yang ditunjukkan pada gambar 3.



Gambar 3. Ilustrasi tahap-tahap pemrosesan teks [25]

Tahap pra-pemrosesan teks yang dilakukan pada penelitian ini adalah:

- Tokenisasi
  Proses untuk mengambil kata-kata secara individu dari sebuah kalimat. Hal ini perlu dilakukan agar sistem lebih mudah dalam memproses teks dalam bentuk kata demi kata. Contohnya seperti pada kalimat 'saya pergi ke sekolah', maka hasil luaran tokenisasinya adalah kata 'saya', 'pergi', 'ke', dan 'sekolah'.
- *Stopword*
  Pada umumnya, metode pemrosesan teks bekerja dengan cara mencari kata-kata unik atau khas yang menjadi ciri dari sebuah dokumen, sehingga semakin banyak sebuah dokumen memiliki kata-kata unik, maka semakin mudah dokumen tersebut untuk diproses. *Stopword* adalah daftar kata-kata yang dianggap tidak memiliki makna penting oleh metode pemrosesan teks karena kata-kata tersebut terdapat hampir di setiap dokumen, sehingga kata-kata tersebut biasanya diabaikan atau dihilangkan saat melakukan pemrosesan teks. *Stopword* biasanya berupa kata sambung atau hubung, kata penunjuk waktu atau tempat, kata ganti orang, kata tanya, dan sejenisnya. Beberapa contoh kata yang termasuk dalam stopword adalah 'dan' (kata hubung), 'sekarang' (kata penunjuk waktu), 'di sana' (kata penunjuk tempat), 'mereka' (kata ganti orang), dan 'mengapa' (kata tanya).
- *Stemming*
  Proses untuk mengubah sebuah kata dengan awalan, sisipan, atau akhiran menjadi bentuk kata dasarnya. Hal ini perlu dilakukan agar sistem lebih mudah dalam mengenali kata yang diproses karena berada dalam bentuk kata dasar. Pada penelitian ini komputer tidak perlu memahami konteks kalimat, sehingga adanya awalan, sisipan, atau akhiran tidak akan berpengaruh terhadap hasil luaran tetapi justru memperlambat pemrosesan. Contohnya seperti pada kata 'memasukkan', maka hasil luaran *stemming*-nya adalah 'masuk'.

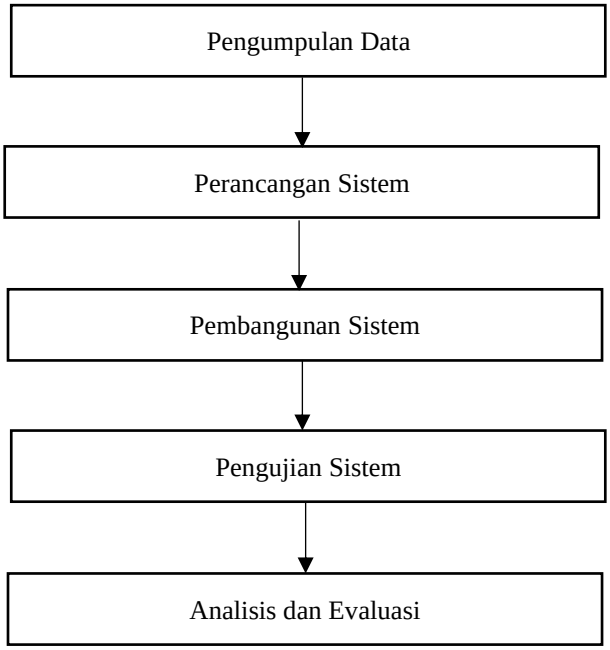
C. Metode *5W1H*

5W1H merupakan enam jenis pertanyaan dasar yang digunakan untuk melakukan pengumpulan informasi (*information gathering*) [26]. Metode ini banyak digunakan oleh para jurnalis untuk mendapatkan informasi yang diinginkan secara detil dan terperinci. Keenam jenis pertanyaan tersebut adalah:

- *What*
  What (apa) merupakan jenis pertanyaan untuk mengetahui tentang apa yang sedang terjadi atau mengenai topik secara umum yang ingin diketahui. Contoh dari pertanyaan bertipe *what* adalah 'apa penyebab kebakaran di gedung tersebut?'.
- *Where*
  Where (di mana) merupakan jenis pertanyaan yang digunakan untuk mencari informasi terkait dengan tempat atau lokasi. Contoh dari pertanyaan bertipe *where* adalah 'Di mana lokasi kebakaran yang sedang terjadi?'.
- *When*
  When (kapan) merupakan jenis pertanyaan yang digunakan untuk mencari informasi terkait dengan waktu mengenai kejadian yang ditanyakan. Contoh dari pertanyaan bertipe *when* adalah 'Kapan api penyebab kebakaran di gedung itu mulai muncul?'.
- *Why*
  Why (mengapa) merupakan jenis pertanyaan yang digunakan untuk mencari informasi dengan menitikberatkan pada alasan atau latar belakang dari kejadian yang ditanyakan. Contoh dari pertanyaan bertipe *why* adalah 'Mengapa kebakaran di gedung tersebut dapat terjadi?'.
- *Who*
  Who (siapa) merupakan jenis pertanyaan yang digunakan untuk mencari informasi terkait dengan subjek, seseorang, atau pelaku dari kejadian yang ditanyakan. Contoh dari pertanyaan bertipe *why* adalah 'Siapa yang bertanggung jawab atas kejadian kebakaran di gedung tersebut?'.
- *How*
  How (bagaimana) merupakan jenis pertanyaan yang digunakan untuk mencari informasi lebih detil mengenai langkah-langkah atau deskripsi dari kejadian yang ditanyakan. Contoh dari pertanyaan bertipe *how* adalah 'Bagaimana kebakaran di gedung tersebut dapat terjadi?'

III. METODE PENELITIAN

Tahap-tahap yang dilakukan pada penelitian ini terdiri dari tahap pengumpulan data, perancangan sistem, pembangunan sistem, pengujian sistem, serta analisis dan evaluasi seperti ditunjukkan pada gambar 4.



Gambar 4. Alur penelitian

A. *Pengumpulan Data*

Pengumpulan data dilakukan dengan meminta mahasiswa dan wali studi mahasiswa sebagai responden untuk menuliskan daftar pertanyaan yang pernah ditanyakan kepada pihak kampus atau universitas terkait dengan aktivitas perkuliahan maupun akademis lainnya. Dari proses pengumpulan data tersebut, berhasil didapatkan data pertanyaan sejumlah 208 pertanyaan yang valid, yaitu pertanyaan yang memiliki struktur lengkap sehingga dapat dipahami manusia serta termasuk ke dalam salah satu kategori yang telah didefinisikan. Dari 208 data pertanyaan valid yang berhasil dikumpulkan, 60 data pertanyaan di antaranya diambil untuk dijadikan basis data latih. Data tersebut terdiri dari 6 kelas, masing-masing kelas memiliki 10 pertanyaan sehingga total keseluruhan berjumlah 60 data. Setiap kelas memiliki 10 data latih dengan pertimbangan nilai k maksimal yang akan diuji adalah 10. Sehingga nilai akurasi yang dihasilkan merupakan nilai berdasarkan batas bawah (*worst case*) karena dibangun dengan jumlah data uji minimal sama dengan nilai k maksimal. Kemudian, sisa data pertanyaan valid sebanyak 148 pertanyaan akan dijadikan data pengujian sistem.

Seluruh data yang dikumpulkan dituliskan menggunakan Bahasa Indonesia. Bentuk bahasa yang digunakan tidak dibatasi, dapat menggunakan bentuk formal/baku maupun bentuk tidak formal. Seluruhnya digunakan baik untuk basis data maupun untuk data uji.

B. *Perancangan Sistem*

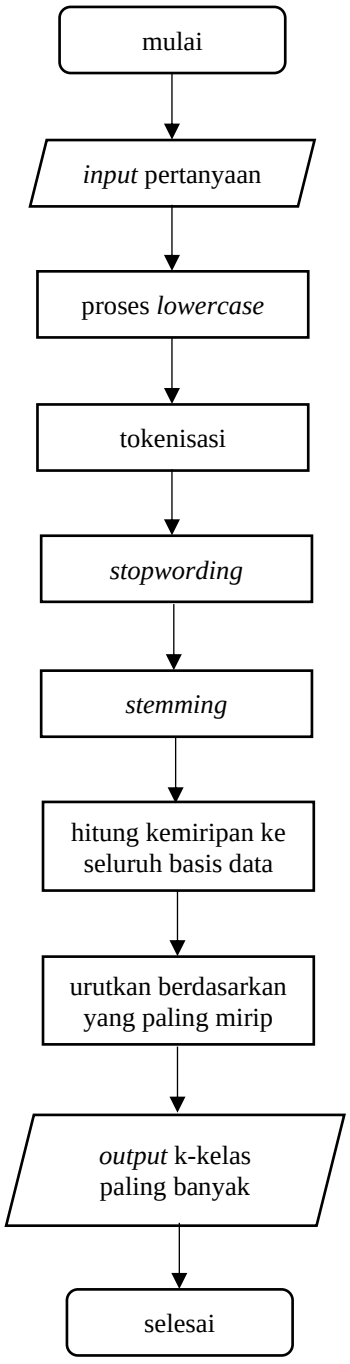
Perancangan sistem dilakukan dengan memperhatikan karakteristik data yang berhasil dikumpulkan. Basis data pertanyaan disusun dalam bentuk struktur data *JavaScript Object Notation* (*JSON*) seperti berikut ini:

```
{
"class": "what",
"patterns": [
        "apakah ada informasi beasiswa baru?",
        "apa saja syarat mendaftar wisuda?",
        "adakah study tour di masa pandemi ini?"
]
},
{
"class": "where",
"patterns": [
        "di mana saya bisa mendaftar wisuda",
        "di mana ruang dekan?",
        "saya bisa lihat info KKN di mana?"
]
},
{
"class": "when",
"patterns": [
        "kapan pendaftaran wisuda dibuka",
        "tanggal berapa deadline pendaftaran skripsi?",
        "kapan saya bisa mengajukan beasiswa?
]
},
{
"class": "why",
"patterns": [
        "mengapa nama saya tidak terdaftar?",
        "mengapa situs kampus tidak bisa diakses?",
        "mengapa perpustakaan belum buka?"
]
},
{
"class": "who",
"patterns": [
        "siapa koordinator skripsi semester ini?",
        "untuk mendaftar cuti studi ke siapa ya?",
        "siapa nama kepala unit perpustakaan?"
]
},
{
"class": "how",
"patterns": [
        "bagaimana cara membatalkan mata kuliah?",
        "bagaimana prosedur pengajuan skripsi?",
        "jika telah terdaftar, bagaimana selanjutnya?"
]
}
```


Setiap entitas pada stuktur data di atas memiliki properti *class* untuk menunjukkan nama kelas dari entitas tersebut, serta properti *patterns* yang berisi daftar pertanyaan-pertanyaan yang dijadikan basis data pengetahuan untuk metode *K-Nearest Neighbors*.

C. Pembangunan Sistem

Pembangunan sistem dilakukan dengan menggunakan bahasa Python versi 3.8. Selain itu terdapat beberapa pustaka (*libraries*) juga turut ditambahkan agar dapat mempermudah proses pembuatan program, beberapa di antaranya adalah pustaka Natural Language Toolkit (NLTK) untuk proses tokenisasi serta pustaka Sastrawi untuk melakukan *stemming*. Diagram alir mengenai cara kerja sistem ditunjukkan pada gambar 5.



Gambar 5. Diagram alir tahap pemrosesan pertanyaan

D. Pengujian Sistem

Pengujian sistem dilakukan dengan menggunakan 148 data pertanyaan di luar data pertanyaan yang dijadikan basis data pengetahuan. Seluruh data pertanyaan dihitung menggunakan persamaan (1) terhadap seluruh data yang terdapat pada basis

data pengetahuan, sehingga akan didapatkan nilai tingkat kemiripan sebanyak 8880 nilai (148 data uji x 6 kelas x 10 data/kelas). Nilai k pada *K-Nearest Neighbors* yang akan diujikan pada sistem mulai dari k paling kecil yaitu 1 sampai dengan nilai k maksimal yaitu 10. Nilai k maksimal tersebut diambil atas dasar pertimbangan jumlah data tiap kelas masing-masing adalah 10, sehingga jika pada algoritma *K-Nearest Neighbors* nilai k diberikan angka 10, maka masih terdapat kemungkinan bahwa 10 data dengan tingkat kemiripian tertinggi berasal dari satu kelas yang sama.

E. Analisis dan Evaluasi

Analisis dan Evaluasi dilakukan setelah seluruh data diujikan pada sistem dan didapatkan hasil luaran kelas untuk masing-masing data berdasarkan nilai k yang telah ditentukan, yaitu k = 1 sampai k = 10. Nilai akurasi rata-rata akan dihitung dengan menjumlahkan seluruh hasil luaran kelas yang didapatkan dengan ketentuan nilai 1 jika kelas sesuai dengan luaran yang diharapkan, dan nilai 0 jika kelas tidak sesuai dengan luaran yang diharapkan.

IV. HASIL DAN PEMBAHASAN

Hasil dari pengujian yang telah dilakukan ditunjukkan pada tabel I dan tabel II.

TABEL I
ACUAN UKURAN TEKS (DENGAN STOPWORDING)

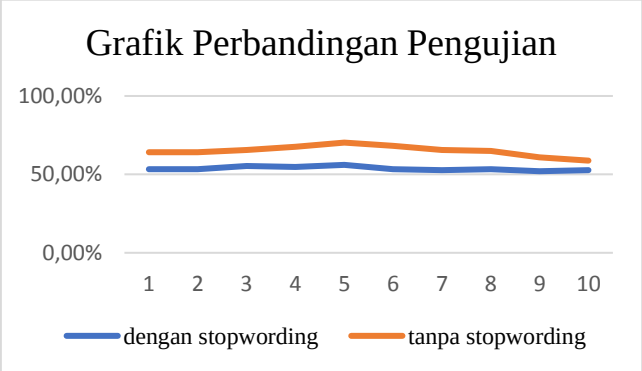
Nilai k	Akurasi (%)
1	53.38%
2	53.38%
3	55.41%
4	54.73%
5	**56.08%**
6	53.38%
7	52.70%
8	53.38%
9	52.03%
10	52.70%

Tabel I berisi data hasil pengujian dengan menggunakan tambahan proses *stopwording*. Sedangkan tabel II berisi data hasil pengujian tanpa adanya proses *stopwording*.

TABEL II
ACUAN UKURAN TEKS (TANPA STOPWORDING)

Nilai k	Akurasi (%)
1	64.19%
2	64.19%
3	65.54%
4	67.57%
5	**70.27%**
6	68.24%
7	65.54%
8	64.86%
9	60.81%
10	58.78%

Berdasarkan hasil pengujian dengan menggunakan dua cara, yaitu dengan *stopwording* dan tanpa *stopwording*, dapat disimpulkan bahwa pengujian dengan *stopwording* menghasilkan nilai akurasi secara keseluruhan yang lebih rendah dibandingkan dengan pengujian tanpa *stopwording*. Nilai akurasi pengujian menggunakan *stopwording* berkisar antara 52.70% sampai dengan 56.08%, sementara nilai akurasi pengujian tanpa menggunakan stopwording berkisar antara 58.78% sampai dengan 70.27% untuk nilai k = 1 sampai k = 10. Hal ini dikarenakan banyak kata-kata kunci yang menjadi penanda jenis pertanyaan, seperti kata 'apa', 'bagaimana', dan 'mengapa', termasuk ke dalam kata-kata *stopword*. Sehingga pada saat proses *stopwording* dilakukan, kata-kata tersebut akan terhapus karena dianggap sebagai kata yang tidak mengandung informasi penting. Grafik perbandingan pengujian dengan menggunakan proses *stopwording* dan tanpa proses *stopwording* ditunjukkan pada gambar 6.



Gambar 6. Grafik perbandingan pengujian dengan *stopwording* dan tanpa *stopword*

Pada tabel II, dapat disimpulkan bahwa persentase nilai akurasi tertinggi sebesar 70.27% didapatkan dengan menggunakan nilai k = 5. Apabila data-data untuk tiap kelas pada nilai k = 5 dijabarkan, maka nilai persentasenya dapat dilihat pada tabel III (diurutkan dari nilai tertinggi).

TABEL III
NILAI AKURASI TIAP KELAS (K = 5)

Kelas	Akurasi (%)
what	75.76%
how	73.33%
who	71.43%
why	55.56%
where	55.56%
when	50.00%

Berdasarkan tabel III, dapat diketahui bahwa tiga kelas pertanyaan tertinggi (*what, how, who*) memiliki nilai persentase akurasi antara 71.43% sampai dengan 75.76%. Sedangkan tiga nilai kelas pertanyaan terendah (*why, where, when*) memiliki nilai persentase akurasi antara 50.00% sampai dengan 55.56%. Hal ini disebabkan data pertanyaan untuk kelas *what, how,* dan *who* memiliki struktur atau format kalimat yang konstan dan variasi yang sedikit, sehingga lebih mudah dikenali oleh sistem. Misalnya pada seluruh pertanyaan dengan kelas *who* selalu terdapat kata 'siapa' di dalamnya, sehingga sistem akan mudah mengenali pertanyaan tersebut sebagai kelas *who*. Tabel IV menunjukkan daftar kata kunci yang muncul hampir di seluruh data pertanyaan.

TABEL IV
DAFTAR KATA KUNCI

Kelas	Akurasi (%)
what	apa, apakah, saja
how	bagaimana, gimana, cara
who	siapa

Sedangkan pada data pertanyaan untuk tiga kelas persentase terendah, yaitu *why*, *where*, dan *when*, ketiganya memiliki pola pertanyaan yang cukup variatif sehingga sulit untuk dikenali oleh sistem. Misalnya pada data pertanyaan untuk menanyakan waktu (*when*), umumnya mengandung kata tanya 'kapan'. Namun pada data pengujian, sebagian besar pertanyaan bertipe *when* tidak selalu selalu mengandung kata tanya 'kapan', melainkan menggunakan kata keterangan waktu seperti 'tanggal', 'jam', dan 'hari' seperti pada contoh berikut ini:

- Selamat siang, saya mau bertanya jadwal ujian kompetensi tanggal berapa ya?
- Acara wisuda berlangsung dari jam berapa sampai berapa?
- Poliklinik kampus buka hari apa saja?

Ketiga pertanyaan di atas memiliki pola yang sangat variatif karena tidak adanya kata tertentu yang selalu muncul di setiap kalimat. Pada kaliamt pertama dan kedua menggunakan kata tanya 'berapa' sehingga sulit dikenali karena kata tanya 'berapa' tidak termasuk ke dalam kategori kelas yang ada, sedangkan pada kalimat ketiga menggunakan kata tanya 'apa' sehingga lebih dikenali sebagai kelas *what*.

V. KESIMPULAN

Berdasarkan seluruh pengujian yang telah dilakukan, sistem dapat memberikan hasil klasifikasi dengan tingkat akurasi sebesar 70.27% untuk nilai k = 5. Pada beberapa kondisi, sistem kesulitan dalam mengenali data pertanyaan karena pola pertanyaan yang sangat variatif dan tidak terdapat pada basis data. Dengan demikian, dibutuhkan adanya tambahan metode khusus agar sebuah kalimat tanya memiliki ciri atau fitur tertentu agar dapat membedakan antara satu kelas dengan kelas yang lain.

REFERENSI

[1] B. Ojokoh, T. Igbe, A. Araoye and F. Ameh, "Question identification and classification on an academic question answering site," in *2016 IEEE/ACM Joint Conference on Digital Libraries (JCDL)*, Newark, NJ, USA, 2016.

[2] Y. Dong, P. Liu, Z. Zhu, Q. Wang and Q. Zhang, "A Fusion Model-Based Label Embedding and Self-Interaction Attention for Text Classification," *IEEE Access* , vol. 8, pp. 30548 - 30559, 2019.

[3] L. Li, Y. Yu, S. Bai, Y. Hou and X. Chen, "An Effective Two-Step Intrusion Detection Approach Based on Binary Classification and k -NN," *IEEE Access* , vol. 6, pp. 12060 - 12073, 2017.

[4] M. Gramajo, L. Ballejos and M. Ale, "Seizing Requirements Engineering Issues through Supervised Learning Techniques," *IEEE Latin America Transactions,* vol. 18, no. 7, pp. 1164 - 1184, 2020.

[5] Y.-F. Li, L.-Z. Guo and Z.-H. Zhou, "Towards Safe Weakly Supervised Learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 43, no. 1, pp. 334 - 346, 2021.

[6] M. A. Zardari and L. T. Jung, "Data classification with k-NN using novel character frequency-direct word frequency (CF-DWF) similarity formula," in *2015 International Symposium on Mathematical Sciences and Computing Research (iSMSC)*, Ipoh, Malaysia, 2015.

[7] K. A. Nugraha, W. Hapsari and N. A. Haryono, "Analisis Tekstur Pada Citra Motif Batik Untuk Klasifikasi Menggunakan K-NN," *Informatika: Jurnal Teknologi Komputer dan Informatika,* vol. 10, no. 2, pp. 135-140, 2014.

[8] K. A. Nugraha, "Deteksi Area Parkir Mobil Berbasis Marker Menggunakan Moment Invariants dan K-NN," *Jurnal Teknik Informatika dan Sistem Informasi* , vol. 5, no. 1, pp. 112-121, 2019.

[9] M. P. Akhter, Z. Jiangbin, I. R. Naqvi, M. Abdelmajeed, A. Mehmood and M. T. Sadiq, "Document-Level Text Classification Using Single-Layer Multisize Filters Convolutional Neural Network," *IEEE Access* , vol. 8, pp. 42689 - 42707, 2020.

[10] S. S. Samant, N. L. B. Murthy and A. Malapati, "Improving Term Weighting Schemes for Short Text Classification in Vector Space Model," *IEEE Access* , vol. 7, pp. 166578 - 166592, 2019.

[11] S. S. Mullick, S. Datta and S. Das, "Adaptive Learning-Based k -Nearest Neighbor Classifiers With Resilience to Class Imbalance," *IEEE Transactions on Neural Networks and Learning Systems,* vol. 29, no. 11, pp. 5713 - 5725, 2018.

[12] H. Ma, J. Gou, X. Wang, J. Ke and S. Zeng, "Sparse Coefficient-Based k -Nearest Neighbor Classification," *IEEE Access* , vol. 5, pp. 16618 - 16634, 2017.

[13] A. Moldagulova and R. B. Sulaiman, "Using KNN algorithm for classification of textual documents," in *2017 8th International Conference on Information Technology (ICIT)*, Amman, 2017.

[14] M. A. Rahman and Y. A. Akter, "Topic Classification from Text Using Decision Tree, K-NN and Multinomial Naïve Bayes," in *2019 1st International Conference on Advances in Science, Engineering and Robotics Technology (ICASERT)*, Dhaka, Bangladesh, 2019.

[15] Z. Jianqiang and G. Xiaolin, "Comparison Research on Text Pre-processing Methods on Twitter Sentiment Analysis," *IEEE Access,* vol. 5, pp. 2870 - 2879, 2017.

[16] D. Sebastian and K. A. Nugraha, "Text normalization for indonesian abbreviated word using crowdsourcing method," in *2019 International Conference on Information and Communications Technology (ICOIACT)*, Yogyakarta, Indonesia, 2019.

[17] X. T. Nguyen, H. Kim and H.-J. Lee, "An Efficient Sampling Algorithm With a K-NN Expanding Operator for Depth Data Acquisition in a LiDAR System," *IEEE Transactions on Circuits and Systems for Video Technology,* vol. 30, no. 12, pp. 4700 - 4714, 2020.

[18] E. T. Maddalena and C. N. Jones, "NSM Converges to a k-NN Regressor Under Loose Lipschitz Estimates," *IEEE Control Systems Letters,* vol. 4, no. 4, pp. 880 - 885, 2020.

[19] A. Navlani, "KNN Classification using Scikit-learn," 2 Agustus 2018. [Online]. Available: https://www.datacamp.com/community/tutorials/k-nearest-neighbor-classification-scikit-learn. [Accessed 20 Februari 2021].

[20] K. A. Nugraha and D. Sebastian, "Pembentukan Dataset Topik Kata Bahasa Indonesia pada Twitter Menggunakan TF-IDF & Cosine Similarity," *Jurnal Teknik Informatika dan Sistem Informasi,* vol. 4, no. 3, pp. 376-386, 2018.

[21] P. Dangeti, "Statistics for Machine Learning by Pratap Dangeti," [Online]. Available: https://www.oreilly.com/library/view/statistics-for-machine/9781788295758/eb9cd609-e44a-40a2-9c3a-f16fc4f5289a.xhtml. [Accessed 20 Februari 2021].

[22] D. Sebastian and K. A. Nugraha, "Sistem Perbaikan Kata Tidak Baku Bahasa Indonesia Menggunakan Metode Crowdsourcing," *Jurnal Teknik Informatika dan Sistem Informasi,* vol. 5, no. 3, pp. 386-396, 2019.

[23] K. A. Nugraha and D. Sebastian, "Analisis Trend Akun Media Sosial Twitter Menggunakan TF-IDF dan Cosine Similarity," in *Rekayasa Teknologi Industri dan Informasi XIII Tahun 2018 (ReTII)*, Yogyakarta, 2018.

[24] T. Xia, "A Constant Time Complexity Spam Detection Algorithm for Boosting Throughput on Rule-Based Filtering Systems," *IEEE Access,* vol. 8, pp. 82653 - 82661, 2020.

[25] D. Harris, "What Is Text Analytics? We Analyze the Jargon," 3 Oktober 2016. [Online]. Available: https://www.softwareadvice.com/resources/what-is-text-analytics/. [Accessed 20 Februari 2021].

[26] Ç. Ç. Karaman, S. Yalıman and S. A. Oto, "Event detection from social media: 5W1H analysis on big data," in *2017 25th Signal Processing and Communications Applications Conference (SIU)*, Antalya, 2017.