BAB II. LANDASAN TEORI

2.1. Dasar Teori

Disini peneliti mengemukakan landasan-landasan teori mendasar yang relevan atau ada kaitannya terhadap penelitian yang akan dilakukan. Diantaranya sebagai berikut:

2.1.1. Tes Uraian

Tes bentuk uraian merupakan alat evaluasi hasil belajar yang paling tua. Tes uraian disebut pula dengan tes esai (essay test) atau tes subjektif. Secara umum tes uraian ini memiliki karakteristik sebagai berikut, pertama, tes uraian adalah tes yang berupa pertanyaan atau perintah yang jawabannya menuntut test mengorganisasikan gagasan atau hal-hal yang telah dipelajarinya dengan cara mengemukakan gagasan tersebut dalam bentuk tulisan. Kedua, jumlah butir soalnya umumnya terbatas, yaitu berkisar empat sampai dengan sepuluh butir. Ketiga, pada mumnya, butir-butir soal tes diawali denga kata-kata: jelaskan, terngkan, uraikan, mengapa, bagaimana, dan kata-kata laian yang menuntut testee memberikan uraian jawaban secara lebih luas. Pada perguruan tinggi, biasanya para dosen menggunakan bentuk uraian tes ini pada saat ujian tengah semester (UTS) atau ujian akhir semester (UAS). Keempat, tes uraian digunakan jika guru ingin mengukur kemampuan menulis (Putri et al., 2022).

2.1.2. Clustering

Clustering dokumen adalah teknik pengelompokan data dokumen yang memiliki konten serupa yang dapat dibandingkan dengan dokumen serupa lainnya. Pengelompokkan ini bertujuan untuk menemukan dan memahami ciri yang membuat suatu dokumen dikelompokkan kedalam kelompok tertentu (N, Al-Obaydy, Hashim, Najm, & Jalal, 2022). Clustering bertujuan untuk mengelompokkan contoh ke dalam kelompok berdasarkan kesamaan mereka. Dalam pengelompokan, contoh-contoh dikelompokkan ke dalam klaster sehingga contoh-contoh yang mirip dikelompokkan dalam klaster yang sama, dan contoh-contoh yang tidak mirip dikelompokkan di luar klaster (Onan, 2019). Clustering juga dapat

berguna untuk mengorganisir koleksi dokumen besar ke dalam kelompok-kelompok yang bermakna, mengidentifikasi topik atau tema dalam kumpulan dokumen, menemukan pola tersembunyi atau hubungan dalam data tak terstruktur, mendeteksi anomali, di mana dokumen yang sangat berbeda dari kelompok utama dapat diidentifikasi untuk diperiksa lebih lanjut.

2.1.3. Teks Preprocessing

Preprocessing sendiri adalah suatu proses mengubah teks asli yang akan digunakan sebagai input dengan menghilangkan unsur tekstual yang kurang bermakna (Najjichah, Syukur, & Subagyo, 2019). Proses ini berfungsi untuk menyiapkan data menjadi data yang bisa dipahami sistem. Adapun library yang akan digunakan untuk proses ini adalah librari Sastrawi (Mashudi & Arief, 2021). Proses ini berfungsi untuk mengatur data agar lebih mudah diproses dan menghilangkan kata-kata yang bukan merupakan kata kunci (H Aris, 2015). Proses ini dibagi menjadi empat tahap sebagai berikut:

1. *Case Folding*

   Case Folding ini berguna untuk menyamaratakan penggunaan huruf kapital, sehingga akan mengurangi inkonsistensi yang akan mempersulit pemrosesan data, contohnya kalimat "Klaster" akan menjadi "klaster".

2. *Tokenization*

   Tokenization merupakan tahapan penguraian string teks menjadi *term* atau kata. Tujuan dari *Tokenization* adalah memisahkan kata-kata dalam sebuah paragraf, kalimat atau halaman ke dalam kata tunggal. Hal ini berguna agar data menjadi lebih halus dan bersifat *general.*

3. *Filtering* dan *Stemming*

   Filtering dilakukan untuk menghilangkan data *string* seperti *hyperlink*, simbol, dan gambar. Sedangkan *stemming* merupakan tahapan pengubahan suatu kata menjadi akar katanya dengan menghilangkan imbuhan awal atau akhir pada kata tersebut.

Stemming akan meningkatkan kecepatan dan akurasi proses clustering karena stemming akan mengurangi jumlah kosakata dan ketergantungan pada bentuk-bentuk kosakata tertentu. Namun stemming bisa selalu efektif karena sebuah kata dasar tidak bisa selalu menjadi perwakilan dan maksud dari kata yang diproses (M. H. Ahmed et al., 2023). Kedua hal ini perlu dilakukkan agar semua data jawaban memiliki format yang sama sehingga mempercepat proses clustering.

4. *Stopword Removal*

Stopword removal, merupakan tahapan penghapusan kata-kata yang tidak relevan dalam penentuan topik sebuah dokumen dan kata yang sering muncul pada dokumen, misalnya “dan”, “atau”, “sebuah”, “adalah”, pada dokumen berbahasa Indonesia. Hal ini bertujuan untuk mengurangi volume data yang akan mempermudah pemrosesan data.

Setelah *preprocessing* selesai data akan di ekstraksi fitur menggunakan metode TF-IDF. Data ini disebut data jawaban terekstrak yang kemudian akan di Clustering menggunakan metode KMeans. Kedua proses ini akan dijelaskan lebih lanjut pada sub bab Desain Sistem.

2.1.4. TF-IDF

TF-IDF merupakan teknik statistikal yang sering digunakan untuk penilaian. Selain itu TF-IDF sering digunakan untuk mengubah data menjadi data vektor. Hal ini disebut vectorizing dengan TF-IDF sebagai vectorizer. Mekanisme penilaian TF-IDF dideskripsikan melalui seberapa pentingnya kata yang muncul pada suatu dokumen. *Term Frequency* (TF) adalah satuan untuk seberapa sering suatu kata muncul dalam sebuah dokumen (N, Al-Obaydy, Hashim, Najm, & Jalal, 2022). Seperti dalam rumus persamaan (1).

$$TF(t,d) = \frac{jumlah\ t\ dalam\ d}{jumlah\ kata\ dalam\ d} \quad (1)$$

Invers Document Frequency (IDF) berguna untuk menghitung frekuensi dari kemunculan suatu kata di semua dokumen seperti dalam rumus persamaan (2).

$$\text{IDF}(t) = \log \frac{jumlah\ keseluruhan\ dokumen}{jumlah\ dokumen\ yang\ memuat\ t} \quad (2)$$

Lalu adapun penilaian TF IDF. Hal ini dilakukan dengan cara mengalikan persamaan TF dengan persamaan IDF seperti yang dicontohkan pada rumus persamaan (3).

$$\text{TF IDF} = TF(t, d) \times \text{IDF}(t) \quad (3)$$

2.1.5. KMeans

Kmeans merupakan metode clustering paling populer karena sangat mudah di gunakan (M. Ahmed et al., 2020). Teknik KMeans menggunakan iteratif *clustering* untuk mengelompokkan sekumpulan objek dalam K klaster berdasarkan atributnya. Dalam kasus ini dokumen dikelompokkan dan kemudian *centroid* dari setiap klaster ditentukan dari rata-rata objek pada klaster, lalu *centroid* dimasukkan ke setiap klaster berdasarkan kemiripan fungsi obyek di dalam klaster yang dihitung menggunakan *euclidean distance* (N, Al-Obaydy, Hashim, Najm, & Jalal, 2022). Kmeans yang digunkakan peneliti merupakan kmeans dari *library scikit-learn*, dimana algoritma *kmeans* pada *scikit-learn* menentukan *centroid* berdasarkan jumlah *cluster* karena *centroid* akan mewakilkan sebuah *cluster* (Ghazal et al., 2021). Berikut langkah menerapkan algoritma KMeans *Clustering*:

1. Pusat klaster atau *centroid* ditentukan berdasarkan nilai, dimana *euclidean distance* digunakan untuk mengatur dokumen d agar dimasukkan ke *centroid* terdekat seperti yang dicontohkan pada rumus persamaan (4).

$$W(C) = \sum_{i=1}^{n} \sum d \epsilon Ci \|d - mi\|^2 \quad (4)$$

2. *Centroid* dari setiap klaster k diperkirakan sama dengan rata-rata data training dokumen d yang dialokasikan pada klaster tersebut seperti pada rumus persamaan (5).

$$m_i^{i+1} = \frac{1}{|Ci|} \sum_{d \epsilon Ci} d \quad (5)$$

3. Jika perbedaan antara *centroid* baru dengan *centroid* lama cukup besar maka ulangi langkah penentuan *centroid* baru sampai

terjadi konvergen atau *centroid* lama hasilnya sama dengan *centroid* baru.

2.1.6. Sastrawi

Sastrawi adalah library stemmer yang berfungsi untuk memperbaiki kesalahan perbaikan kata dari sebuah kata menjadi kata dasar. Sastrawi menggunakan algoritma yang dibuat Nazief dan Adriani, lalu dioptimalisasikan dengan algoritma CS (Rosid et al., 2020).

2.1.7. Bahasa Pemrograman Python

Python adalah bahasa pemrograman yang pada implementasinya berbasis objek. Bahasa pemrograman python adalah *object-oriented programming language* yang bersifat *high level* atau bahasa yang mudah untuk diterjemahkan dan pada penggunaanya mendekati bahasa manusia. Python dapat digunakan untuk merancang berbagai macam sistem seperti *mobile programming*, CLI (*command line interface*), *desktop, web* dan game (Nugraha, Sanjaya, & Pamungkas, 2022)

2.1.8. Metode Elbow

Metode elbow digunakan untuk menentukan jumlah klaster yang paling optimum atau yang terbaik menggunakan SSE atau *sum of squared distance* dari data dan *centroid* klaster (Yuliana Sari et al., 2022). Berikut rumus metode elbow pada persamaan (6) dan langkah metode elbow:

A. Memasukkan nilai klaster
B. Menaikkan nilai klaster
C. Menghitung hasil SSE dari setiap klaster dengan rumus (6)

$$\text{SSE} = \sum_{k=1}^{k} \sum xi \, [xi - ck]^2 \qquad (6)$$

D. Melihat hasil SSE dari nilai klaster yang turun secara drastis
E. Menetapkan nilai klaster yang berbentuk siku

2.1.9. Metode Silhouette

Metode *silhouette coefficient* atau SC. SC sendiri adalah metrik yang dibangun untuk menilai atau mengukur kualitas dari suatu teknik pengelompokkan data, mirip dengan metode *elbow* hanya saja metrik ini dinyatakan dalam bentuk tampilan grafis. Setiap kalster dinyatakan dalam

siluet, yang didasarkan pada perbandingan kerapatan dan pemisahnya. Siluet ini menunjukkan keberadaan objek di dalam klaster (Widaningrum et al., 2022.). Berikut rumus metode *silhouette* pada persamaan (7):

$$S(i) = \frac{b(i)-a(i)}{\max(a(i),b(i))} \quad (7)$$

Dimana, S(i) adalah hasil nilai SC, b(i) adalah nilai minimum dari rata-rata jarak, a(i) adalah perbedaan rata-rata objek (i) ke semua objek lain pada a.

2.1.10. React

ReactJS adalah front-end library yang dikembangkan oleh Facebook. ReactJS digunakan sebagai pendukung dari web-framework. ReactJS memiliki beberapa keunggulan diantaranya memberikan kecepatan, simplicity, dan scalability. ReactJS memungkinkan pengembang dapat membangun sebuah komponen UI yang lebih interaktif, stateful, & reusable (Panjaitan & Pakpahan, 2021).

2.1.11. Flask

Flask adalah microframework yang dipelopori oleh Armin Ronacher. Flask jauh lebih ringan dan cepat karena Flask dibuat dengan ide menyederhanakan inti framework-nya seminimal mungkin. Dengan tagline "web development, one drop at a time", Flask dapat membantu kita membuat situs dengan sangat cepat meskipun dengan librari yang sederhana (Putra & Putera, 2019).

2.1.12. Use Case Diagram

Use case mendefinisikan apa yang akan diproses oleh sistem dan komponen–komponennya. *Use case* bekerja dengan menggunakan skenario yang merupakan deskripsi dari urutan atau langkah–langkah yang menjelaskan apa yang dilakukan oleh *user* terhadap sistem maupun sebaliknya (Setiyani, 2021).

2.1.13. Activity Diagram

Activity diagram, dalam bahasa Indonesia diagram aktivitas, yaitu diagram yang dapat memodelkan proses-proses yang terjadi pada sebuah sistem. Runtutan proses dari suatu sistem digambarkan secara vertikal. Activity diagram merupakan pengembangan dari Use Case yang memiliki

alur aktivitas. Alur atau aktivitas berupa bisa berupa runtutan menu-menu atau proses bisnis yang terdapat di dalam sistem tersebut (Prasetya et al., 2022).

2.2. Tinjauan Pustaka

Tinjauan pustaka disini berisi penelitian terdahulu yang menjadi referensi peneliti. Rincian penelitian terdahulu dapat dilihat pada tabel 2.1.

Tabel 2. 1. Penelitian Terdahulu

No	Judul	Penulis	Metode	Hasil
1	ANALISIS CLUSTERING UNTUK PENGELOMPOKAN JUDUL SKRIPSI MAHASISWA MENGGUNAKAN METODE TF-IDF DAN ALGORITMA KMEANS (STUDI KASUS: STT WASTUKANCANA).	(Nugroho & Hermanto, 2021).	*text preprocessing*, *feature Selection*, TF-IDF, dan *text mining*, dan Kmeans	Hasil dari penelitian ini adalah pengelompokkan judul skripsi mahasiswa yang didapat berdasarkan *cluster* yang terbentuk dan dari *cluster* ini dapat menjadi acuan sebagai rekomendasi dalam penyimpanan skripsi yang sudah dibuat dan penentuan judul skripsi yang akan datang
2	*Document classification using term frequency-inverse document frequency and KMeans clustering*	(Ibrahem Al-Obaydy et al., 2022)	TF-IDF, dan Kmeans	Metode yang digunakan menghasilkan tingkat akurasi lebih tinggi daripada metode *K-Nearest Neighbors* dalam mengolah informasi
3	PENGGUNAAN METODE K-MEANS DENGAN PROBABILITY SAMPLING DAN PARAMETER	(Aisyiyah, 2018)	Kmeans, probability sampling, Silhouette	Metode inisialisasi dengan parameter densitas menghasilkan klaster yang lebih baik

	DENSITAS UNTUK CLUSTERING JAWABAN URAIAN			dibandingan dengan metode probability sampling. Namun dengan menggunakan pengukuran nilai Silhouette, metode probability sampling menghasilkan klaster yang lebih baik dibandingkan metode parameter densitas. Dengan karakteristik dataset yaitu adanya kemiripan antar jawaban, metode parameter densitas lebih tepat digunakan karena mengutamakan kedekatan data dengan data-data lain dalam membentuk klaster.
4	Implementasi Text Mining Pengelompokkan Dokumen Skripsi Menggunakan Metode K-Means Clustering	(Adhe et al., 2020)	Kmeans, TF-IDF, Preprocessing, Silhouette	1. Banyaknya kelompok optimal yang terbentuk dari dokumen skripsi menggunakan metode KMeans Clustering adalah 2 cluster dengan nilai silhouette coefficient 0,12 yang berarti no structure. Hal ini dapat dikarenakan penelitian ini menggunakan dataset dokumen skripsi sebanyak 119, sehingga term-

				term yang didapatkan tidak sepenuhnya mewakili dan menjadi ciri dari dokumen-dokumen yang mengandung term-term tersebut. 2. Hasil pengelompokkan yang terbentuk dari dokumen skripsi menggunakan metode KMeans Clustering adalah sebanyak 2 kelompok dengan anggota cluster ke-1 sebanyak 85 dokumen dan anggota cluster ke2 sebanyak 34 dokumen. Dokumen-dokumen skripsi yang masuk ke cluster 1 didominasi penelitian dengan metode data mining terutama tentang klasifikasi, analisis runtun waktu, analisis regresi, analisis data uji hidup, analisis spasial dan operasi riset. Sedangkan dokumen-dokumen skripsi yang masuk ke cluster 2 didominasi penelitian dengan metode analisis multivariat, pengendalian mutu dan matematika asuransi.

Perbedaan penelitian yang akan dilakukan dengan penelitian dahulu adalah penelitian yang akan dilakukan bertujuan untuk melakukan analisis *clustering* jawaban uraian mahasiswa dengan membandingkan kemiripan kata yang terdapat dalam jawaban mahasiswa tersebut menggunakan metode TF-IDF dan algoritma KMeans *Clustering*, dimana TF-IDF digunakan untuk melakukan membobot jawaban uraian mahasiswa dan KMeans *Clustering* digunakan untuk mengelompokkan jawaban uraian ke dalam klaster berdasarkan kemiripan kata yang terdapat pada jawaban uraian tersebut. Hasil yang diharapkan dari penelitian ini adalah pengelompokkan jawaban uraian mahasiswa yang didapat berdasarkan klaster yang terbentuk dan dari klaster ini dapat menjadi acuan untuk memberikan nilai kepada jawaban uraian mahasiswa tersebut.