BAB III

METODOLOGI PENELITIAN

3.1 Pendahuluan

Bab ini akan menjabarkan setiap tahapan penelitian yang dimulai dari tahap pengumpulan data, metodologi penelitian serta manajemen proyek penelitian. Tahapan penelitian akan dijadikan pedoman dalam setiap fase pengembangan perangkat lunak dalam memberikan solusi untuk rumusan masalah dan tujuan yang ingin dicapai dari penelitian ini.

3.2 Data Penelitian

Data yang digunakan dalam penelitian ini berupa kalimat tanya berbahasa Indonesia. Data yang digunakan terdiri dari 3 kategori kumpulan data yang telah dianalisis dan diberi label. Kategori data yang digunakan dalam penelitian ini diantaranya kalimat tanya *factoid* dan *non-factoid.* Untuk kategori data yang tidak termasuk dua kategori sebelumnya akan diberi label *others.* Sekumpulan data tersebut dilatih menggunakan algoritma *machine learning* untuk mendapatkan hasil model pelatihan, dalam hal ini algoritma yang akan digunakan ialah *Support Vector Machine*. Selanjutnya, data tersebut diuji untuk mengetahui hasil kinerja system untuk tiap metode yang digunakan. Contoh pelabelan pada kalimat tanya berbahasa Indonesia dideskripsikan melalui tabel III-1.

Tabel III-1. Contoh Pelabelan pada Kalimat Tanya

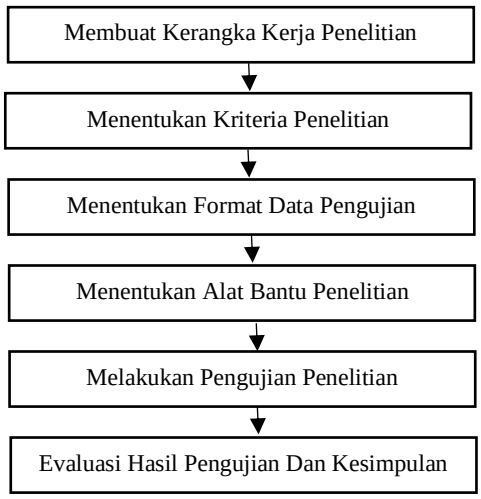
No.	Kalimat Tanya	Label
1.	benua apa yang memiliki sebutan benua biru	*Factoid*
2.	binatang apa yang merupakan khas indonesia dan hanya ditemukan di Indonesia	*Factoid*
3.	bagaimana cara membedakan kata hubung dan kata kerja	*Non-Factoid*
4.	mengapa hari minggu dijadikan sebagai hari libur	*Non-Factoid*
5.	ingat tidak kapan pertama kali kita jalan berdua	*Others*
6.	apa saja bahan untuk membuat sop	*Others*

3.2.1 Jenis dan Sumber Data Penelitian

Data yang digunakan pada penelitian ini merupakan data sekunder yang berisikan kumpulan pertanyaan berbahasa Indonesia berlabel yang didapatkan dari penelitian sebelumnya yang dilakukan oleh Yusliani et al., 2021 dan dimuat dalam format file berupa .xls.

3.3 Tahapan Penelitian

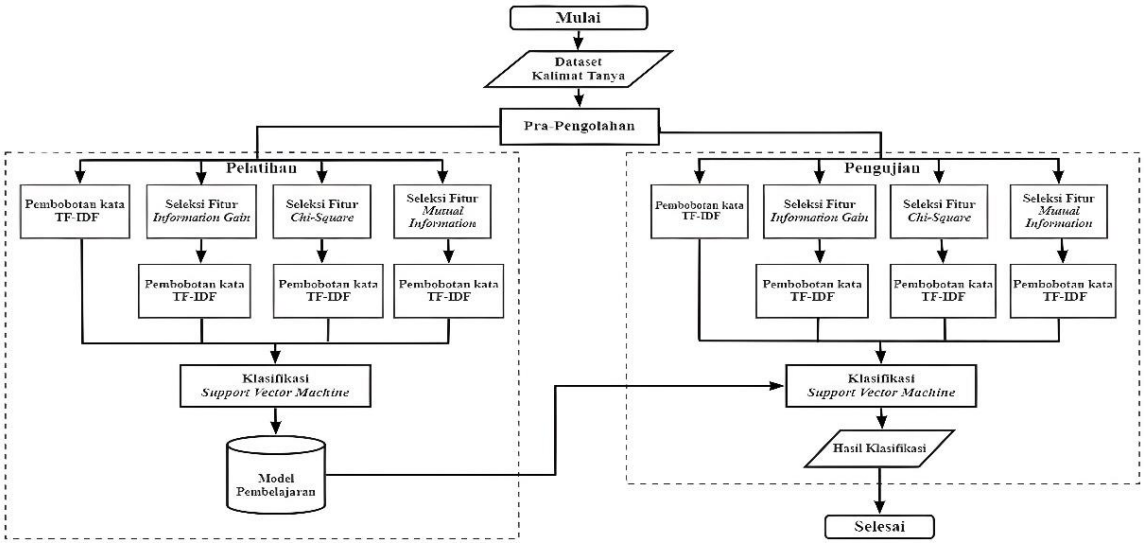
Tahapan penelitian adalah rangkaian kegiatan yang akan dilakukan selama proses penelitian berlangsung. Tahapan tersebut diuraikan pada Gambar III-1.



Gambar III-1. Diagram Tahapan Penelitian

3.3.1 Kerangka Kerja Penelitian

Rangkaian kerangka kerja pada proses pengembangan perangkat lunak dalam penelitian ini diuraikan pada gambar berikut.



Gambar III-2. Diagram Kerangka Kerja

Berdasarkan diagram kerangka kerja yang ditampilkan pada Gambar III-2, terdapat rangkaian proses yang terdiri dari 4 jenis tahapan seleksi fitur yakni *Information Gain, Chi-Square, Mutual Information* serta TF-IDF tanpa seleksi fitur. Selanjutnya, hasil seleksi diklasifikasi menggunakan algoritma *Support Vector Machine*. Rincian tahapan proses diuraikan sebagai berikut.

1. Pra-Pengolahan

   Pra-pengolahan merupakan tahapan awal pada klasifikasi teks, tahapan ini mengubah data yang masih dalam bentuk tidak terstruktur menjadi data yang terstruktur. Tujuan dari tahapan pra-pengolahan ialah mempermudah proses untuk membaca data yang di *input* dan juga berperan dalam penyeragaman data teks. pra-pengolahan diawali dengan proses data *cleaning*. Pada tahap data *cleaning*, data akan dibersihkan dengan cara menghilangkan tanda baca dan karakter yang mengandung *noise*. Contoh karakter yang dihilangkan dalam data teks diantaranya tanda baca, angka, *emoticon* dan simbol.

   Proses kedua dalam pra-pengolahan ialah *case folding*. Dalam proses *case folding* semua huruf yang terdapat pada dokumen akan dibuat sama menjadi bentuk *lowercase* sehingga hanya karakter ‘a’ sampai ‘z’ yang diterima. Proses terakhir ialah *tokenizing* yaitu proses memisahkan dokumen kedalam bentuk satuan kata atau *token*.

   Pada tahap seleksi fitur digunakan 3 metode seleksi fitur yaitu *Information Gain, Chi-Square,* dan *Mutual Information*. Masing-masing metode seleksi fitur ini nantinya akan dilakukan perbandingan pada hasil

kinerja yang didapat. Metode *Information Gain, Chi-Square* dan *Mutual Information* melakukan perhitungan sesuai dengan rumus perhitungan yang ada di masing-masing metode seleksi fitur untuk memberikan nilai atau melakukan *scoring* pada setiap kata. Setelah dilakukan proses *scoring* pada setiap kata. Kemudian, kata-kata yang dipilih untuk digunakan pada tahap pelatihan ialah kata-kata yang lebih dari atau sama dengan nilai *threshold* yang digunakan.

2. Ekstraksi fitur menggunakan TF-IDF

   Pada tahap ekraksi fitur, akan dilakukan pembobotan untuk setiap kata pada data dokumen menggunakan metode gabungan *term-frequency* dan *inverse document frequency* (TF-IDF). Bobot nilai yang diberikan menggunakan metode TF-IDF akan berguna untuk proses klasifikasi pertanyaan berbahasa Indonesia.

3. Klasifikasi menggunakan SVM

   Pada tahap klasifikasi menggunakan algoritma SVM dilakukan dua proses tahapan yaitu pelatihan dan pengujian. Proses yang dilakukan pada pelatihan menggunakan fungsi kernel linear, polinomial dan RBF. Proses ini menggunakan data latih sebagai masukan. Hasil klasifikasi didapat dari pemodelan berupa nilai margin paling besar dari setiap kelas. Penggunaan kernel disesuaikan dengan parameter yang ada disetiap kernel. Setelah mendapatkan model pelatihan, selanjutnya akan dilakukan pengujian algoritma SVM menggunakan data uji sebagai data masukan.

3.3.2 Kriteria Pengujian

Tahap pengujian bertujuan untuk mengetahui hasil kinerja algoritma *Support Vector Machine* dalam melakukan klasifikasi menggunakan seleksi fitur. Kriteria pengujian yang dilakukan pada penelitian ini adalah :

1. Klasifikasi data masukan menggunakan algoritma *Support Vector Machine* tanpa seleksi fitur.
2. Fitur pada data masukan diseleksi menggunakan metode *Information Gain*, selanjutnya data yang telah diseleksi dilakukan proses klasifikasi menggunakan *Support Vector Machine*.
3. Fitur pada data masukan diseleksi menggunakan metode *Chi-Square*, selanjutnya data yang telah diseleksi dilakukan proses klasifikasi menggunakan *Support Vector Machine*.
4. Fitur pada data masukan diseleksi menggunakan metode *Mutual Information*, selanjutnya data yang telah diseleksi dilakukan proses klasifikasi menggunakan *Support Vector Machine*.

Pengujian dilakukan menggunakan metode pengujian *k-fold cross validation*. Nilai *k* yang digunakan pada penelitian ini sebanyak 10 *fold*, untuk melakukan pengujian sebanyak 10 kali. Untuk mengukur hasil kinerja dari klasifikasi dirangkum menggunakan *confusion matrix* untuk menghitung dan menampilkan hasil berupa *accuracy, precision, recall* dan *f-measure.*

3.3.3 Format Data Pengujian

Hasil data pengujian akan dianalisa menggunakan *Confusion Matrix*. Tabel *Confusion Matrix* untuk hasil dari setiap kriteria pengujian di ilustrasikan pada rancangan tabel III-2.

Tabel III-2. Rancangan Tabel *Confusion Matrix* Hasil Klasifikasi

Uji ke-	Kernel :			C :			K :		
	Factoid			Non-Factoid			Others		
	TA	FA1	FA2	TB1	TB	TB2	FC1	FC2	TC
1									
.									
10									
Rata-Rata									

Untuk rancangan hasil kinerja dari klasifikasi setiap pengujian di ilsutrasikan pada tabel III-3.

Tabel III-3. Rancangan Tabel Hasil Pengujian

Uji ke-	Kernel :		C :		K :
	Accuracy	*Precision*	*Recall*	*F-Measure*	Waktu Komputasi
1					
.					
.					
10					
Rata-Rata					

3.3.4 Alat yang Digunakan dalam Pelaksanaan Penelitian

Pada pelaksanaan penelitian perbandingan metode seleksi fitur untuk klasifikasi pertanyaan berbahasa Indonesia menggunakan algoritma SVM, terdapat beberapa alat yang dibutuhkan dalam rangka membantu pengerjaan penelitian. Adapun alat yang dibutuhkan diantaranya :

1. *Hardware*

Processor	: Intel Core i3-5005U
HDD	: 500 GB
RAM	: 8 GB

2. *Software*

Sistem Operasi	: Windows 10 64-bit
Teks Editor	: Spyder
Library	: Pyqt ver 5, NLTK, Pandas, Math, Sklearn

3.3.5 Pengujian Penelitian

Pada penelitian ini terdiri dari 4 jenis pengujian untuk masing-masing metode seleksi fitur, yakni pengujian tanpa seleksi fitur dan pengujian dengan beberapa metode seleksi fitur yakni *Information Gain*, *Mutual Information* dan *Information Gain*. Untuk tahapan pengujian tanpa seleksi fitur diawali dengan tahapan pra-pengolahan data, data yang telah diolah akan dilakukan pembobotan langsung dengan perhitungan TF-IDF. Selanjutnya data yang telah diberi bobot akan diklasifikasi menggunakan algoritma *Support Vector Machine* menggunakan fungsi kernel linear, polinomial dan RBF dengan nilai parameter C yang ditentukan.

Pengujian lainnya ialah dengan menggunakan masing-masing metode seleksi fitur dalam tahapan klasifikasi, tahapan diawali dengan proses pra-pengolahan data. Tahap selanjutnya ialah melakukan penyeleksian fitur menggunakan 3 metode seleksi fitur dengan perhitungan seleksi fitur yang berbeda untuk setiap proses klasifikasi yang dilakukan. Setelah dilakukan *scoring* untuk tiap fitur dan dilakukan pengurutan untuk fitur dengan *score* terbaik, selanjutnya dilakukan pembobotan fitur menggunakan TF-IDF untuk tiap fitur yang telah diseleksi. Selanjutnya, data akan diklasifikasi menggunakan algoritma *Support Vector Machine* menggunakan fungsi kernel linear, polinomial dan RBF dengan parameter berupa nilai C yang ditentukan. Hasil untuk tiap pengujian akan dibandingkan untuk memperoleh hasil kinerja terbaik dari setiap penggunaan seleksi fitur melalui pengujian yang dilakukan.

3.3.6 Analisis Hasil Pengujian dan Membuat Kesimpulan

Agar dapat mengetahui perbandingan hasil kinerja dari setiap metode seleksi fitur diantaranya *Information Gain, Chi-Square, Mutual Information* dan penggunaan algoritma SVM pada klasifikasi pertanyaan berbahasa Indonesia. Maka dilakukan proses evaluasi menggunakan metode evaluasi yang standar dalam mengukur hasil kinerja klasifikasi. Evaluasi tersebut terdiri dari hasil pengukuran *accuracy, recall, precision*, dan *f-measure*. Gambaran dari hasil analisis dapat dilihat pada Tabel III-4.

Tabel III-4. Rancangan Tabel Hasil Analisis Klasifikasi

Algoritma	Nilai C	Nilai K	Kernel :				
			Accuracy	*Precision*	*Recall*	*F-Measure*	Waktu Komputasi
SVM							
SVM+IG							
SVM+CS							
SVM+MI							

3.4 Metode Pengembangan Perangkat Lunak

Pada penelitian ini, metode pengembangan perangkat lunak menggunakan metode RUP (*Rational Unified Process*) yang terdiri dari 4 fase yaitu fase insepsi, fase elaborasi, fase konstruksi, dan fase transisi.

3.4.1 Fase Insepsi

Beberapa proses yang terjadi pada fase insepsi ialah :

1. Pemodelan Binis : Menentukan batasan dan ruang lingkup masalah.
2. Kebutuhan : Menentukan *system requiment* untuk mendapatkan kebutuhan fungsional dan non-fungsional pada pengembangan perangkat lunak.
3. Analisa Perancangan : Melakukan analisis terhadap tahapan kerja, data yang digunakan, kebutuhan fungsional dan non-fungsional yang sebelumnya telah ditentukan pada pengembangan perangkat lunak.

4. Implementasi : Merancang diagram berdasarkan alur kerja dan kebutuhan yang telah ditentukan.

3.4.2 Fase Elaborasi

Beberapa proses yang terjadi pada fase elaborasi ialah :

1. Pemodelan Binis : Menentukan batasan dan ruang lingkup masalah.
2. Kebutuhan : Menentukan *system requiment* untuk mendapatkan kebutuhan fungsional dan non-fungsional pada pengembangan perangkat lunak.
3. Analisa Perancangan : Melakukan analisis terhadap tahapan kerja, data yang digunakan, kebutuhan fungsional dan non-fungsional yang sebelumnya telah ditentukan pada pengembangan perangkat lunak.
4. Implementasi : Merancang diagram *use case* berdasarkan alur kerja dan kebutuhan yang ditentukan.

3.4.3 Fase Konstruksi

Beberapa proses yang terjadi pada fase kontruksi ialah :

1. Kebutuhan : Membuat *class diagram* sesuai model yang telah dirancang.

2. Implementasi : Membangun kode program yang telah dirancang menggunakan bahasa pemrograman yang dipilih.

3.4.4 Fase Transisi

Beberapa proses yang terjadi pada fase transisi ialah :

1. Pemodelan Binis : Menentukan metode pengujian perangkat lunak yang akan digunakan.
2. Kebutuhan : Menentukan kebutuhan alat yang diperlukan untuk tahapan pengujian.
3. Analisa Perancangan : Membuat pengujian melalui skenario *use case*.
4. Implementasi : Melakukan pengujian menggunakan scenario *use case* yang telah dibuat.

3.5 Kesimpulan

Pada bab ini telah dijelaskan secara rinci terkait tahapan penelitian yang akan dilakukan serta metodologi penelitian yang digunakan untuk membangun perangkat lunak yang dapat membandingkan hasil kinerja metode seleksi fitur pada pertanyaan berbahasa Indonesia menggunakan algoritma *Support Vector Machine*.