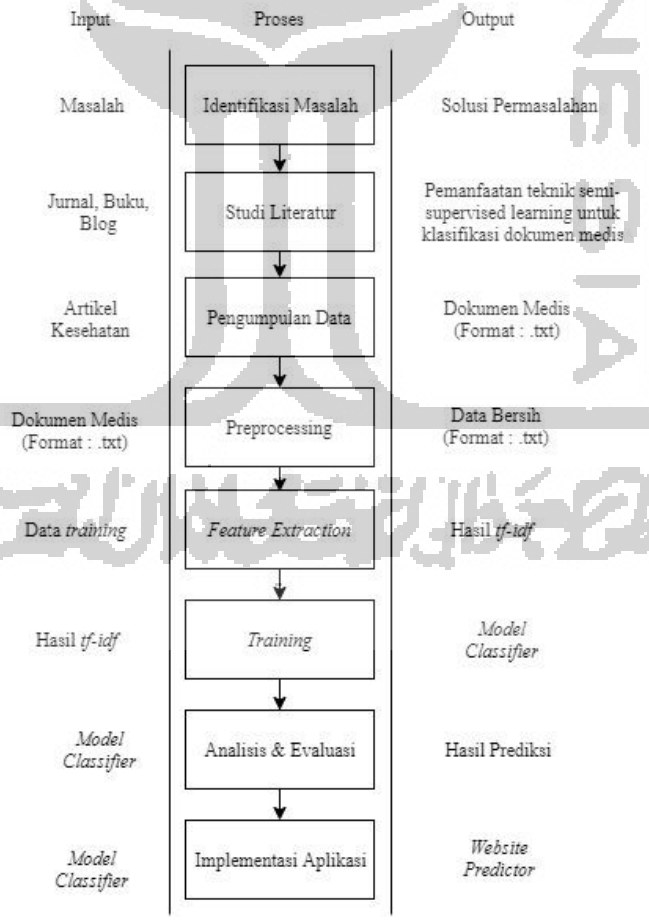
BAB III
METODE PENELITIAN

Pada bab ini dijelaskan mengenai metode penelitian untuk melakukan penelitian klasifikasi dokumen teks. Klasifikasi dilakukan dengan menentukan kategori dari semua dokumen *testing* yang ada. Metode penelitian klasifikasi dokumen teks ini meliputi beberapa langkah pengerjaan yang akan dilakukan (lihat subbab 3.1).

3.1 Langkah-langkah Pengerjaan Tugas Akhir

Langkah-langkah dalam pengerjaan tugas akhir ini dapat dilihat pada Gambar 3.1. Gambar 3.1 menjelaskan secara singkat melalui alur proses penyelesaian tugas akhir yang akan dilakukan. Selain alur proses, Gambar 3.1 juga menampilkan input serta output dari proses pengerjaan tugas akhir yang akan dilakukan.



Gambar 3. 1 Langkah pengerjaan tugas akhir

3.2 Uraian Metodologi

Metodologi yang digunakan dalam mengerjakan tugas akhir ini dapat dilihat pada gambar 3.1 adapun penjelasan dari tiap proses yang dilakukan adalah sebagai berikut:

3.2.1 Identifikasi Masalah

Langkah pertama yang dilakukan dalam menyelesaikan tugas akhir ini adalah mengidentifikasi masalah. Berdasarkan masalah pada penelitian ini yaitu belum adanya metode klasifikasi yang digunakan untuk menyelesaikan persoalan klasifikasi dokumen medis dengan akurasi yang cukup baik. Solusi yang ditawarkan adalah dengan melakukan klasifikasi dokumen medis pada dokumen artikel website kesehatan di Indonesia. Kemudian, mencari tahu lebih detail pemanfaatan teknik *semi-supervised learning* dalam membantu klasifikasi dokumen teks. Pemanfaatan teknik *semi-supervised learning* pada penelitian ini menggunakan *labeled documents* dengan metode *Naïve Bayes* yaitu *Multinomial Naive Bayes* dan percobaan yang memanfaatkan *unlabeled documents* dengan menggunakan *Pseudo-Labeling*.

3.2.2 Studi Literatur

Langah selanjutnya adalah melakukan studi literatur terkait penelitian sebelumnya yang serupa dalam hal klasifikasi dokumen, teknik *semi-supervised learning* untuk mengetahui pola dan hubungan antar kata yang nantinya dapat digunakan untuk mengetahui topik pembahasan dalam dokumen medis. Hasil dari tahap ini adalah menemukan akurasi terbaik untuk menentukan manfaatan teknik *semi-supervised learning* yang menggunakan *labeled documents* dan *unlabeled documents* untuk klasifikasi dokumen medis, dan bagaimana pengaruh dari porsi dokumen medis pada *labeled documents* dan *unlabeled documents* untuk menghasilkan akurasi yang baik dari hasil klasifikasi. Tujuan lain dari tahapan ini adalah untuk mencari tahu konsep dan cara kerja dari teknik *semi-supervised learning* dan mencari tahu tahapan lainnya untuk mendukung dalam menghasilkan hasil analisa yang bagus.

3.2.3 Pengumpulan Data

Langkah selanjutnya adalah melakukan pengumpulan data. Data yang digunakan pada percobaan tugas akhir ini terdiri dari 700 data training dan 100 data testing yang dikumpulkan dan disimpan ke dalam dokumen teks. Data dengan jumlah 800 dokumen tersebut kemudian dibagi menjadi data training dan data testing, dimana data testing jumlahnya sama setiap percobaan yaitu 100 dokumen. Kemudian, data training dengan jumlah 700 dokumen dibagi

menjadi labeled documents dan unlabeled documents dengan porsi awal 400:300. Untuk setiap jenis dokumen di atas, dilakukan uji coba sebanyak 4 kali dengan proporsi dokumen training sebesar 400 dokumen berlabel sampai dengan 100 dokumen berlabel. Data dokumen tersebut berisi artikel kesehatan yang terdiri dari judul dan isi artikel. Data yang digunakan terbagi menjadi 10 kategori yang berkaitan dengan medis yaitu kesehatan bayi, diabetes, diet, jantung, kecantikan, kehamilan, kesehatan gigi dan mulut, kolesterol, kulit, mata. Data training dari 10 kategori medis tersebut juga sudah dibagi menjadi data ber-label (*labeled data*) dengan jumlah 500 data dan yang tidak ber-label (*unlabeled data*) dengan jumlah 300 data.

Proses pengujian dilakukan dengan membagi proses pengujian ke dalam lima tahap. Pengujian pertama menggunakan data acak untuk 400 data berlabel dan 300 data tidak berlabel (model 4:3) sebagai data training dan data acak untuk 100 data berlabel sebagai data testing, pengujian kedua menggunakan data acak untuk 300 data berlabel dan 400 data tidak berlabel (model 3:4) sebagai data training dan data acak untuk 100 data berlabel sebagai data testing, pengujian ketiga menggunakan data acak untuk 200 data berlabel dan 500 data tidak berlabel (model 2:5) sebagai data training dan data acak untuk 100 data berlabel sebagai data testing, dan yang terkahir dilakukan adalah pengujian keempat yaitu menggunakan data acak untuk 100 data berlabel dan 600 data tidak berlabel (model 1:6) sebagai data training dan data acak untuk 100 data berlabel sebagai data testing. Keempat pengujian tersebut akan di *training* dan model yang dihasilkan akan digunakan untuk melakukan klasifikasi dokumen medis.

Keempat pengujian yang dilakukan pada masing-masing porsi dokumen melibatkan 5 kali percobaan. Perbedaan dari masing-masing percobaan adalah untuk data berlabel di acak pada setiap percobaan. Namun, untuk data tidak berlabel menggunakan data yang sama pada setiap percobaan. Dari hasil akurasi 5 kali percobaan yang dilakukan akan dibandingkan nilai akurasi yang dihasilkan baik untuk data berlabel, data tidak berlabel, dan data kombinasi. Nilai akurasi pada tiga tipe data akan diambil nilai yang terbesar sebagai nilai hasil dari pengujian yang dilakukan.

Data *training* dan *testing* berisi artikel medis yang diambil secara manual pada website-website kesehatan di Indonesia, seperti alodokter.com, halodoc.com, sehatq.com, klikdokter.com, hellosehat.com, doktersehat.com. Beberapa dokumen yang diambil secara manual tersebut kemudian disimpan menjadi format dokumen teks berekstensi txt agar dapat dilakukan proses klasifikasi. Pengambilan data pada website di sesuaikan dengan kategori yang telah ditentukan penulis dengan kategori yang ada dalam website. Contoh dokumen yang digunakan pada tugas akhir ini seperti pada Tabel 3.1.

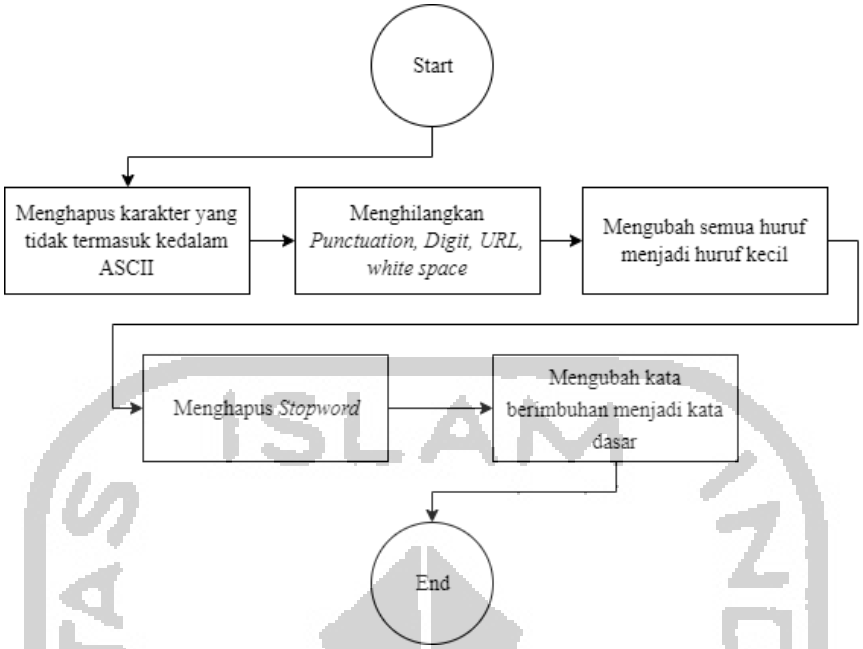
Tabel 3. 1 Jumlah dokumen klasifikasi

Data Training				**Data Testing**	
Labeled Documents		Unlabeled Documents		Labeled Documents	
Kategori	Jumlah Dokumen	Kategori	Jumlah Dokumen	Kategori	Jumlah Dokumen
Bayi	40	Bayi	30	Bayi	10
Diabetes	40	Diabetes	30	Diabetes	10
Diet	40	Diet	30	Diet	10
Jantung	40	Jantung	30	Jantung	10
Kecantikan	40	Kecantikan	30	Kecantikan	10
Kehamilan	40	Kehamilan	30	Kehamilan	10
Gigi&mulut	40	Gigi&mulut	30	Gigi&mulut	10
Kolesterol	40	Kolesterol	30	Kolesterol	10
Kulit	40	Kulit	30	Kulit	10
Mata	40	Mata	30	Mata	10

Unjuk kerja yang diukur dalam percobaan ini adalah tingkat akurasi algoritma dalam melakukan klasifikasi dokumen teks secara benar ke dalam sepuluh kategori yang disebutkan di atas dengan memanfaatkan unlabeled documents dengan porsi yang lebih banyak dibandingkan dengan labeled documents. Oleh karena itu, tugas akhir ini dilakukan untuk melihat seberapa besar manfaat *unlabeled documents* dalam klasifikasi dokumen teks.

3.2.4 Preprocessing

Preprocessing adalah langkah yang harus dilakukan untuk membersihkan data. Data yang sudah dikumpulkan seringkali mengandung karakter-karakter yang tidak perlu yang harus dihilangkan atau biasa disebut dengan *noise*. Tidak hanya dihilangkan beberapa kata yang masih belum baku secara penulisan, *slang word*, dan bahasa asing pun juga perlu dinormalisasi dengan harapan dengan semakin baiknya data yang dimiliki hasil yang akan diperoleh dari proses selanjutnya dalam proses pengolahan bahasa ini akan menjadi lebih baik (Hidayatullah & Ma'arif, 2017). Dalam mengerjakan tugas akhir ini tahapan *preprocessing* yang dilakukan dapat dilihat pada Gambar 3.2.



Gambar 3. 2 Skema preprocessing

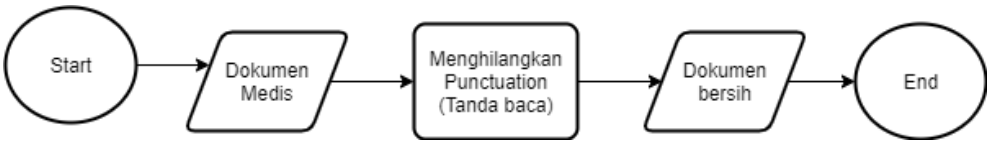
Adapun penjelasan dari tiap-tiap proses adalah sebagai berikut:

a. Menghilangkan karakter yang tidak termasuk ke dalam ASCII

ASCII (*American Standard Code for Information Interchange*) merupakan suatu standar internasional dalam kode huruf dan simbol seperti *Hex* dan *Unicode* tetapi ASCII lebih bersifat universal, contohnya 124 adalah untuk karakter "|". Ia selalu digunakan oleh komputer dan alat komunikasi lain untuk menunjukkan teks. Langkah ini dilakukan untuk menghapus karakter *non*-ASCII.

b. Menghilangkan *Punctuation, Digit, URL, White space*

Menghilangkan *punctuation* dilakukan untuk menghilangkan atau menghapus tanda baca seperti tanda baca titik, koma, tanda seru, tanda tanya, dan lain-lain. Alur proses menghilangkan *punctuation* (tanda baca) dapat dilihat pada Gambar 3.3.



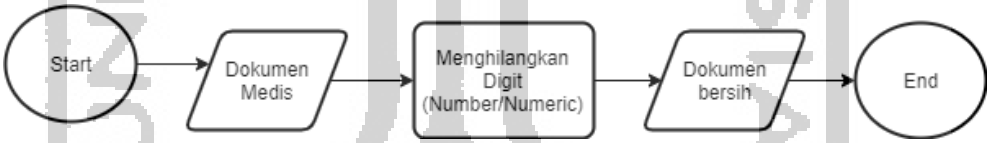
Gambar 3. 3 Alur proses menghilangkan *punctuation*

Contoh penerapan *remove punctuation* pada teks dapat dilihat di Tabel 3.2 yang menjelaskan kata sebelum dan sesudah penerapan *remove punctuation*. Kolom sebelah kiri meunjukan kalimat sebelum dilakukan *remove punctuation*. Kemudian, kolom sebelah kanan menunjukan kalimat yang sudah bersih.

Tabel 3. 2 Menghilangkan *punctuation*

Sebelum	**Sesudah**
Posisi lutut-dada?!@#$%%** Cara alami yang banyak disarankan oleh dokter adalah posisi lutut-dada atau posisi menungging (sujud).	Posisi lutut dada Cara alami yang banyak disarankan oleh dokter adalah posisi lutut dada atau posisi menungging sujud
Mums bisa melakukan gerakan ini kurang lebih 5 kali sehari dalam waktu 3-5 menit.()&^%$%. Olahraga yang paling mudah dilakukan adalah berjalan.	Mums bisa melakukan gerakan ini kurang lebih 5 kali sehari dalam waktu 3 5 menit Olahraga yang paling mudah dilakukan adalah berjalan

Menghilangkan *digit* atau angka, ini bertujuan untuk menghapus angka yang terkandung di dalam data. Alur proses menghilangkan *digit* dapat dilihat pada Gambar 3.4.



Gambar 3. 4 Alur proses menghilangkan *digit*

Contoh penerapan *remove digit* ini pada teks dapat dilihat pada Tabel 3.3 yang menjelaskan kata sebelum dan sesudah penerapan *remove digit*. Kolom sebelah kiri meunjukan kalimat sebelum dilakukan *remove digit*. Kemudian, kolom sebelah kanan menunjukan kalimat yang sudah bersih.

Tabel 3. 3 Menghilangkan *digit*

Sebelum	Sesudah
4 Jenis Alergi yang Sering Dialami Bayi. ilansir dari WebMD, hampir sekitar 6 juta anak memiliki alergi terhadap makanan.	Jenis Alergi yang Sering Dialami Bayi. ilansir dari WebMD, hampir sekitar juta anak memiliki alergi terhadap makanan.
Diperkirakan ada sekitar 1 dari 10 bayi dilaporkan mengalami kondisi dermatitis atopik atau yang lebih sering dikenal dengan eksem.	Diperkirakan ada sekitar dari bayi dilaporkan mengalami kondisi dermatitis atopik atau yang lebih sering dikenal dengan eksem.
Selain 3 faktor yang telah disebutkan di atas, ada juga beberapa faktor lain yang bisa menyebabkan Si Kecil mengalami alergi.	Selain faktor yang telah disebutkan di atas, ada juga beberapa faktor lain yang bisa menyebabkan Si Kecil mengalami alergi. bisa menyebabkan alergi pada kulit Si Kecil.

Menghilangkan *URL* digunakan untuk menghapus atau menghilangkan karakter *url* dari data yang dimiliki. Alur proses menghilangkan *URL* dapat dilihat pada Gambar 3.5.

Gambar 3. 5 Alur proses menghilangkan *URL*

Contoh penerapan *remove url* ini pada teks dapat dilihat pada Tabel 3.4 yang menjelaskan kata sebelum dan sesudah penerapan *remove url*. Kolom sebelah kiri meunjukan kalimat sebelum dilakukan *remove url*. Kemudian, kolom sebelah kanan menunjukan kalimat yang sudah bersih..

Tabel 3. 4 Menghilangkan *URL*

Sebelum	Sesudah
Apa yang Menyebabkan Bayi Sungsang? www.coba.com Hampir setiap ibu #hamil menginginkan proses persalinan secara normal. Namun, ada beberapa faktor yang membuat Mums mau tidak mau harus menjalani operasi Caesar.	Apa yang Menyebabkan Bayi Sungsang? Hampir setiap ibu #hamil menginginkan proses persalinan secara normal. Namun, ada beberapa faktor yang membuat Mums mau tidak mau harus menjalani operasi Caesar.

Menghilangkan *whitespace* digunakan untuk menghapus spasi yang lebih dari satu yang menyebabkan jarak antar kata yang terlalu jauh. Alur proses menghilangkan *whitespace* dapat dilihat pada Gambar 3.6.

Gambar 3. 6 Alur proses menghilangkan *whitespace*

c. Mengubah semua huruf menjadi huruf kecil

Mengubah semua huruf menjadi huruf kecil atau yang sering dikenal dengan *casefolding* dibutuhkan dalam mengkonversi keseluruhan teks dalam dokumen menjadi suatu bentuk standar (biasanya huruf kecil atau *lowercase*). Tujuan dari proses *casefolding* adalah untuk membuat semua teks dalam dokumen konsisten. Contoh penerapan menghilangkan *casefolding* dapat dilihat pada Tabel 3.5.

Tabel 3. 5 Penerapan *casefolding*

Sebelum	Sesudah
Kenali 3 Jenis Temperamen Bayi Berikut!. Setiap manusia dilahirkan dengan karakter dan sifat atau temperamen yang berbeda-beda.	kenali 3 jenis temperamen bayi berikut!. setiap manusia dilahirkan dengan karakter dan sifat atau temperamen yang berbeda-beda.
Ketiga temperamen ini tentunya memiliki ciri yang berbeda, mulai dari yang mudah hingga yang sulit. Apa saja sih ciri dan perbedaannya. EASY. Slow-to-warm. DIFFICULT.	ketiga temperamen ini tentunya memiliki ciri yang berbeda, mulai dari yang mudah hingga yang sulit. apa saja sih ciri dan perbedaannya. easy. slow-to-warm. difficult.

d. Menghapus stopword

Stopword adalah kata-kata yang tidak deskriptif yang dapat dibuang dalam pendekatan bag-of-words. Contoh stopwords adalah "yang", "dan", "di", "dari" dan seterusnya. Tahap ini adalah tahap mengambil kata-kata penting dari hasil token. Berikut beberapa contoh *stop words* yang digunakan pada penelitian ini dapat dilihat pada Tabel 3.6.

Tabel 3. 6 *Stop words*

Kata
yang
saya
oleh
dalam
nya
serasa
dan
Dia
akan
Juga

Contoh penerapan menghapus *stopword* pada teks dapat dilihat pada Tabel 3.7 yang menjelaskan kata sebelum dan sesudah penerapan *remove stopword*. Kolom sebelah kiri meunjukan kalimat sebelum dilakukan *remove stopword*. Kemudian, kolom sebelah kanan menunjukan kalimat yang sudah bersih.

Tabel 3. 7 Menghapus *stopword*

Sebelum	**Sesudah**
Bayi yang baru lahir dan diberikan ASI eksklusif.	Bayi baru lahir diberikan ASI eksklusif .
Pencegahan dan Penanggulangan gerakan seperti mengayuh sepeda pada kaki bayi dan mencoba pijat perut.	Pencegahan Penanggulangan gerakan mengayuh sepeda kaki bayi mencoba pijat perut.
Anda mungkin akan langsung bersikap waspada dan khawatir.	Anda langsung bersikap waspada khawatir. bayi tampak nyaman , mudah marah , menangis .

e. Mengubah kata berimbuhan menjadi kata dasar

Mengubah kata berimbuhan menjadi kata dasar atau yang sering disebut dengan *stemming* merupakan teknik yang diperlukan untuk memperkecil jumlah indeks yang berbeda dari suatu dokumen, juga untuk melakukan pengelompokan kata-kata lain yang memiliki kata dasar dan arti yang serupa namun memiliki bentuk atau form yang berbeda karena mendapatkan imbuhan yang berbeda. Contoh penerapan menghapus *stemming* dapat dilihat pada Tabel 3.8.

Tabel 3. 8 Penerapan *stemming*

Sebelum	Sesudah
Kerugian dan Keuntungan Bayi Tabung. Memperbesar Kemungkinan Hamil. menunggu kehamilan secara alami dengan rajin berkonsultasi pada dokter sehingga bisa mengetahui penyebab susah hamil.	rugi dan untung bayi tabung besar mungkin hamil tunggu hamil cara alami dengan rajin konsultasi pada dokter sehingga bisa tahu sebab susah hamil
Selain itu, pertimbangkan pemilihan rumah sakit, dokter, juga gaya hidup yang Anda jalankan juga penting diperhatikan untuk mendukung tingkat keberhasilan program bayi tabung.	selain itu timbang pilih rumah sakit dokter juga gaya hidup yang anda jalan juga penting perhati untuk dukung tingkat hasil program bayi tabung
Namun, tak adalah salahnya pula untuk memperbesar dan mempercepat peluang mendapatkan anak melalui program bayi tabung ini.	namun tak adalah salah pula untuk besar dan cepat peluang dapat anak lalu program bayi tabung ini
Seperti yang telah disebutkan sebelumnya, bahwa metode bayi tabung memang terbukti mampu memperbesar kemungkinan hamil Anda.	seperti yang telah sebut belum bahwa metode bayi tabung memang bukti mampu besar mungkin hamil anda

Setelah melakukan preprocessing semua dokumen disimpan ke dalam dokumen baru sebagai dokumen bersih. Namun, untuk mempermudah proses klasifikasi dokumen bersih dari hasil preprocessing disimpan menjadi satu ke dalam sebuah dokumen baru. Proses ini perlu dilakukan karena dokumen-dokumen akan lebih mudah melalui proses klasifikasi.

3.2.5 *Feature Extraction*

Feature Extraction adalah proses dari penggunaan informasi dari sekumpulan data untuk menciptakan fitur yang membuat algoritma machine learning bekerja. Penelitian ini menggunakan satu jenis fitur yaitu pembobotan *tf-idf*.

a. Nilai sebuah fitur berdasarkan pembobotan *tf-idf*

Dalam proses klasifikasi, setelah data dipanggil, maka akan dilakukan *feature extraction*. *Feature extraction* adalah salah satu metode yang digunakan untuk memilih kata unik dari

korpus yang dimiliki. Tujuan dari *feature extraction* ini diantaranya adalah agar *classifier* lebih efektif dengan mengurangi ukuran kosa kata dan meningkatkan akurasi dengan menghilangkan noise atau kata yang tidak perlu. Pada penelitian ini metode *feature extraction* yang digunakan adalah *tf-idf.* Berikut ini adalah Tabel 3.9 ilustrasi penghitungan *tf-idf.*

Tabel 3. 9 Ilustrasi data yang digunakan

d1	"kesehatan kulit wanita"
d2	"kesehatan kulit wajah"
d3	"diabetes mellitus wanita"

Setelah data tersedia, selanjutnya adalah mencari kata unik dalam korpus dan menghitung *tf (term frequency)* dalam korpus tersebut seperti yang dapat dilihat pada Tabel 3.10.

Tabel 3. 10 *Term frequency*

	mellitus	diabetes	kesehatan	wajah	wanita	kulit
d1	0	0	1	0	1	1
d2	0	0	1	1	0	1
d3	1	1	0	0	1	0

Selanjutnya adalah menghitung nilai *idf* pada dokumen. Beberapa kata muncul dalam dua dokumen, beberapa hanya muncul dalam satu dokumen. Jumlah total dokumen adalah N = 3. Oleh karena itu, nilai idf seperti yang dilihat pada Tabel 3.11.

Tabel 3. 11 *Idf value*

mellitus	$\log_2\left(\frac{3}{1}\right) = 1.584$
diabetes	$\log_2\left(\frac{3}{1}\right) = 1.584$
kesehatan	$\log_2\left(\frac{3}{2}\right) = 0.584$
wajah	$\log_2\left(\frac{3}{1}\right) = 1.584$
wanita	$\log_2\left(\frac{3}{2}\right) = 0.584$

kulit	$\log_2\left(\frac{3}{2}\right) = 0.584$

Selanjutnya adalah menghitung *tf-idf* dengan menggunakan rumus logaritma seperti yang dapat dilihat pada persamaan 3.1.

$$w_{i,j} = tf_{i,j} \times log(\frac{N}{df_i}) \quad (3.1)$$

Pada persamaan 3.1, maka di peroleh hasil dari *tf –idf* dapat dilihat pada Tabel 3.12.

Tabel 3. 12 Hasil *tf-idf*

	mellitus	diabetes	kesehatan	wajah	wanita	Kulit
d1	0	0	0.584	0	0.584	0.584
d2	0	0	0.584	1.584	0	0.584
d3	1.584	1.584	0	0	0.584	0

Berdasarkan tabel diatas dapat kita lihat bahwa nilai tertinggi adalah jika sebuah kata jarang muncul dalam banyak dokumen, artinya kata ini memiliki kekuatan diskriminatif yang tinggi dalam sebuah dokumen. Untuk nilai yang terendah setelahnya menunjukkan kata ini muncul beberapa kali dalam dokumen yang berbeda sehingga menjadikan kata tersebut memiliki relevansi yang tidak jelas. Dan nilai yang paling rendah adalah kata yang selalu muncul di dalam banyak dokumen atau kata yang sama sekali tidak ada dalam dokumen.

3.2.6 *Training*

Training atau pelatihan adalah salah satu langkah penyelesaian untuk mengerjakan tugas akhir. *Training* dilakukan untuk melatih data yang sudah dipersiapkan agar dapat diklasifikasi. *Training* data yang dilakukan adalah dengan membaca data yang sebelumnya sudah di *convert* menjadi satu dokumen baru. Data yang dibaca akan siap untuk ditraining, salah satunya adalah untuk mengetahui jumlah dokumen tiap-tiap kategori pada dokumen. Selain itu, *training* data juga berguna untuk mencari tau kata-kata yang saling berhubungan dalam suatu dokumen. *Training* data juga berguna untuk mempersiapkan data agar dapat diklasifikasi menggunakan model *machine learning*.

Metode *machine learning* yang akan digunakan pada percobaan tugas akhir ini adalah *Multinomial Naïve Bayes* dan untuk menerapkan metode *Multinomial Naïve Bayes* ke dalam

teknik *semi-supervised learning* yang menggunakan *labeled data* dan *unlabeled data* secara bersamaan, *Pseudo Labeling* adalah teknik yang mendukung. Dua metode tersebut merupakan metode yang bisa digunakan pada teknik *semi-supervised learning* untuk klasifikasi dokumen medis. Pada pembahasan ini, akan dijelaskan metode-metode dalam melakukan klasifikasi dokumen medis:

a. *Multinomial Naïve Bayes*

Naïve Bayes merupakan metode *fully supervised learning* yang memerlukan tahap pembelajaran untuk membangun model probabilistik. Model probabilistik tersebut nantinya akan digunakan untuk melakukan perhitungan *prior* dan *conditional probability* dokumen *testing* dalam menentukan kategori dari dokumen *testing* tersebut. Naïve Bayes membangun model probabilistik dari *feature extraction* yang digunakan dari data *labeled*. Berikut ini adalah contoh penerapan algoritma *Multinomial Naïve Bayes* pada Table 3.13. Pada contoh ini, akan ditunjukkan bagaimana proses penentuan kategori untuk dokumen3 berdasarkan teori yang dibahas pada persamaan (2.2).

Tabel 3. 13 Penerapan algoritma *multinomial naïve bayes*

Dokumen	**Kategori**	**Fitur (Kemunculan)**
dokumen1	bayi	kecil (2), asi (3), bunda (2)
dokumen2	kehamilan	lahir (3), ibu (2), janin (4)
dokumen3	?	lahir (2), kecil (1), baru (2)

Langkah selanjutnya adalah pembuatan model probabilistik dengan melakukan perhitungan. Model probabilistik yang terbentuk adalah seperti Tabel 3.14 berikut:

Tabel 3. 14 Model probabilistik

Kategori	$p(c_i)$	$p(w_{kj}\|c_i)$						
		asi	janin	bunda	kecil	lahir	ibu	baru
bayi	$^1/_2$	$^4/_{14}$	$^1/_{14}$	$^3/_{14}$	$^3/_{14}$	$^1/_{14}$	$^1/_{14}$	$^1/_{14}$
kehamilan	$^1/_2$	$^1/_{16}$	$^5/_{16}$	$^1/_{16}$	$^1/_{16}$	$^4/_{16}$	$^3/_{16}$	$^1/_{16}$

Setelah pembuatan model probabilistik selesai dilakukan, langkah terakhir yang dilakukan adalah penentuan kategori untuk dokumen3:

$$c^* = \arg\max_{c_i \in C} \prod_k p(w_{kj} \mid c_i) \times p(c_i)$$

$p(\text{“bayi”} \mid \text{”dokumen3”}) = p(\text{“bayi”}) \times p(\text{“lahir”} \mid \text{”bayi”}) \times p(\text{“kecil”} \mid \text{”bayi”}) \times p(\text{“baru”} \mid \text{”bayi”})$

$= {}^1/_2 \times {}^1/_{14} \times {}^3/_{14} \times {}^1/_{14}$

$= {}^3/_{5488} \approx 0{,}0000594$

$p(\text{“kehamilan”} \mid \text{”dokumen3”}) = p(\text{“kehamilan”}) \times p(\text{“lahir”} \mid \text{” kehamilan”}) \times p(\text{“kecil”} \mid \text{” kehamilan”}) \times p(\text{“baru”} \mid \text{” kehamilan”})$

$= {}^1/_2 \times {}^4/_{16} \times {}^1/_{16} \times {}^1/_{16}$

$= {}^1/_{2048} \approx 0{,}0004882$

karena p(“*kehamilan*”|”dokumen3”) > p(“bayi”|”dokumen3”), maka kategori dari dokumen3 adalah **kehamilan**.

b. Pseudo Labeling

Proses klasifikasi dokumen medis menggunakan teknik *Pseudo-Labeling* juga melibatkan *Multinomial Naïve bayes* sebagai algoritma untuk melakukan klasifikasi pada *labeled data*. Cara kerja teknik *pseudo-labeling* ini adalah dengan cara menggunakan model yang dibangun dari proses klasifikasi menggunakan *labeled data* untuk melakukan prediksi terhadap *unlabeled data*. Hasil prediksi dari *unlabeled data* ini lah yang dmaksudkan sebagai *pseudo-label data* yang mana label hasil prediksi dianggap sebagai label sebenarnya dari data. Dan dengan memanfaatkan *labeled data* dan *unlabeled data* yang ada, maka dibangun model *classifier* baru. Model *classifier* tersebut yang nanti akan digunakan sebagai model *classifier* dari teknik *semi-supervised learning*.

3.2.7 Analisis dan Evaluasi

Pada percobaan untuk tugas akhir ini digunakan 100 data untuk melakukan uji validitas untuk mengetahui efektifitas model yang sudah dibangun sebelumnya. 100 data tersebut merupakan data testing yang akan digunakan untuk menguji validitas dari model classifier yang dibangun. Tabel 3.15 adalah data testing yang dilabeli sesuai dengan kategori masing-masing teks dokumen medis.

Tabel 3. 15 Dokumen *testing*

Kategori	Jumlah Dokumen
Data Testing	
Labeled Documents	
Bayi	10
Diabetes	10
Diet	10
Jantung	10
Kecantikan	10
Kehamilan	10
Gigi dan mulut	10
Kolesterol	10
Kulit	10
Mata	10

Pada penelitian ini, analisis yang dilakukan adalah dengan menggunakan model 4:3, model 3:4, model 2:5, dan model 1:6. Empat model tersebut akan digunakan untuk melakukan klasifikasi pada seluruh data testing yang ada. Pada keempat model tersebut terdapat tiga model classifier yang dibangun. Model tersebut adalah model yang dibangun dengan data berlabel, data tidak berlabel, dan data kombinasi dari data berlabel dan tidak. Analisis akan dilakukan dengan menggunakan skenario seperti dibawah ini:

a. Percobaan 1, menggunakan model 4:3
    1. Data berlabel

       Klasifikasi data testing dilakukan menggnakan model classifier yang dibangun dari data berlabel.
    2. Data tidak berlabel

       Klasifikasi data testing dilakukan menggunakan model classifier yang dibangun dari data tidak berlabel
    3. Data kombinasi

       Klasifikasi data testing dilakukan menggunakan model classifier yang dibangun dari kombinasi data berlabel dan data tidak berlabel.

b. Percobaan 2, menggunakan model 3:4
    1. Data berlabel

       Klasifikasi data testing dilakukan menggnakan model classifier yang dibangun dari data berlabel.
    2. Data tidak berlabel

       Klasifikasi data testing dilakukan menggunakan model classifier yang dibangun dari data tidak berlabel
    3. Data kombinasi

       Klasifikasi data testing dilakukan menggunakan model classifier yang dibangun dari kombinasi data berlabel dan data tidak berlabel.
c. Percobaan 3, menggunakan model 2:5
    1. Data berlabel

       Klasifikasi data testing dilakukan menggnakan model classifier yang dibangun dari data berlabel.
    2. Data tidak berlabel

       Klasifikasi data testing dilakukan menggunakan model classifier yang dibangun dari data tidak berlabel
    3. Data kombinasi

       Klasifikasi data testing dilakukan menggunakan model classifier yang dibangun dari kombinasi data berlabel dan data tidak berlabel.
d. Percobaan 4, menggunakan model 1:6
    1. Data berlabel

       Klasifikasi data testing dilakukan menggnakan model classifier yang dibangun dari data berlabel.
    2. Data tidak berlabel

       Klasifikasi data testing dilakukan menggunakan model classifier yang dibangun dari data tidak berlabel
    3. Data kombinasi

       Klasifikasi data testing dilakukan menggunakan model classifier yang dibangun dari kombinasi data berlabel dan data tidak berlabel.

Langkah terakhir penelitian ini adalah mendapatkan akurasi yang cukup baik untuk melakukan klasifikasi dokumen medis menggunakan teknik *semi-supervised learning*. Teknik

semi-supervised learning yang akan dimanfaatkan pada penelitian ini adalah dimana jumlah *unlabeled documents* lebih besar dari jumlah *labeled documents*. Percobaan ini dimaksudkan apakah teknik *semi-supervised learning* memiliki kinerja yang baik walaupun hanya memiliki sedikit informasi dari dokumen berlabel. Namun, mampu mengklasifikasi kategori dokumen tidak berlabel dengan jumlah yang besar dan mampu memperoleh akurasi yang cukup baik untuk klasifikasi.

3.2.8 Implementasi Aplikasi

Langkah ini digunakan untuk membangun aplikasi berbasis website yang akan digunakan untuk melakukan klasifikasi dokumen medis. Dalam pengembangannya, website menggunakan Flask sebagai framework dalam Bahasa pemrograman python. Website yang dibangun akan digunakan sebagai predictor. Model yang digunakan untuk melakukan prediksi pada teks dokumen medis adalah model kombinasi data yang dibangun berdasarkan dua jenis data yaitu data berlabel dan data tidak berlabel. Model yang dibangun dibagi menjadi 4 porsi dokumen. Porsi pertama adalah model 4:3, porsi kedua adalah model 3:4, porsi ketiga adalah model 2:5, dan yang terakhir adalah porsi keempat berupa model 1:6.