BAB III
METODOLOGI PENELITIAN

1.1 Objek dan Subjek Penelitian

Objek penelitian merupakan sesuatu yang dijadikan fokus dalam melakukan suatu penelitian untuk mendapatkan jawaban maupun solusi dari permasalahan yang terjadi. Fokus dalam penelitian ini adalah menguji tingkat keakuratan model 1D-Convnet dalam studi kasus klasifikasi kelompok obat SSRI dan atipikal.

Subjek penelitian adalah batasan penelitian dimana peneliti dapat menentukan benda, hal, atau seseorang, untuk melekatkannya menjadi variabel penelitian. Subjek dari penelitian ini adalah kelompok obat SSRI dan atipikal.

1.2 Alat dan Bahan Penelitian

Bahan penelitian yang akan digunakan pada penelitian ini berupa kumpulan data obat yang berasal dari situs *Drugbank online* kemudian mengkategorikan data file dengan format ekstensi .txt ke dalam label SSRI dan atipikal. Sedangkan alat yang akan digunakan pada penelitian ini dijabarkan sebagai berikut:

3.2.1 Perangkat Keras

Perangkat keras yang akan digunakan dalam penelitian ini berupa laptop yang memiliki spesifikasi seperti yang dicantumkan pada Tabel 3.1.

Tabel 3.1 Spesifikasi Laptop Acer Nitro 5

Nama Device	Spesifikasi
Processor	AMD Ryzen 5 3550H
VGA	Radeon Vega Mobile Gfx 210 GHz
RAM	8.00 GB
SSD	224 GB

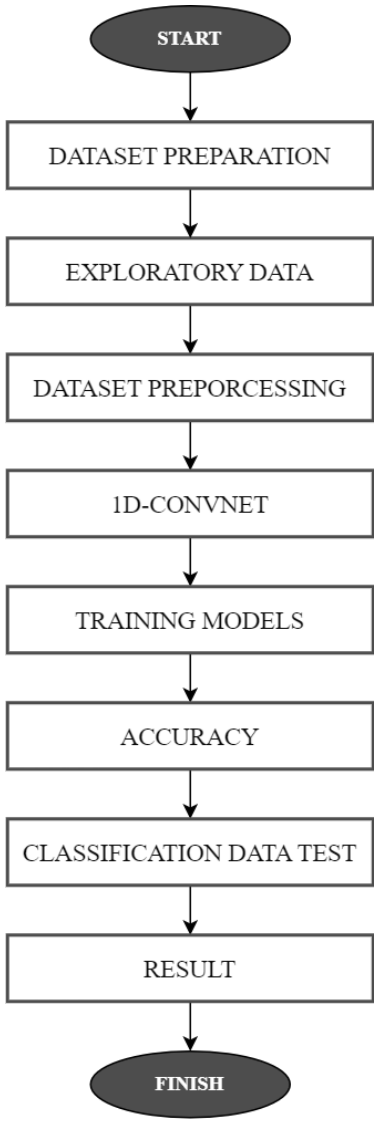
3.2.2 Perangkat Lunak

Perangkat lunak yang akan digunakan untuk memfasilitasi pengembangan penelitian ini dapat dilihat pada Tabel 3.2

Tabel 3.2 Perangkat Lunak Kebutuhan

Nama Perangkat Lunak	Type / Version
Sistem Operasi	Windows 10 64 bit
Browser	Google Chrome
Python	3.7.13
Aplikasi Python	Jupyter Notebook 6.4.12

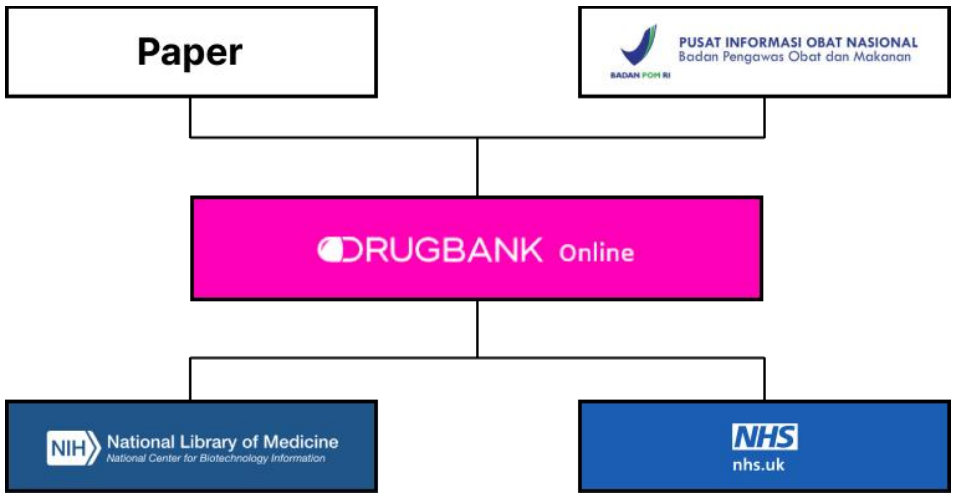
1.3 Diagram Alir Penelitian



Gambar 3.1 Diagram alir penelitian

1.3.1 *Dataset Preparation*

Sebelum membuat program, hal yang paling utama dilakukan adalah mempersiapkan kumpulan data karena program pembelajaran mendalam tidak akan berjalan tanpa adanya kumpulan data. Kumpulan data yang akan digunakan pada penelitian ini berupa informasi obat – obatan untuk penderita gangguan mental khususnya pada kelompok obat SSRI dan atipikal yang tersedia pada database *Drugbank online*. *Drugbank online* merupakan situs yang menyediakan kumpulan informasi kimia terbesar di dunia dan dapat diakses secara bebas dengan mengklik pencarian kemudian pengguna dapat mencari berdasarkan nama obat, rumus molekul, struktur, atau pengenal lainnya. Pada penelitian ini, kumpulan data dicari dengan cara mengetikkan nama obat pada situs *Drugbank online* kemudian akan muncul seluruh keterangan lengkap mengenai obat tersebut.



Gambar 3.2 Sumber informasi obat [6], [40], [41].

Terdapat informasi lengkap mengenai obat – obatan yang ada di dalam situs *Drugbank online*. Dari sekian banyaknya informasi yang ada di dalam situs *Drugbank online*, beberapa informasi yang digunakan untuk kebutuhan penelitian ini yaitu informasi mengenai nama obat di dalam resep, dosis obat dalam satuan miligram (mg), kegunaan obat, dan kelompok obat SSRI atau atipikal. Selain *Drugbank*, terdapat beberapa situs resmi lainnya yang dapat digunakan untuk menambah referensi pencarian data obat seperti *paper* nasional maupun internasional, Pusat Informasi Obat Nasional Badan Pengawas Obat dan Makanan (Pionas BPOM) yang merupakan situs kesehatan resmi milik pemerintahan Indonesia, *National Center for Biotechnology*

Information (NCBI) yang merupakan situs kesehatan resmi milik pemerintahan Amerika Serikat, dan *National Health Service United Kingdom* (NHS UK) yang merupakan situs kesehatan resmi milik pemerintahan Inggris. Meskipun demikian, sumber utama pengumpulan data tetap menggunakan *Drugbank online*.

1.3.2 *Exploratory Data*

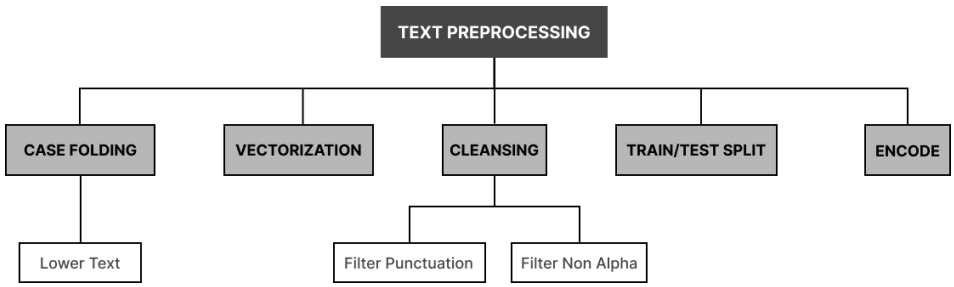
Exploratory Data merupakan suatu proses melakukan analisa awal pada data yang bertujuan untuk mendeteksi kesalahan data di awal dan mengetahui hubungan antar data sehingga analisis data dapat menjadi lebih valid dan relevan. Karena data yang dikumpulkan berupa informasi yang berbentuk susunan kata dan tidak terstruktur maka data dari penelitian ini bersifat kualitatif. Oleh karena itu, teknik yang akan digunakan untuk melakukan *exploratory data* pada penelitian ini berupa teknik statistik deskriptif karena kumpulan data yang digunakan dalam penelitian ini berupa file dengan format ekstensi .txt dan terdapat beragam kata maupun karakter pada setiap file data. Teknik ini dapat digunakan untuk melihat ringkasan dari data secara keseluruhan menggunakan tampilan tabel, diagram, grafik, dan lainnya sehingga hubungan antar data dapat lebih mudah dipahami. Selain itu, program juga dapat mengetahui jika terdapat beberapa kesalahan dalam data.

1.3.3 *Data Preprocessing*

Tahapan *data preprocessing* merupakan suatu proses menyeleksi data agar menjadi lebih terstruktur guna memperoleh hasil yang akurat pada saat pelatihan maupun pengujian. Karena data yang akan digunakan berupa teks maka tahapan *preprocessing* yang dilakukan terdiri dari beberapa macam *text preprocessing.* Sebelum memulai tahapan, kumpulan data dibagi menjadi dua elemen yaitu data target (y) dan data fitur (x). Data target merupakan data yang berisi kelas atau label atau jawaban yang ingin diprediksi sedangkan data fitur adalah data yang berisi atribut untuk mengidentifikasi pola dari model sehingga dapat digunakan untuk memprediksi jawaban target yang tepat. Data fitur (x) bisa disebut dengan istilah data / variabel independen dan data target (y) bisa disebut dengan istilah data / variabel dependen. Setelah data terbagi menjadi dua elemen, selanjutnya data dipisahkan menjadi dua bagian yaitu data latih dan data uji. Data latih merupakan data yang digunakan untuk

melatih model supaya dapat mengenali pola, sedangkan data uji merupakan data yang digunakan untuk menguji hasil dari pelatihan yang telah dilakukan oleh model. Pemisahan data ini dilakukan dengan tujuan untuk menghindari terjadinya *overfitting*. Namun pemilahan kumpulan data ini bisa saja dilakukan pada tahap akhir setelah dilakukan *text preprocessing*.

Karena data yang akan digunakan harus berupa bentuk satu dimensi maka data teks harus diubah terlebih dahulu menjadi bentuk vektor dengan cara melakukan teknik *encode*. Terdapat beberapa macam *text preprocessing* yang dilakukan pada penelitian ini mulai dari *case folding* hingga melakukan tahapan *encode data*. *Text preprocessing* pada penelitian ini dilakukan dengan tujuan untuk memudahkan teknik *encode data* sehingga dapat meminimalisir terjadinya *overfitting* dan mampu menghasilkan akurasi yang optimal. Hasil akhir dari *text preprocessing* adalah kosakata yang telah diencode menjadi numerik.



Gambar 3.3 Arsitektur informasi *text preprocessing*

1. *Case folding*

Case folding merupakan proses mengubah teks menjadi format yang sesuai dengan tujuan mengurangi redudansi data pada saat pelatihan model sehingga perhitungan menjadi optimal [42]. Pada penelitian ini, teks diubah menjadi format *lower text* atau mengubah seluruh teks ke dalam huruf kecil.

2. *Vectorization*

Pada penelitian ini, *vectorization* digunakan sebagai proses untuk memisahkan kata dan karakter dalam kalimat menggunakan delimiter berupa tanda koma. Hal ini bertujuan untuk memudahkan pembersihan data (*cleansing*) dan proses tokenisasi data karena masing – masing kata telah dipisahkan.

3. *Cleansing*

Cleansing merupakan proses pembersihan data yang bertujuan untuk menghilangkan *noise* dan memperbaiki ketidakkonsistenan data dengan cara melakukan penyaringan [43]. Selain itu, *cleansing* dapat memudahkan model dalam menganalisis tingkat keterkaitan hubungan antar kata. Penelitian ini menggunakan dua teknik penyaringan yaitu *filter punctuation* untuk menyaring tanda baca dan *filter non alpha* untuk menyaring kata atau karakter yang memiliki format non – alfabet.

4. *Train / test split*

Membagi dataset menjadi dua bagian yaitu data latih dan data uji merupakan salah satu metode yang dapat digunakan untuk mengevaluasi performa model. Dengan menggunakan metode ini, hasil klasifikasi untuk data baru akan menjadi lebih akurat. Data latih digunakan untuk melatih dan mengembangkan model sedangkan data uji digunakan setelah proses pelatihan model selesai. Kedua jenis data ini akan dibandingkan untuk memeriksa apakah model yang telah digunakan dapat bekerja dengan benar atau tidak.

5. *Encode*

Proses *encode* merupakan proses mengubah teks menjadi bentuk numerik yang bertujuan untuk memudahkan pelatihan model menggunakan 1D-Convnet. Terdapat beberapa cara untuk melakukan *encode data.* Cara pertama adalah melakukan *encode data* menggunakan *categorical encoding*. Cara ini sering digunakan untuk variabel yang hanya terdiri dari satu data namun memiliki beberapa nilai yang sama, biasanya digunakan untuk mengelompokkan nilai dari variabel dependen (y). Cara kedua adalah dengan menggunakan *function tokenization().* Fungsi ini biasa digunakan untuk mengenkode data yang memiliki banyak nilai, biasanya digunakan pada variabel independen (x). Di dalam proses tokenisasi terdapat *function* yang digunakan untuk mengubah kata ke dalam bentuk numerik yaitu *texts_to_sequences()* dan *pad_sequences()* yang digunakan untuk menambahkan nilai nol sebanyak ruang dimensi yang kosong untuk mengisi nilai dari kalimat [44].

1.3.4 1D-Convnet

Setelah kumpulan data selesai diproses, langkah selanjutnya adalah melatih model menggunakan 1D-Convnet yang merupakan turunan model dari arsitektur

CNN. Penelitian ini menggunakan beberapa lapisan untuk membangun model 1D-Convnet, seperti *max pooling*, *flatten*, dan *dense.* Selain itu terdapat juga parameter yang harus dipenuhi pada 1D-Convnet sendiri, diantaranya *filter* / kernel, ukuran kernel, dan *activation.*

Tabel 3.3 Sampel parameter model 1D-Convnet

Lapisan	Parameter
1D – Convnet	Filter = 32, Kernel_size = 8, Activation = ”relu”
MaxPooling1D	Pool_size = 2
Dense	Units = 10, activation = “relu”
Dense	Units = 1, activation = “sigmoid”

Pada saat membangun model 1D-Convnet, penelitian ini menambahkan lapisan *embedding* dengan tujuan untuk menjadikan data teks yang telah di *encode* menjadi lebih terstruktur karena *embedding* mempelajari representasi kata dalam ruang dimensi dimana kata – kata yang memiliki makna hampir sama akan diwakilkan oleh vektor yang nilai dan letaknya berdekatan.

Tabel 3.4 Sampel parameter lapisan *embedding*

Parameter	Jumlah
Input_dim	vocab_size
Output_dim	300
Input_length	Max_length

Pada tabel 3.4, terdapat tiga buah parameter yang harus dipenuhi pada *embedding* diantaranya *input_dim* (ukuran kosakata dalam teks), *output_dim* (ukuran ruang vektor tempat kosakata akan disematkan) yang bisa ditentukan sebelumnya, dan *input length* (panjang keseluruhan inputan kata dalam data). Nilai *vocab_size* didapat dari total ukuran data yang telah ditokenisasi sedangkan *max_length* didapat dari panjang maksimum kalimat yang ada pada data latih. Hasil keluaran dari lapisan *embedding* berbentuk tiga dimensi yang terdiri dari ukuran *batch size* (opsional), jumlah baris yang berisi jumlah sampel kata, jumlah kolom yang berisi jumlah fitur. Hasil keluaran dari lapisan *embedding* dapat digunakan sebagai masukan untuk lapisan selanjutnya yakni lapisan konvolusi.

1.3.5 *Training Models*

Sebelum model dilatih menggunakan *function fit()*, model di *compile* terlebih dahulu dengan tujuan menganalisa *loss function* menggunakan *optimizer* "adam" dan *loss* "*binary_crossentropy*". Model dapat dikatakan sempurna apabila memiliki nilai *cross entropy loss* senilai nol atau mendekati nol. Penelitian ini menggunakan *binary cross entropy* karena perhitungan *loss* ini dikhususkan untuk klasifikasi data dengan nilai biner dan data target memiliki nilai biner sehingga cocok untuk menggunakan jenis *loss* ini. Selain itu, pelatihan model menggunakan *optimizer* berupa *Adam Optimizer* dengan tujuan untuk memperbarui bobot secara iteratif yang didasarkan pada data training.

1.3.6 *Accuracy*

Perhitungan akurasi pada model yang telah dilatih, dilakukan dengan menggunakan perhitungan *confusion matrix* supaya dapat memahami dengan jelas apakah masih ada data yang tidak akurat atau nihil dan bertujuan untuk mengukur performa dalam permasalahan klasifikasi sehingga model dapat digunakan untuk memprediksi data.

1.3.7 *Classification Data Test*

Untuk membuktikan hasil keakuratan model, penelitian ini menguji keakuratan model dengan cara mengklasifikasikan data uji. Hasil keluaran dari klasifikasi ini berupa nilai nol (0) dan satu (1). Nilai nol memiliki arti bahwa inputan tersebut termasuk ke dalam kategori obat atipikal dan nilai satu memiliki arti bahwa inputan tersebut termasuk ke dalam kategori obat SSRI.

1.3.8 *Result*

Hasil yang dijabarkan pada penelitian ini berupa analisis dari akurasi yang diperoleh baik pada saat pelatihan model maupun ketika menggunakan *confusion matrix* dan hasil klasifikasi data uji kemudian dilakukan perbandingan dengan penelitian sebelumnya menggunakan studi kasus yang sama sehingga dapat menyimpulkan perbedaan antara penelitian yang sedang dilakukan dengan penelitian terdahulu dan keunggulan dari penelitian yang sedang dilakukan.

1.4 Hipotesis Penelitian

Penelitian ini dilakukan untuk membuktikan bahwa model 1D-Convnet dapat digunakan untuk mengklasifikasikan kelompok obat SSRI dan atipikal berdasarkan kumpulan data teks. Berikut merupakan hipotesis dari penelitian ini:

H_0 : Pasien gangguan mental yang mengonsumsi obat secara berlebihan dan tidak sesuai anjuran dokter, tidak dapat membedakan jenis obat yang sedang dikonsumsinya sehingga mengalami efek samping yang tidak terduga.

H_1 : Untuk memperoleh target akurasi minimal 95% maka beberapa hal dapat dilakukan seperti jumlah kumpulan data minimal 400 data, percobaan evaluasi model menggunakan *epoch* dan *batch size* yang berbeda, serta percobaan ukuran parameter yang berbeda pada saat pelatihan model.