# BAB III
# METODOLOGI PENELITIAN

## 3.1 Jenis Penelitian

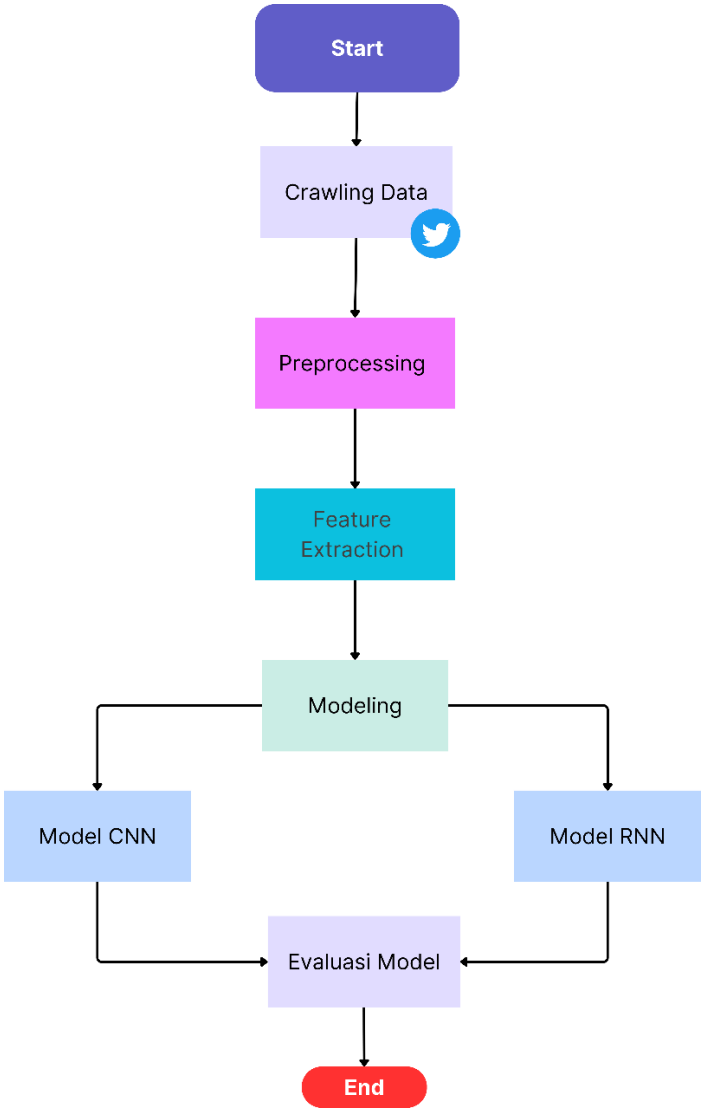
Penelitian ini termasuk dalam jenis penelitian Kuantitatif Eksperimental karena pendekatannya melibatkan percobaan terhadap variabel-variabel tertentu untuk memahami pengaruhnya terhadap hasil analisis sentimen dalam penelitian ini menggunakan 3 *feature extraction* yaitu *FastText*, CBOW dan TF-IDF. Pada penelitian ini merancang eksperimen pada beberapa *feature extraction* dalam arsitektur neural network yang digunakan, yaitu *Convolutional Neural Network* (CNN) dan *Recurrent Neural Network* (RNN) dengan tujuan untuk membandingkan kinerja keduanya.

## 3.2 Objek, Populasi dan Sample Penelitian

Objek dari penelitian ini yaitu ulasan mengenai layanan Internet First Media di Twitter, Sedangkan populasi dalam penelitian ini yaitu seluruh data yang didapatkan selama proses *Crawling* data yaitu sebanyak 48,722 data tweet Twitter mengenai opini penguna layanan Internet First Media di Twitter. Setelah melalui tahap *Preprocessing*, Sample yang digunakan dalam penelitian yaitu sebanyak 27,451 data tweet Twitter.

## 3.5 Alur Penelitian

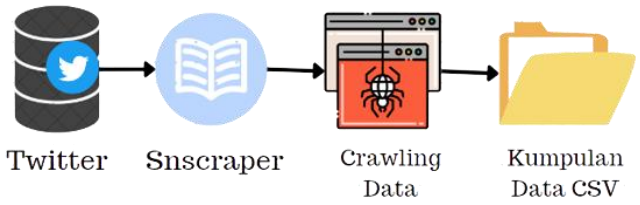
Tahapan penelitian dalam penelitian tugas akhir ini meliputi pengumpulan data, preprosessing, implementasi Metode CNN dan RNN, evaluasi kinerja dan menganalisis hasil dari data. Hasil dari penelitian ini akan disajikan dalam statistik dan digunakan untuk memberikan wawasan yang mendalam tentang sentimen pelanggan dan saran-saran untuk meningkatkan kepuasan dan kualitas layanan Internet provider First Media. Tahapan alur penelitian ini dapat dilihat pada Gambar 3.1 dibawah ini.




Gambar 3.1 Alur Penelitian

### 3.5.1. Pengumpulan Data

Pada tahap ini, akan dilakukan proses pengambilan data dari platfrom media sosial Twitter dengan menggunakan *library snscrapper*. Data yang di ambil berkaitan dengan opini pengguna media sosial Twitter terhadap layanan Internet provider First Media dan seluruh data berbahasa Indonesia. Proses pengambilan data dapat dilihat pada Gambar 3.2 di bawah ini.




Gambar 3.2 Proses *Crawling* Data

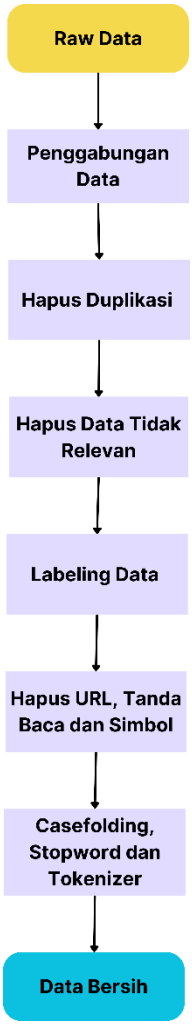
Dalam proses pengambilan data untuk memperoleh data yang sesuai, proses pengambilan data dilakukan beberapa kali menggunakan beberapa kata kunci yang relevan yang dapat dilihat pada Tabel 3.1 berikut.

Tabel 3. 1
Kata Kunci Pencarian

| No. | Kata Kunci |
|---|---|
| 1. | Internet First Media |
| 2. | First Media Cepat |
| 3. | First Media Lancar |
| 4. | First Media Lambat |
| 5. | First Media Lemot |
| 6. | First Media Gangguan |

### 3.5.2. Pra-pemrosesan Data

Pada tahap pra-pemrosesan data, data yang telah diperoleh dari proses pengumpulan data akan mengalami beberapa langkah pra-pemrosesan untuk mempersiapka data sebelum digunakan dalam proses pelatihan model. Tahapan yang dilakukan pada pra-pemrosesan data diantaranya yaitu, penggabungan data, penghapusan data duplikasi, penghapusan data tidak reelevan, labeling data, hapus url, simbol dan tanda baca, *casefolding*, *stopword* dan *tokenizer*. Visualisasi dari pra-pemrosesan data dapat dilihat dalam Diagram 3.3 berikut:




Gambar 3.3 Tahapan Pra-pemrosesan Data

1. Penggabungan Data

Pada tahap ini akan di lakukan proses penggabungan seluruh data hasil *Crawling*, karena data yang diperoleh dari proses pengumpulan data terdiri dari beberapa file csv terpisah berdasarkan kata kunci yang telah digunakan dalam proses *Crawling*. Oleh karena itu file-file tersebut akan digabungkan dan disimpan dalam satu file CSV.

2. Menghapus Data Duplikasi

Setelah semua data berhasil digabungkan menjadi satu file csv, tahap selanjutnya adalah melakukan proses deteksi duplikasi atau kesamaan konten teks pada data tweet. Data yang terdeteksi sebagai duplikasi akan dihapus secara otomatis, dan hanya satu data asli yang akan disimpan.

3. Menghapus Data Tidak Relevan

Pada tahap ini akan dilakukan penghapusan data yang tidak relevan pada *dataset*. Data yang tidak relevan dengan topik akan diidentifikasi dan dihapus dari *dataset* untuk memastikan bahwa data yang akan digunakan dalam analisis sentimen adalah data yang valid dan relevan. Berikut merupakan contoh data yang relevan dan tidak relevan dapat dilihat pada Tabel 3.2 dibawah ini

Tabel 3. 2

Identifikasi Data Tidak Relevan

| **Text** | **Relevan** | **Tidak Relevan** |
|---|---|---|
| Lowongan Kerja Marketing Firstmedia Internet dan TV Kabel https://t.co/D05jZuWh4H | | ✓ |
| Baru kali ini bangga sama first media, yang lain pada lemot, gangguan dll, First Media lancar jaya 🥰🥰🥰 luv | ✓ | |
| PROMO FIRST MEDIA BULAN INI (Internet Unlimited+Tv kabel)khusus utk 10 orang pertama( AREA :SBY,SDA DAN GRESIK )yaitu :<br>Gratis pasang,Gratis daftar,Gratis kabel 40 meter,Open all channel 3 bulan.<br>Syarat mudah ( foto ktp & email)<br>Cp : HADI<br>Tlp + Wa.085231912889<br>Buruan Daftar... | | ✓ |
| Kenapa Internet First Media lemot banget di rumah gue???😖 😖 😖 | ✓ | |

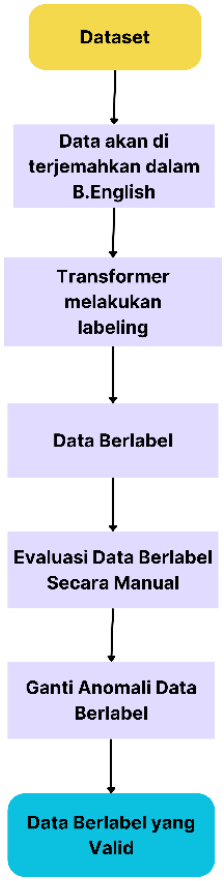
4. *Casefolding*

*Casefolding* merupakan proses untuk mengubah semua karakter dalam teks menjadi huruf kecil atau lowercase. Tujuan dari *Casefolding* adalah untuk membuat

data teks konsisten dalam hal kapitalisasi, sehingga huruf besar dan huruf kecil dianggap sama

5. Data *Labeling*

Berikut ini pada Gambar 3.4 adalah tahap alur pelanbelan *dataset* kedalam kelas sentimen positif dan negatif




Gambar 3.4 Proses Pelebelan Data

Tahap pemberian label pada data dilakukan secar automatic labeling. *Dataset* yang digunakan sebanyak 27.451 data. Langkah pertama yang dilakukan yaitu terjemahkan data kedalam bahasa Inggris, setelah itu data akan di beri pelabelan secara otomatis menggunakan teknik NLP *Transformer*. Setelah melalui tahap labeling otomatis, selanjutnya data akan di r*eview* secara manual dengan memfilter data secara random untuk memvalidasi hasil dari pemberian label pada data dengan kategori positif dan negatif apakah sudah dilakukan dengan benar. Apabila ada anomali maka label data akan diganti secar manual. Data yang telah diberikan label akan digunakan untuk modeling *Deep Learning* dengann CNN dan

RNN untuk belajar polaritas setimen dari *dataset*. Klasifikasi kelas sentimen positif dan negatif dapat dilihat pada Tabel 3.3 dibawah ini

Tabel 3. 3
Pelabelan Data

| Text | Label |
|---|---|
| Baru kali ini bangga sama *First Media*, yang lain pada lemot, gangguan dll, First Media lancar jaya luv | *Positive* |
| Alhamdullilah First Media lancar wkwk | *Positive* |
| @solehsolihun di gue First Media lancar bangett smpe tetangga yg pake indihome aja ganti fifst media | *Positive* |
| @kardusnivora First Media lancar jaya | *Positive* |
| @Firstmedia ini konsisten dan berdedikasi tinggi dalam gangguan jaringan. | *Negative* |
| @FirstMediaCares bayar selalu *ontime*, tapi Internet sering bermasalah. besok fix cabut ganti provider. mantap firstmedia batam hari gangguan terus. wifi untuk kerja padahal. | *Negative* |
| Bener bener firstmedia paling first kalo gangguan, gaada x jam gangguan lagi | *Negative* |

6. Menghapus tanda baca dan simbol

Langkah berikutnya dalam pra-pemrosesan data adalah menghapus tanda baca dan simbol dari teks dalam *dataset*. Tanda baca dan simbol seperti koma, titik, tanda tanya, tanda seru, tanda kurung, dan emotikon akan di hapus pada teks karena tidak memberikan kontribusi penting dalam analisis sentimen dan dapat mengganggu proses pemrosesan teks lebih lanjut. Khususnya, simbol yang berbentuk emotikon seperti sedih, senang dan marah harus di hapus pada teks karena dapat memiliki makna ganda dan mempengaruhi hasil analisis sentimen. Berikut adalah contoh dari proses menghapus tanda baca dan simbol yang dapat dilihat pada Tabel 3.4 di bawah ini.

Tabel 3.4

Proses Penghapusan Tanda Baca dan Simbol

| Sebelum | Sesudah |
|---|---|
| untung First Media lancar jaya walaupun ujan😭😭😭😭 | untung First Media lancar jaya walaupun ujan |
| sumpah First Media lemot banget ya allah, buka youtube aja loading terusssssss 😭😭😭 | sumpah First Media lemot banget ya allah buka youtube aja loading terusssssss |

7. *Stopword*

Pada tahap *Stopword* akan dilakukan penghapusan pada kata-kata umum yang sering muncul dalam teks tetapi tidak memberikan informasi yang berguna dalam pemahaman konten teks. Contoh *Stopword*s dalam bahasa Indonesia adalah "dan", "atau", "yang", "dari", "di", "ke", "dengan", "untuk", "saya", "kamu", "dia", "mereka", "kita", "aku", dan lain-lain. Kata-kata tersebut akan dihapus dengan tujuan untuk menyederhanakan teks dan menghilangkan kata-kata yang tidak relevan agar fokus memberikan kata-kata yang lebih informatif.

8. *Tokenizing*

Tahap tokenisasi digunakan untuk memecah teks menjadi unit-unit terkecil berupa kata, frasa, atau karakter yang disebut token. Tujuan dari tokenisasi adalah untuk mengubah teks menjadi bentuk yang lebih terstruktur untuk memudahkan dalam analisis teks dan pemrosesan lebih lanjut. Misalnya, kalimat "Internet First media lancar." akan dipecah menjadi token-token seperti "Internet", "First", "Media", "lancar".

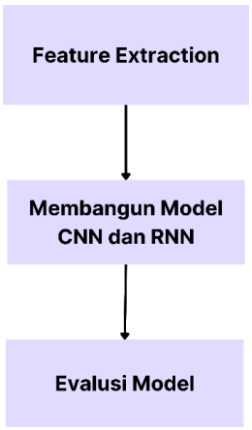
9. Data *Split*

Tahap pembagian data pada penelitian *dataset* akan di bagi menjadi tiga yaitu Data *traning*, Data *validasi dan* Data *testing*. Dengan pembagian data ini, model dapat mengatasi potensi masalah *overfitting* dan *underfitting* serta memberikan gambaran kinerja yang lebih akurat pada data baru. *Dataset* yang digunakan berjumlah 27.451 maka proses pembagian data akan di jelaskan sebagai berikut:

1. Data *Training* sebesar 64%: Sebanyak 17.568 data akan digunakan untuk proses pelatihan model. Data *Training* ini adalah yang paling banyak digunakan dalam proses pembelajaran. Model akan "belajar" dari data ini dan mencoba untuk menyesuaikan pola-pola yang ada dalam data tersebut.
2. Data Validasi 16%: Sebanyak 4.392 data akan digunakan sebagai data validasi. Data ini akan digunakan selama proses pelatihan untuk memantau kinerja model di setiap *Epoch* atau iterasi pelatihan. Validasi membantu dalam mengidentifikasi apakah model terlalu banyak mempelajari pola yang mungkin tidak umum (*overfitting*) atau terlalu umum (*underfitting*) dari data *Training*.
3. Data *Testing* 20%: Sebanyak 5.491 data akan digunakan sebagai data *Testing*. Data ini tidak digunakan selama proses pelatihan atau validasi. Namun, setelah model dilatih, data *Testing* akan digunakan untuk menguji performa model secara objektif pada data yang belum pernah dilihat sebelumnya. Hasil pengujian ini memberikan gambaran tentang seberapa baik model dapat menggeneralisasi pada situasi dunia nyata.

### 3.5.3. *Modeling*

Pada tahap *Modeling*, akan dibangun menggunakan dua metode yang berbeda, yaitu *Convolutional Neural Network* (CNN) dan *Recurrent Neural Network* (RNN), untuk melakukan analisis sentimen terhadap *dataset* yang berasal dari tweet pengguna Twitter yang berkaitan dengan opini terhadap layanan Internet provider First Media. Proses *Modeling* yang dilakukan akan dapat di lihat pada Gambar 3.5 dibawah ini.




Gambar 3.5 Alur *Modeling* CNN

1. *Feature Extraction*

Pada penelitian ini *Feature Extraction* akan digunakan ada 3 yaitu TF-IDF dan dua jenis *Word Embedding* diantaranya *CBOW dan FastText. Feature Extraction* digunakan untuk merubah kata-kata pada teks kedalam representatif vektor numerik yang bermakna.

2. Pembangunan Model CNN

Dalam membangun medel CNN pada penelitian ini menggunakan *Layer* CNN sebagai berikut:

1. *Layer Embedding* , Ini adalah *Layer* pertama dalam model. Fungsi utama dari *Layer* ini adalah untuk mengubah indeks kata dalam kalimat menjadi vektor dengan dimensi tertentu.
2. *Layer* Conv1D, *Layer* konvolusi satu dimensi. Fungsi utama dari *Layer* ini adalah untuk mengidentifikasi pola-pola fitur dalam data dengan menggunakan filter konvolusi. Pada *Layer* ini menggunkana fungsi aktivasi ReLU (*Rectified Linear Unit*)
3. *Layer MaxPooling1D*, *Layer pooling* yang digunakan untuk mengurangi dimensi dari keluaran *Layer* konvolusi. *MaxPooling* akan mengambil nilai maksimum dalam setiap kelompok nilai dari keluaran *Layer* konvolusi, membantu mengurangi kompleksitas dan mencegah *overfitting*.
4. *Layer Flatten*, Ini adalah langkah untuk mengubah *output* dari *Layer pooling* menjadi bentuk vektor satu dimensi yang dapat dihubungkan ke *Layer Dense* berikutnya.
5. *Layer Dropout*, digunakan untuk mencegah *overfitting*
6. *Layer Dense*, Ini adalah *Layer fully connected*. dengan fungsi aktivasi softmax. *Layer* ini melakukan klasifikasi berdasarkan fitur yang telah dipelajari oleh *Layer* sebelumnya dan menghasilkan probabilitas distribusi kelas sentimen yang mungkin.

3. Membangun Model RNN

   Pada modeling RNN pada penelitian ini menggunakan beberapa *Layer* RNN sebagai berikut

1. *Layer Embedding* , *Layer* pertama untuk mengubah indeks kata dalam kalimat menjadi vektor dengan dimensi tertentu. Disini lah penerapan Feature Extraction.
2. *Layer* LSTM (*Long Short-Term Memory*), LSTM mampu menangkap konteks temporal yang panjang dalam teks. Setiap langkah waktu dalam LSTM memproses input saat ini, *output* dari langkah sebelumnya, dan memori internal.
3. *Layer Dropout*, digunakan untuk mencegah *overfitting* dalam model.
4. *Layer Dense*, *Layer output* yang menghasilkan probabilitas distribusi kelas sentimen yang mungkin. *Layer* ini menggunakan aktivasi softmax, untuk mengonversi nilai-nilai dari neuron dalam *Layer* ini menjadi probabilitas untuk masing-masing kelas sentimen.

### 3.5.4. Evaluasi Model

Dalam tahap Evaluasi Model, model yang telah dilatih akan diuji dan dievaluasi kinerjanya menggunakan metrik evaluasi untuk mengevaluasi performa kedua model yaitu *Accuracy*, *Precision*, *Recall*, dan *F1-score*.

1. *Accuracy* mengukur sejauh mana model dapat melakukan klasifikasi dengan benar secara keseluruhan.
2. *Precision* mengukur sejauh mana model memberikan hasil positif yang benar dari seluruh hasil positif yang diprediksi.
3. *Recall* mengukur sejauh mana model dapat mengidentifikasi dengan benar seluruh hasil positif dari seluruh data positif yang sebenarnya.
4. Sedangkan *F1-score* merupakan harmonic mean dari *Precision* dan *Recall*, yang memberikan gambaran keseluruhan tentang performa model. Evaluasi model akan memberikan gambaran seberapa baik kedua model ini dapat memprediksi sentimen dari tweet pengguna terhadap layanan Internet First Media.

Evaluasi model bertujuan untuk mengukur sejauh mana model dapat memahami dan memberikan klasifikasi yang akurat pada data baru yang tidak ada dalam *dataset* pelatihan. Berikut adalah beberapa langkah yang dilakukan dalam tahap Evaluasi Model.

### 3.5.5. Perangkat Penelitian

Pada Penelitian menggunakan metode CNN dan RNN untuk mengklasifikasi sentimen teks pada data twitter. Seluruh proses pengolahan data dalam penelitian ini menggunakan platfrom Google Colaboratory. Pada Tabel 3.5 dibawah ini akan menyajikan informasi mengenai perangkat lunak yang digunakan dalam penelitian.

Tabel 3. 5

*Software* yang di Gunakan

| **Bahasa Pemrograman** | **Python** |
|---|---|
| *Editor* | *Google Colaboratory* |
| *Library* | *Pandas, Numpy, TensorFlow, Keras, Gensim , sklearn., Scikit-Learn* |