

BAB III

METODOLOGI PENELITIAN

3.1 Jenis dan Sumber Data

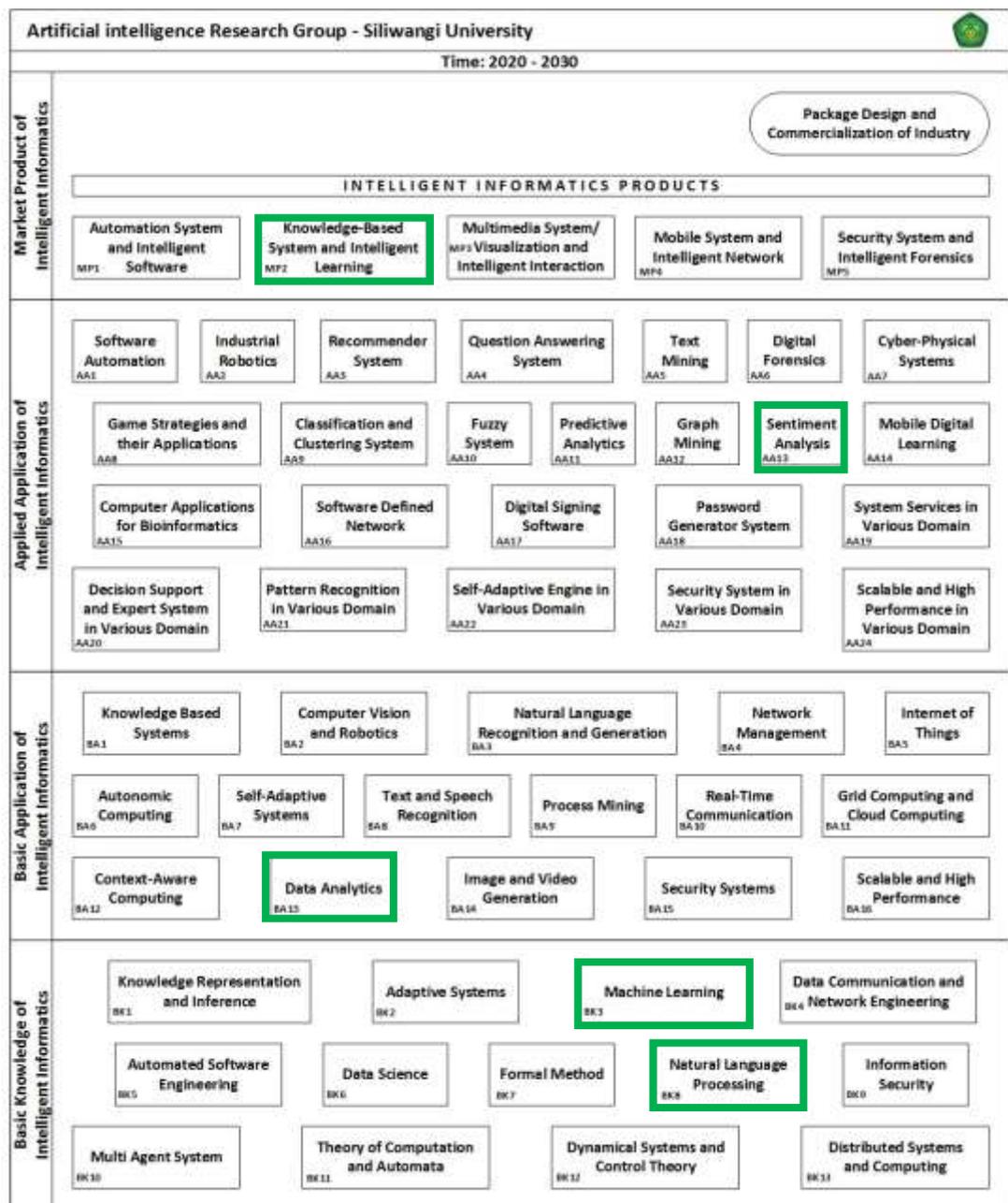
Jenis data yang digunakan pada penelitian ini adalah data primer. Data primer merupakan data yang diperoleh langsung dari sumber datanya (Faisal et al., 2020). Data ulasan aplikasi Al-Qur'an Indonesia berasal dari situs *Google Play Store* yang diambil dengan metode *web scraping* yang dilakukan dengan menggunakan *library* dari *python* yaitu *google-play-scraper*. *Scraping* data dilakukan pada tanggal 9 Agustus 2023. Data yang diambil untuk penelitian ini berada pada rentang waktu 1 Januari sampai 9 Agustus tahun 2023. Jumlah data yang terkumpul pada rentang waktu tersebut sebanyak 13.436 data ulasan berbahasa Indonesia. Tipe data yang dikumpulkan adalah sebagai berikut:

Tabel 3.1 Tipe Data

Kolom	Tipe Data
Username	Object
Score	Integer
At	Datetime
Content	Object

3.2 Peta Jalan Penelitian

Peta Jalan Penelitian atau *Roadmap* Penelitian merupakan peta yang memetakan perjalanan dari suatu penelitian. Berikut adalah *Roadmap* pada penelitian ini.

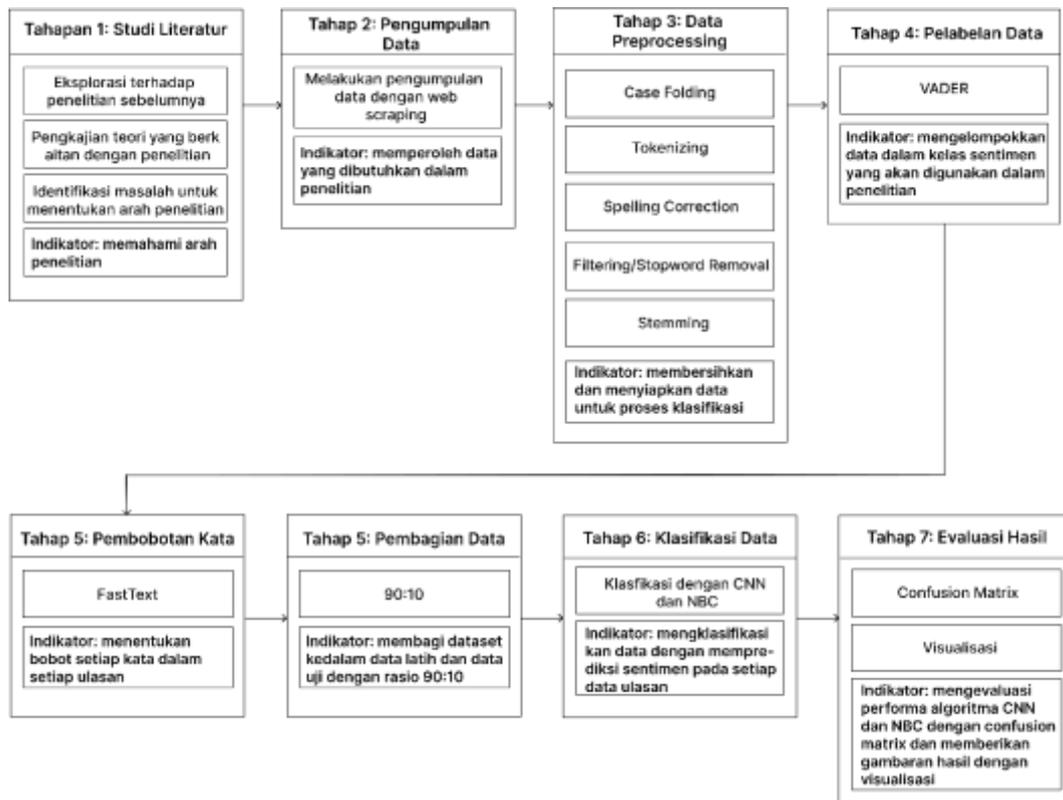


Gambar 3.1 *Roadmap* Penelitian

Berdasarkan Gambar 3.1 terdapat beberapa kajian yang berkaitan dengan penelitian yang akan dilakukan. Kajian yang berkaitan ditandai dengan kotak hijau, diantaranya *Natural Language Processing (NLP)*, *Machine Learning*, *Data Analytics*, *Sentiment Analysis* dan *Knowledge-Based System and Intellegent Learning*. dalam penelitian ini yaitu terkait perbandingan algoritma *machine learning* dan *deep learning* pada analisis sentimen. Pada tahap inisiasi penelitian dimulai dengan studi literatur terkait proses penerapan algoritma *machine learning* dan *deep learning* dalam analisis sentimen, dan juga pengumpulan data yang akan digunakan dalam proses penelitian. Tahap pengembangan merupakan tahap menyiapkan dan membersihkan data yang selanjutnya akan digunakan dalam Tahap Lanjut yaitu dalam proses implementasi algoritma Naïve Bayes Classifier dan Convolutional Neural Network dalam analisis sentimen.

3.3 Tahapan Penelitian

Berikut merupakan tahapan dari penelitian yang dimulai dengan studi literatur, pengumpulan data, *data preprocessing*, pelabelan data, pembobotan kata, klasifikasi data dan evaluasi hasil seperti yang tertera pada gambar 3.2



Gambar 3.2 Tahapan Penelitian

Berdasarkan Gambar 3.2 Tahapan penelitian yang dilakukan dijelaskan sebagai berikut.

1. Studi Literatur

Studi literatur dilakukan dengan mengkaji beberapa penelitian sebelumnya dari beberapa sumber seperti jurnal, buku serta laporan penelitian yang berkaitan dengan analisis sentimen, algoritma *Naïve Bayes Classifier*, *Convolution Neural Network*, pelabelan lexicon dan aplikasi Al-Qur'an digital. Tahapan ini dilakukan untuk mengetahui arah dari penelitian yang dilakukan

2. Pengumpulan Data

Pengumpulan data dilakukan dengan teknik *web scraping* yang dilakukan dengan menggunakan *library* dari *python* yaitu *google-play-scraper* dengan

parameter *lang= 'id', country = 'id', sort = Sort.NEWEST*. *Scraping* data dilakukan pada tanggal 9 Agustus 2023. Data hasil *scraping* berjumlah 13.436 data ulasan. Pengumpulan data dilakukan dengan tujuan untuk memperoleh data yang diperlukan dalam proses penelitian.

3. *Data Preprocessing*

Data Preprocessing dilakukan untuk membersihkan dan menyiapkan data untuk proses klasifikasi. Pada penelitian ini, *data preprocessing* dilakukan dengan beberapa proses diantaranya *Case Folding, Tokenizing, Spelling Correction, Filtering/Stopword Removal* dan *Stemming*. Pada tahap *case folding* seluruh data diubah kedalam bentuk *lower case* atau huruf kecil, dan karakter selain huruf a-z dihapus. Pada tahap *tokenizing*, data dibagi kedalam bagian yang lebih kecil yang biasa disebut dengan token. Pada tahap *spelling correction*, dilakukan perbaikan untuk kata-kata seperti singkata, *typo* atau bahasa gaul yang jika tidak diperbaiki dapat mempengaruhi hasil sentimen, dengan *spelling correction* kata-kata tersebut akan diubah menjadi kata baku. Pada tahap *filtering*, dilakukan penghapusan untuk kata-kata yang tidak mempengaruhi nilai sentimen, kata tersebut akan dihapus karena dianggap sebagai *noise*. Dan pada tahap *stemming*, setiap kata akan diubah menjadi kata dasarnya dengan cara menghapus setiap imbuhan yang ada.

4. Pelabelan Data

Pelabelan data dilakukan dengan tujuan untuk memberikan label pada setiap data ulasan dengan kelas sentimen. Data ulasan dapat memiliki kelas positif, negatif atau netral. Pelabelan diberikan berdasarkan kata pada ulasan.

Pelabelan dilakukan secara otomatis dengan menggunakan *library* python yaitu VADER. VADER akan membagi data kedalam kelas sentimen berdasarkan dengan *compound score*.

5. Pembobotan Kata

Pembobotan kata dilakukan untuk menentukan bobot kata pada setiap data ulasan. Metode yang digunakan untuk pembobotan kata adalah *FastText*. *FastText* merupakan metode yang mampu mengatasi masalah OOV (*Out of Vocabulary*) dengan cara mengurai setiap kata kedalam *n-gram*, ini membuat *FastText* memungkinkan untuk menangkap makna kata-kata dan memahami sufiks serta prefiks kata. Nilai vektor untuk setiap kata merupakan penjumlahan dari nilai vektor setiap *n-gram*-nya.

6. Klasifikasi Data

Klasifikasi data meliputi aktivitas untuk membagi data menjadi data latih dan data uji. Data latih digunakan untuk membangun dan melatih model klasifikasi. Sementara data uji digunakan untuk menguji model klasifikasi untuk memperoleh tingkat akurasi algoritma *Naive Bayes Classifier* dan *Convolutional Neural Network*. Pada penelitian ini, data dibagi kedalam data latih dan data uji dengan rasio 90:10, dimana 90% sebagai data latih dan 10% sebagai data uji.

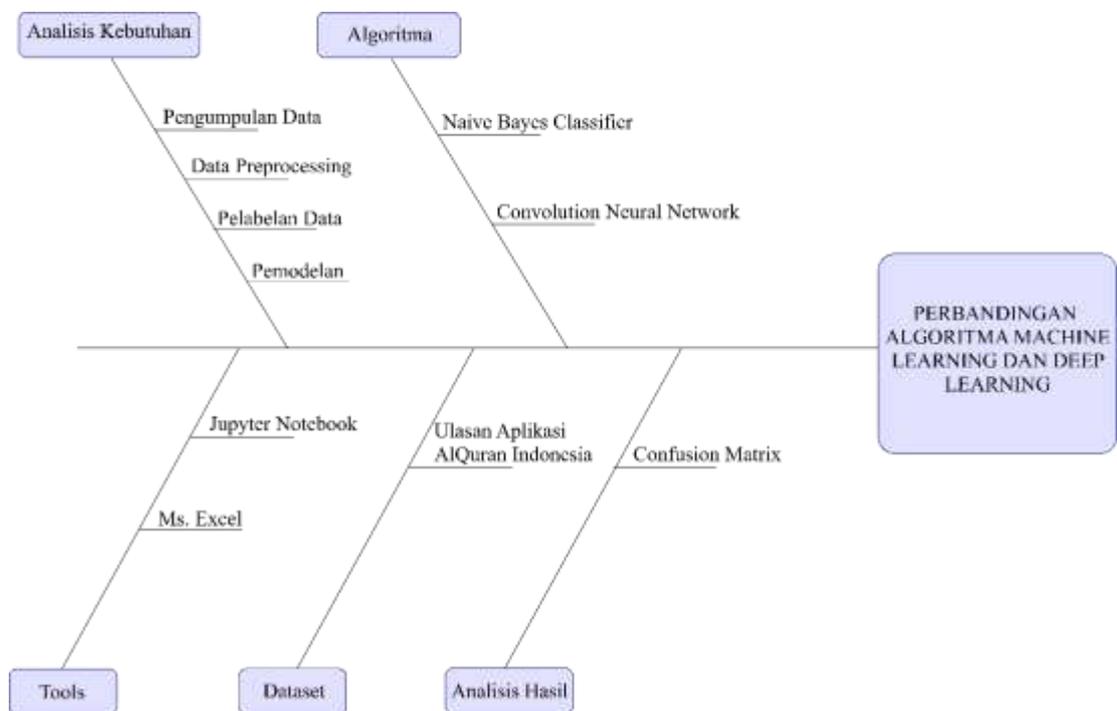
7. Evaluasi Hasil

Evaluasi hasil meliputi proses menghitung nilai akurasi algoritma CNN dan NBC dengan menggunakan *confusion matrix* dan memberikan gambaran hasil klasifikasi sentimen dengan visualisasi berdasarkan data penelitian yang

telah dikumpulkan. Dari hasil analisis menggunakan *confusion matrix*, metode CNN mendapatkan akurasi sebesar 93,89%, presisi 93,70%, *recall* 93,89% dan *f1-score* 93,75%. Dan dari total 1.278 data uji yang digunakan, CNN berhasil mengklasifikasikan 1.200 data ulasan dengan benar, dimana 1.019 ulasan termasuk sentimen positif, 169 ulasan termasuk sentimen netral, dan 12 ulasan termasuk sentimen negatif. Sementara metode NBC mendapatkan akurasi 76,83%, presisi, 85,40%, *recall* 76,83% dan *f1-score* 79,90%. NBC berhasil mengklasifikasikan 982 data ulasan dengan benar, dimana 829 ulasan termasuk sentimen positif, 137 ulasan termasuk sentimen netral, dan 16 ulasan termasuk sentimen negatif.

3.4 Fishbone Diagram

Berikut merupakan diagram *fishbone* pada penelitian ini



Gambar 3.3 Diagram *Fishbone*

Berdasarkan Gambar 3.3 Diagram *Fishbone* pada penelitian ini dijelaskan sebagai berikut.

1. Analisis Kebutuhan

Analisis kebutuhan dilakukan untuk menganalisis aktivitas yang dilakukan selama proses penelitian. Penelitian didasarkan pada proses pengumpulan data yang dilakukan dengan teknik *web scraping*, kemudian tahap *Data Preprocessing* yang meliputi tahapan *case folding*, *tokenizing*, *spelling correction*, *filtering* dan *stemming*. Pada tahap pelabelan, data diberikan label berdasarkan kata yang terdapat pada setiap ulasan. Selanjutnya, data diklasifikasikan menggunakan algoritma *Naïve Bayes Classifier* dan *Convolutional Neural Network*.

2. Algoritma

Proses klasifikasi pada penelitian ini menggunakan algoritma *machine learning* dan *deep learning*, yaitu *Naïve Bayes Classifier* dan *Convolutional Neural Network*.

3. Tools

Dalam mendukung proses penelitian yang dilakukan terdapat beberapa *tools* yang digunakan, yaitu *Microsoft Excel* untuk membaca dan mengecek hasil *scraping* yang memiliki format *.csv*. Selain itu, untuk proses pengumpulan data, *data preprocessing*, pelabelan data dan klasifikasi digunakan aplikasi *Jupyter Notebook*.

4. Dataset

Dataset yang digunakan pada penelitian ini diambil dari ulasan aplikasi Al-Qur'an Indonesia pada situs *Google Play Store* yang diambil dengan menggunakan teknik *web scraping*.

5. Analisis Hasil

Analisis hasil dilakukan dengan tujuan menganalisis perbandingan performa kedua algoritma klasifikasi yang digunakan. Proses analisis hasil dilakukan dengan menggunakan *Confusion Matrix*.