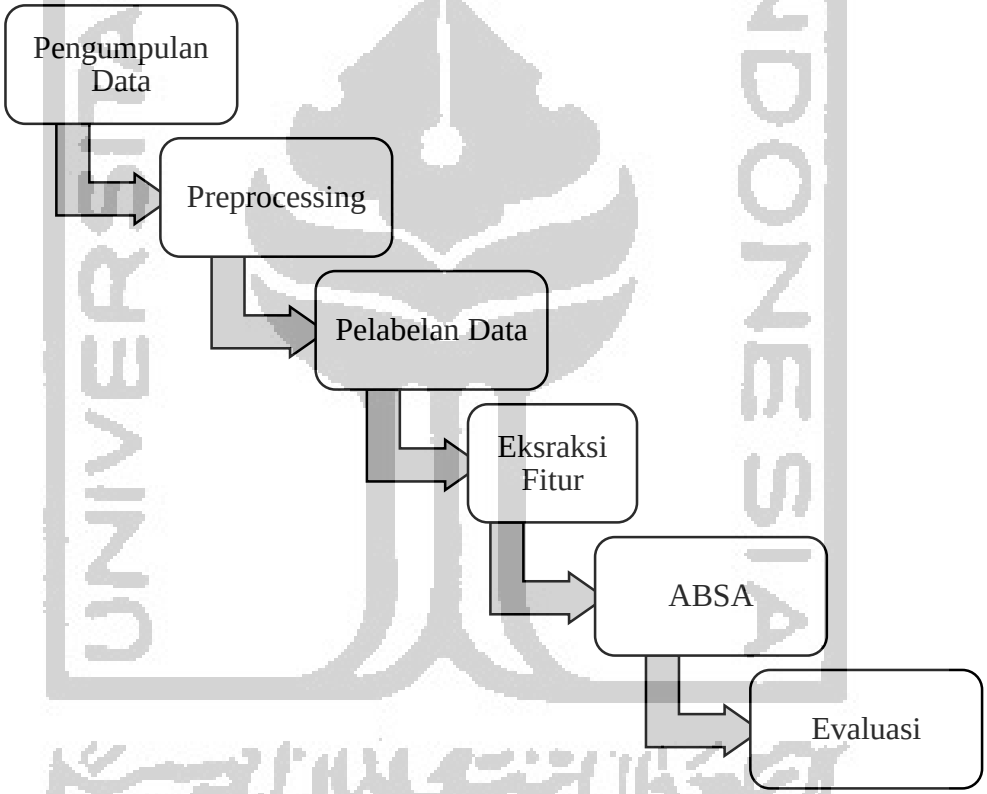
BAB III
METODOLOGI PENELITIAN

3.1 Langkah Pengerjaan Tugas Akhir

Berikut ini merupakan langkah-langkah yang digunakan dalam mengerjakan analisis sentimen berbasis fitur. Langkah-langkah pengerjaan tugas akhir ini diawali dengan pengumpulan data, kemudian diikuti *preprocessing*, pelabelan data, ABSA, dan diakhiri dengan evaluasi. Gambar 3.1 di bawah ini menunjukkan langkah-langkah untuk ABSA.



Gambar 3.1 Langkah pengerjaan tugas akhir

3.2 Uraian Metodologi

3.2.1 Pengumpulan Data

Pengumpulan data dilakukan dengan mengambil ulasan dari situs TripAdvisor (tripadvisor.co.id) dan menyimpan data tersebut dalam file berekstensi .csv. Data diambil secara acak pada pariwisata di seluruh Indonesia. Setelah data didapatkan, maka data dipisahkan menjadi kalimat-kalimat untuk diolah selanjutnya. Data ulasan akan digunakan

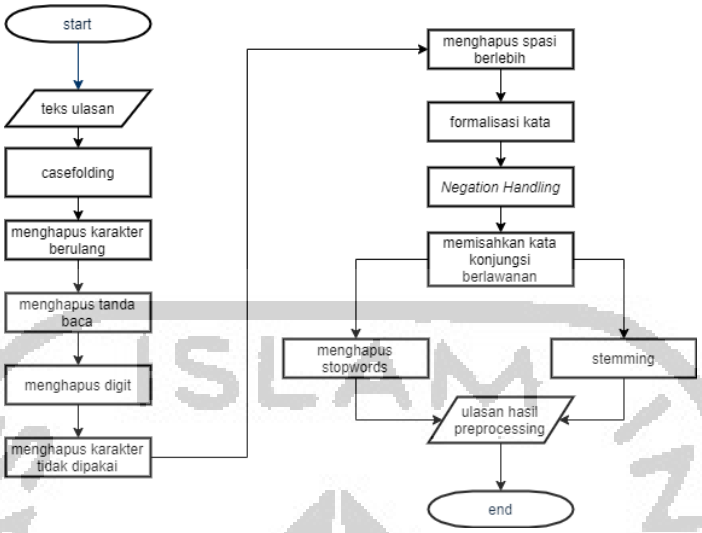
sebagai data pelatihan untuk pembuatan model dan data uji untuk melihat hasil akhir dari klasifikasi dan uji akurasi.

Pemilihan label untuk setiap kalimat ulasan dipengaruhi oleh penelitian yang dilakukan oleh Nurifan, Sarno, & Sungkono (2019) yang mengkategorikan kalimat ulasan restoran yang sebelumnya telah dilabeli oleh annotator profesional. Kategori yang digunakan pada penelitian mereka yaitu *service* (pelayanan), *physical environment* (lingkungan fisik), *food quality* (kualitas makanan), dan *price fairness* (harga). Pada penelitian ini hanya digunakan beberapa aspek kategori yang dipengaruhi *physical environment* pada penelitian sebelumnya, sehingga label dibagi menjadi fasilitas, suasana dan lokasi.

3.2.2 *Preprocessing*

Preprocessing termasuk langkah yang krusial untuk dilakukan karena pada langkah ini akan dilakukan penghapusan kata ataupun simbol yang tidak diperlukan pada kalimat ulasan yang diberikan. Bentuk *preprocessing* yang dilakukan antara lain *casefolding*, menghapus karakter berulang, menghapus tanda baca, menghapus digit, menghapus karakter tidak dipakai, menghapus spasi berlebih, formalisasi kata, *negation handling*, memisahkan kata konjungsi berlawanan, menghapus *stopwords* dan *stemming*.

Pada proses ini harus dilakukan secara berurutan dan tidak bisa dilakukan secara acak, oleh karena itu, susunan *preprocessing* pada setiap kasus memiliki susunan yang berbeda bergantung dengan kebutuhan pembersihan data pada data yang dimiliki. Gambar 3.2 dibawah ini merupakan susunan *preprocessing* yang digunakan pada kasus ABSA pariwisata.



Gambar 3.2 Urutan langkah pembersihan data

a. *Casefolding*

Casefolding digunakan untuk menyamaratakan semua huruf pada teks menjadi huruf kecil atau besar seluruhnya. Pada kasus ini, semua huruf teks diubah menjadi huruf kecil. Tabel 3.1 merupakan contoh *casefolding* dengan menyamaratakan semua huruf menjadi huruf kecil.

Tabel 3.1 Contoh penerapan proses *casefolding*

Sebelum	Sesudah
Pengunjung di Candi Borobodur cukup banyak dan padat.	pengunjung di candi borobodur cukup banyak dan padat.
Kami mengunjungi Candi Borobudur pada siang-sore hari di bulan Desember.	kami mengunjungi candi borobudur pada siang-sore hari di bulan desember.
ditepi pantai kami disuguhi tarian KECAK yang bisa ditonton secara gratis	ditepi pantai kami disuguhi tarian kecak yang bisa ditonton secara gratis

b. Menghapus karakter berulang

Menghapus karakter berulang yaitu pembersihan data dengan menghilangkan beberapa huruf yang muncul secara berulang sebanyak lebih dari 2 kali. Tabel 3.2 merupakan beberapa contoh menghapus karakter berulang pada data pariwisata.

Tabel 3.2 Contoh penerapan penghapusan karakter berulang

Sebelum	Sesudah
jogja tunggu aku yaaaaaaaaaa.	jogja tunggu aku ya
Menjelang matahari tenggelam juga kereee eennn banget pemandangan nya.	Menjelang matahari tenggelam juga keren banget pemandangan nya.
hiduuup sultan dan panjang umur buar pak Sultan TOOOPPPP!!!	hidup sultan dan panjang umur buar pak Sultan TOP!

c. Menghapus tanda baca

Tahap ini digunakan untuk menghilangkan tanda baca pada data teks seperti !@#$%^&*() dan tanda baca lain. Tabel 3.3 menunjukkan contoh penggunaan penghapusan tanda baca.

Tabel 3.3 Contoh penerapan menghapus tanda baca

Sebelum	Sesudah
Tiket masuknya relatif murah :)	Tiket masuknya relatif murah
Situs budaya ini masih satu wilayah dengan salah satu candi yang sangat fenomenal ' Candi Prambanan'	Situs budaya ini masih satu wilayah dengan salah satu candi yang sangat fenomenal Candi Prambanan
saya datang ke candi ini jam setengah 6 pagi!!	saya datang ke candi ini jam setengah 6 pagi

d. Menghapus digit

Tahap ini digunakan untuk menghapus digit 0-9 yang terkandung pada data teks. Tahap ini diperlukan karena tidak adanya label “harga”, sehingga digit pada ulasan tidak terlalu signifikan untuk penelitian ini. Tabel 3.4 di bawah ini merupakan contoh perubahan dari pengaplikasian menghapus digit.

Tabel 3.4 Contoh penerapan menghapus digit

Sebelum	Sesudah
saya datang ke candi ini jam setengah 6 pagi	saya datang ke candi ini jam setengah pagi
Berjarak tempuh sekitar 40km dari Kota Pa ngkalpinang	Berjarak tempuh sekitar km dari Kota Pangkalpinang
konon pembangunan Kuil ini menghabiskan dana sekitar 13 milyar	konon pembangunan Kuil ini menghabiskan dana sekitar milyar

e. Menghapus karakter tidak dipakai

Tahap ini digunakan untuk menghapus karakter yang tidak dipakai. Karakter tidak dipakai ditetapkan jika karakter berjumlah kurang dari 4 angka. Tabel 3.5 merupakan contoh dari penerepan menghapus spasi berlebih.

Tabel 3.5 Contoh penerapan karakter tidak dipakai

Sebelum	Sesudah
Disini indah sekali ya	Disini indah sekali
Yang paling menarik, di pulau ini kita bisa menemukan menara mercusuar.	Yang paling menarik, pulau kita bisa menemukan menara mercusuar.
wah ternyata candi borobudur megah sekali	ternyata candi borobudur megah sekali

f. Menghapus spasi berlebih

Tahap ini digunakan untuk menghapus data teks yang memiliki spasi berlebih. Tabel 3.6 merupakan contoh dari penerepan menghapus spasi berlebih.

Tabel 3.6 Contoh penerapan menghapus spasi berlebih

Sebelum	Sesudah
Masyarakat setempat menyebutnya sebagai Kuil Shaolin.	Masyarakat setempat menyebutnya sebagai Kuil Shaolin.
Indah banget !!!	Indah banget !!!
Bangunan Viharanya besar, bagus/ indah dan megah.	Bangunan Viharanya besar, bagus/indah dan megah.

g. Formalisasi kata

Formalisasi kata merupakan tahap untuk mengubah kata yang merupakan kata tidak baku menjadi kata baku. Tabel 3.7 merupakan contoh dari penerapan formalisasi kata.

Tabel 3.7 Contoh penerapan formalisasi kata

Sebelum	Sesudah
Mmg pantai ini sgt indah!	Memang pantai ini sangat indah
Berada di keraton ini (yg penuh dengan bar ang2 kuno) membuat kita seperti kembali k e zaman dahulu kala.	Berada di keraton ini (yang penuh dengan barang2 kuno) membuat kita seperti kembali ke zaman dahulu kala.
ternyata keren bgt.	ternyata keren banget.

Sedangkan Tabel 3.8 merupakan beberapa kamus kata informal beserta kata formalnya yang digunakan untuk mengubah data teks agar menjadi lebih mudah untuk diolah pada tahap selanjutnya

Tabel 3.8 Kamus kata tidak baku beserta kata baku

Sebelum	Sesudah
smpe	sampai
smpt	sempat
pp	pulang pergi
priksa	periksa
puanas	panas
mksh	terima kasih
kudu	harus
yg	yang
lansia	lanjut usia
renov	renovasi

.

h. *Negation Handling*

Pada tahap ini dilakukan *negation handling*. *Negation* Handling yaitu menggabungkan kata yang didalamnya mengandung salah satu kata “tidak”, “kurang”, “jangan”, atau “bukan”

yang diikuti oleh kata berikutnya. Tabel 3.9 merupakan contoh dari *negation* handling yang digunakan pada kasus ini.

Tabel 3.9 Contoh penerapan *negation handling*

Contoh data mentah	‘Koleksi di museum ini tidak masif.’
Preprocessing	‘koleksi museum tidak masif’
***Negation* handling (Neg_Term)**	‘koleksi museum tidak_masif’’

i. Memisahkan Kata Konjungsi Berlawanan

Pada tahap ini dibutuhkan agar dapat memisahkan kalimat yang bisa jadi mengandung lebih dari 1 aspek atau 1 sentimen. Kata konjungsi yang digunakan untuk memisahkan kalimat yaitu kata “tetapi”, “tapi”, “meskipun”, “walaupun”, “padahal”, “namun”. Tabel 3.10 merupakan contoh dari memisahkan kata konjungsi berlawanan pada kasus ini.

Tabel 3.10 Contoh pemisahan kalimat dengan kata konjungsi berlawanan

Sebelum	**Sesudah**
"indah sekali pemandangannya meskipun fasilitas disana tidak terlalu memadai"	['indah sekali pemandangannya', 'fasilitas disana tidak terlalu memadai']
"terdapat banyak penjual yang menyediakan suvenir bagus tetapi penjualnya sangat memaksa sehingga mengganggu"	['terdapat banyak penjual yang menyediakan suvenir bagus', 'penjualnya sangat memaksa sehingga mengganggu']

j. Menghapus *stopwords*

Menghapus *stopwords* diperlukan untuk menghapus kata-kata yang tidak terlalu berpengaruh dalam proses mendapatkan sentimen ulasan. Tahap bagian ini hanya diperlukan untuk mengolah teks untuk mendapatkan sentimen saja. Daftar *Stopwords* ditentukan dengan melihat frekuensi kata yang sering muncul pada data ulasan yang dimiliki. Tabel 3.11 merupakan contoh dari penghapusan *stopwords* pada kasus ini.

Tabel 3.11 Contoh penggunaan menghapus *stopwords*

Sebelum	Sesudah
'terdapat banyak penjual yang menyediakan souvenir'	'penjual menyediakan souvenir'
'penjualnya sangat memaksa sehingga mengganggu'	'penjualnya memaksa mengganggu'

k. *Stemming*

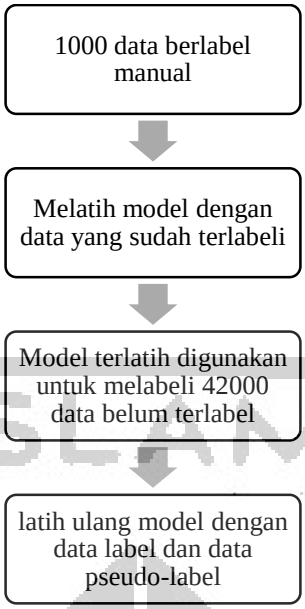
Stemming diperlukan untuk menghapus kata-kata yang tidak terlalu berpengaruh dalam proses mendapatkan aspek kategori ulasan. Tahap bagian ini hanya diperlukan untuk mengolah teks untuk mendapatkan aspek kategori saja. Tabel 3.12 merupakan contoh dari penghapusan *stopwords* pada kasus ini.

Tabel 3.12 Contoh penerapan *stemming*

Sebelum	Sesudah
'terdapat banyak penjual yang menyediakan souvenir'	['dapat banyak jual yang sedia souvenir']
'penjualnya sangat memaksa sehingga mengganggu'	['jual sangat paksa sehingga ganggu']

3.2.3 Pelabelan Data

Pelabelan dilakukan pada data training setelah data berhasil didapatkan. Pelabelan data yang diperlukan yaitu pelabelan sentimen dan juga aspek kategori. Pelabelan sentimen ulasan dibagi menjadi 2, positif dan negatif. Sedangkan pelabelan aspek kategori dibagi menjadi lokasi, fasilitas, serbaneka, dan suasana. Metode yang digunakan untuk melakukan pelabelan data yaitu *pseudo-labeling*, yaitu memanfaatkan sejumlah data untuk memprediksi data lain yang belum terlabeli.



Gambar 3.3 Langkah pseudo-labeling pada data pariwisata

Gambar 3.3 menunjukkan langkah untuk mendapatkan *pseudo-label* pada data yang belum terlabeli. Data label yang digunakan untuk melatih model *pseudo* berjumlah 1000 data yang dilabeli secara manual. Setiap data tersebut memiliki satu label aspek, dan satu label sentimen. Jumlah data pada setiap label tidak sama, sehingga data pada 1000 data di atas *imbalance* (tidak berimbang), maka dari itu dilakukan teknik *oversampling* untuk membuat data pada setiap kelas seimbang. Model ini menggunakan metode CNN untuk mendapatkan *pseudo-label*. Dalam membuat model CNN, digunakan menggunakan 64 *embedding dimentions*, 64 *batch size* dan memanfaatkan *early stopping* dengan memonitor nilai *val_loss*. Untuk mengatasi *overfitting*, digunakan juga teknik *dropout*. Nilai *kernel* yang digunakan yaitu 9, *filter* 200, *hidden_dims* 128, dan *dropout* 0.25 (Wallace, 2014).

Tabel 3.13 merupakan beberapa data dengan label *pseudo* yang dihasilkan setelah melalui pelatihan model.

Tabel 3.13 Contoh hasil pseudo-labeling

Data	Sentimen	Aspek
pemandangan terbaik kalau datang saat matahari terbit antara pagi	Positif	Suasana
candi borobudur merupakan warisan dunia yang wajib dikunjungi ketika berada yogyakarta	Positif	Serbaneka
sekarang candi borobudur benar benar bersih adannya pengetatan kebersihan membuat candi borobudur semakin terlihat teratur baik	Positif	Lokasi
tapi yang membuat tidak nyaman adalah pedagang yang selalu mengejar kita padahal sudah menolak	Negatif	Fasilitas
tapi sayang banyak tangan jahil yang tidak bertanggung jawab merusak sebagian arca	Negatif	Serbaneka

3.2.4 Ekstraksi Fitur

POS *tagging* (*part-of-speech tagging*) yaitu proses untuk menandai kata-kata dalam sebuah teks/korpus sesuai dengan bagian tertentu dari pembicaraan, berdasarkan definisi dan konteksnya. Ekstraksi fitur dilakukan untuk masing-masing aspek dan sentimen. Pada ekstraksi sentimen, digunakan tag VB, NEG, RB, dan JJ. Sedangkan pada ekstraksi aspek digunakan tag NN, VB, dan NNP. Tidak hanya POS *Tag*, pada tahap ini juda dilakukan pengambilan kembali kata yang sudah melalui tahap *negation handling*. Tabel 3.14 merupakan contoh penggunaan POS Tagging, dan Tabel 3.15 merupakan tabel keterangan dari tag yang digunakan untuk ekstraksi aspek dan sentimen.

Tabel 3.14 Contoh Penggunaan POS Tagging

Data Mentah	Hasil *preprocess* dan POS Tag	Ekstraksi Sentimen	Ekstraksi Aspek
Udara sangat sejuk di pegunungan	[('Udara', 'NN'), ('Sangat', 'RB'), ('sejuk', 'VB'), ('pegunungan', 'NN')]	Sangat sejuk	Udara pegunungan
kendaraan menuju lokasi sepi serta mudah ada dari semua arah	[('kendaraan', 'NN'), ('menuju', 'VB'), ('lokasi', 'NN'), ('sepi', 'NN'), ('mudah', 'JJ'), ('arah', 'NN')]	Menuju mudah	Kendaraan lokasi sepi arah

Tabel 3.15 Keterangan Tag

Tag	Keterangan
NEG	Negasi
RB	Adverb (Sebagai kata keterangan kata sifat/kata kerja)
JJ	Kata sifat
VB	Kata kerja
NN	Kata benda
NNP	Kata benda yang menggunakan huruf kapital di awal kata

3.2.5 Aspect Based Sentiment Analysis (ABSA)

Pada langkah ini akan dilakukan ekstraksi aspek dan ekstraksi sentimen dalam ulasan pariwisata. Kemudian menyiapkan model untuk mendapatkan aspek kategori, serta model untuk mendapatkan sentimen. Analisis sentimen berbasis fitur berfokus pada ekstraksi kalimat atau ulasan dengan aspek dan nilai polaritas sentimennya. Model yang digunakan yaitu memanfaatkan metode convolutional neural network (CNN).

Pada metode CNN digunakan beberapa layer, yaitu *embedding*, konvolusi, *pooling*, dan lapisan *dense* atau *fully connected* (FC) *layer*. Lapisan *embedding* digunakan untuk memberikan representasi kata yang padat dan makna relatifnya. *Embedding* akan memetakan setiap kata menjadi ruang vektor berkelanjutan sehingga membetuk sebuah kosa kata yang berkelanjutan dan terdistribusi. Lapisan konvolusi digunakan untuk ekstraksi fitur. Didalamnya dilakukan perhitungan untuk mendapatkan keluaran yang disebut dengan *feature map*, sehingga didapatkan *feature map* yang merepresentasikan karakteristik masukan. Lapisan

pooling secara progresif mengurangi ukuran spasial dari representasi untuk mengurangi jumlah parameter dan perhitungan dalam jaringan. Lapisan FC berperan sebagai lapisan keluaran dari jaringan syaraf tiruan.

Pembuatan model CNN untuk mendapatkan hasil terbaik ditambahkan 2 model berbeda yaitu CNN + LSTM dan CNN + GRU. *Convolutional Neural Networks* (CNN) menawarkan keuntungan dalam memilih fitur yang baik, *Long-Short Term Memory* (LSTM) telah membuktikan kemampuan yang baik untuk belajar data sekuensial (Alayba, England, & Iqbal, 2018). Baik LSTM maupun *Gated Recurrent Unit* (GRU) sering digunakan untuk representasi data sekuensial (He & Lu, 2019). Seluruh model skenario yang digunakan menggunakan 64 *embedding dimentions*, 64 *batch size* dan memanfaatkan *early stopping* dengan memonitor nilai val_loss. Untuk mengatasi *overfitting*, digunakan juga teknik *dropout*.

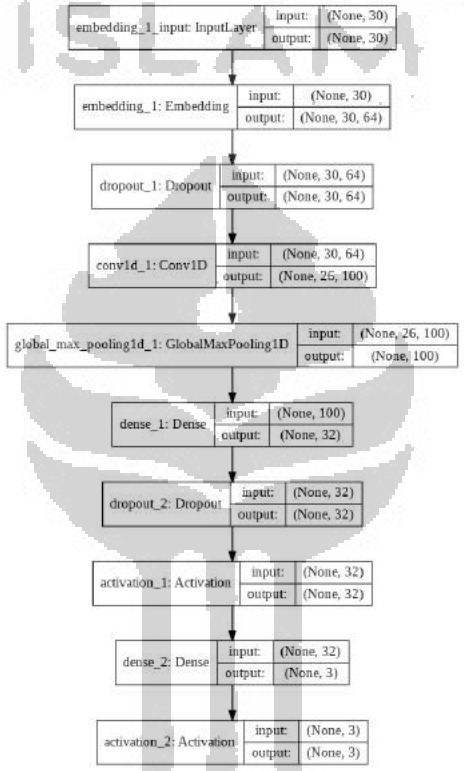
Karena ukuran filter yang digunakan memiliki pengaruh besar terhadap performa CNN, oleh karena itu nilai filter harus disesuaikan. Untuk klasifikasi kalimat, ukuran filter yang sebaiknya digunakan yaitu berada pada rentang 1-10. Selain itu juga mencari jumlah filter dengan rentang 100-600 sekaligus menerapkan drop out antara 0,0-0,5 untuk mendapatkan performa terbaik (Wallace, 2014). Nilai neuron dan unit lstm maupun gru memanfaatkan nilai yang sudah sering digunakan diantaranya yaitu 32, 64, dan 128. Tabel 3.16 merupakan skenario yang telah dibuat :

Tabel 3.16 Skenario Model Sentimen Analisis Berbasis Fitur

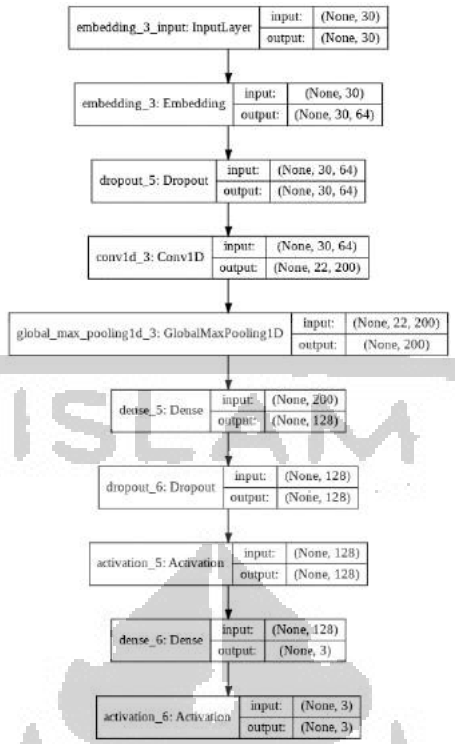
Skenario	Model	Metode	Kernel	Filter	Hidden_Dims	Dropout	Unit
1	M1	CNN	5	100	32	0.25	
	M2		9	200	128	0.25	
	M3		9	200	128	0.5	
2	M4	CNN + LSTM	9	200	128	0.25	64
	M5		9	200	128	0.5	64
3	M6	CNN + GRU	9	200	128	0.25	64
	M7		9	200	128	0.5	64

Arsitektur seringkali digunakan untuk meringkas, memvisualisasikan dan lebih memahami model jaringan saraf yang digunakan. Arsitektur layer aspek Gambar 3.4, Gambar 3.5, dan arsitektur lapisan sentimen Gambar 3.8, dan Gambar 3.9, merupakan susunan lapisan CNN yang digunakan. Setiap arsitektur terdapat urutan lapisan beserta ukuran masukan dan ukuran luaran. Urutan lapisan dimulai dari lapisan masukan, kemudian dilanjutkan dengan lapisan

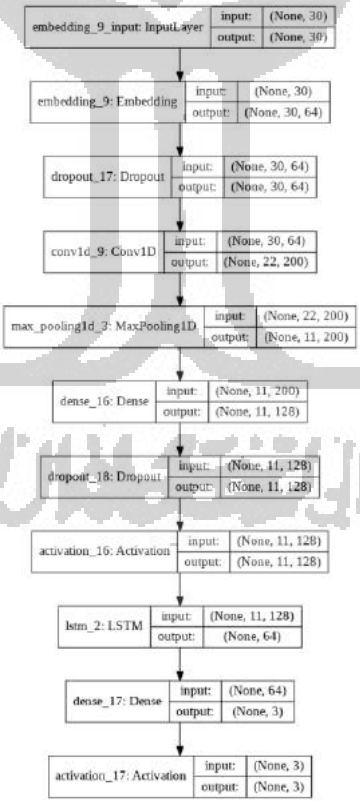
embedding, lapisan *dropout*, lapisan konvolusi 1D, lapisan *global max pooling* 1D, lapisan *dense*, lapisan *dropout*, lapisan aktivasi, lapisan *fully connected* beserta aktivasi nya. Sedangkan arsitektur lapisan aspek Gambar 3.6, Gambar 3.7, dan arsitektur lapisan sentimen Gambar 3.10, dan Gambar 3.11 merupakan susunan lapisan metode gabungan CNN + LSTM dan CNN + GRU. Sama dengan susunan lapisan CNN sebelumnya, hanya saja setelah melalui lapisan CNN diikuti oleh lapisan metode LSTM atau metode GRU sebelum diakhiri dengan lapisan *fully connected*.



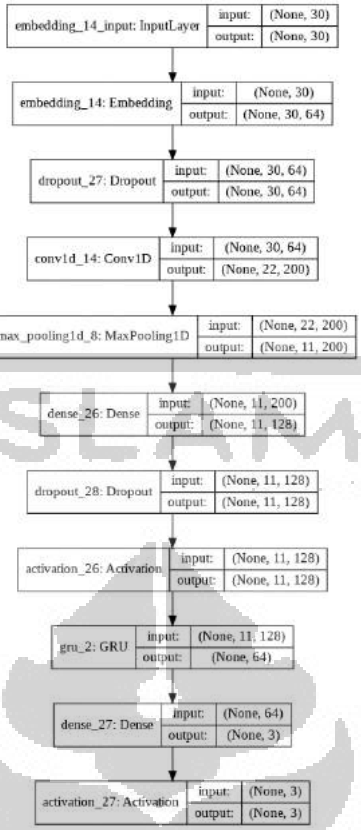
Gambar 3.4 Arsitektur Aspek M1



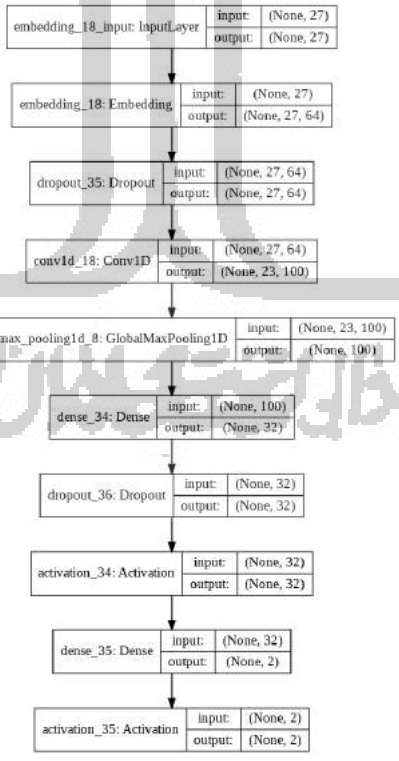
Gambar 3.5 Arsitektur Aspek M2 & M3



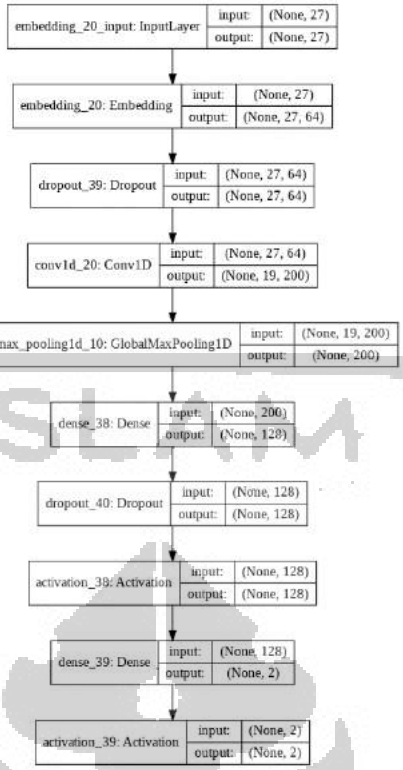
Gambar 3.6 Arsitektur Aspek Skenario 2



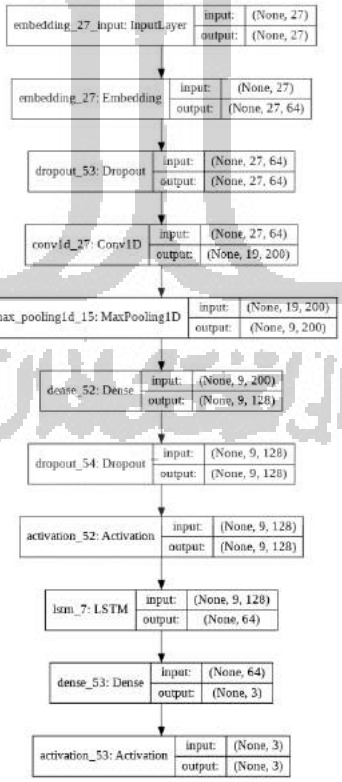
Gambar 3.7 Arsitektur Aspek Skenario 3



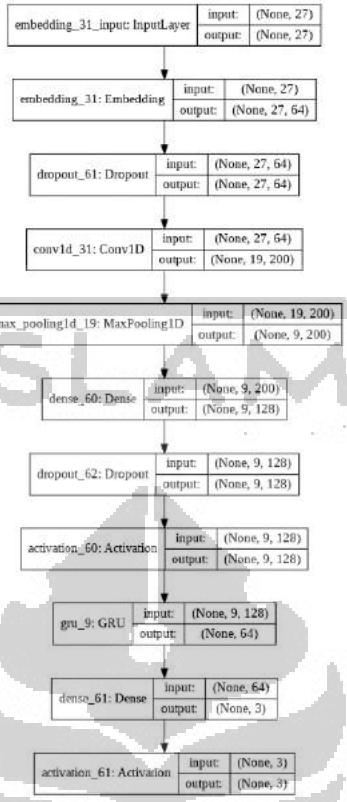
Gambar 3.8 Arsitektur Sentimen M1



Gambar 3.9 Arsitektur Sentimen M2 & M3



Gambar 3.10 Arsitektur Sentimen Skenario 2



Gambar 3.11 Arsitektur Sentimen Skenario 3

3.2.6 Evaluasi

Evaluasi dilakukan dengan menggunakan metode *hold-out* dan confusion matrix. Metode *hold-out* membagi data kedalam 2 bagian yakni data latih dan data uji. Data latih digunakan untuk melatih model. Data validasi digunakan untuk melihat seberapa baik performa model. *Confusion matrix* biasa digunakan untuk mengukur kinerja suatu algoritma klasifikasi. Contoh *confusion matrix* dalam masalah klasifikasi 2 kelas dapat dilihat Tabel 3.17. Dapat disimpulkan bahwa gambar tersebut memiliki 4 hasil klasifikasi berbeda. *True positive* dan *true negative* adalah hasil klasifikasi yang benar, sementara *false positive* dan *false negative* merupakan 2 kemungkinan tipe kesalahan. Penjelasan dari *confusion matrix* dapat dilihat di bawah ini (Kohavi & Provost, 1998):

Tabel 3.17 Confusion Matrix

		Predicted class	
		Negatives	*Positives*
Actual Class	*Negatives*	a	b
	Positives	c	d

Beberapa istilah untuk confusion matrix yaitu diantaranya : akurasi, *true positive, true negative, false positive, false negative,* dan presisi. Akurasi dapat didapatkan dari bagian hasil benar(true) dibandingkan dengan jumlah total data. Perhitungan akurasi dapat dilihat pada persamaan (3.1) :

$$\text{Akurasi} = \frac{a + d}{a + b + c + d} \tag{3.1}$$

Perhitungan *true positive rate (Recall, Sensitivity)* didapatkan dari kelas positif yang teridentifikasi benar. Persamaan (3.2) merupakan cara menghitung *true positive* :

$$Recall = \frac{d}{c + d} \tag{3.2}$$

Presisi atau nilai prediktif positif didapat dari kasus positif prediktif yang akurat, dan dihitung menggunakan persamaan (3.3):

$$\text{Presisi} = \frac{d}{b + d} \tag{3.3}$$