Nama : Jiddy Abdillah
NIM : 1301162765
Kelas : SIDE-40-GAB Penambangan Teks

Analisis Performa BiLSTM dan CNN menggunakan IMDB Review Dataset

A. Strategi Pendekatan Masalah

Pada penyelesaian masalah klasifikasi teks, ada beberapa komponen penting, yaitu:

1. Training Text and Labels

Pada masalah ini saya menggunakan dataset IMDB berisi review sebuah film dan label sentiment dari review tersebut. Jumlah data pada dataset ini yaitu sebanyak 25000 data.

	id	sentiment	review
0	5814_8	1	With all this stuff going down at the moment w...
1	2381_9	1	\The Classic War of the Worlds\" by Timothy Hi...
2	7759_3	0	The film starts with a manager (Nicholas Bell)...

2. Feature Vector

Feature Vector yang digunakan pada masalah ini adalah GloVe 6B 100d, berisi 6 miliar kata dan memiliki 100 dimensi yang nantinya digunakan pada layer embedding dari model prediktif yang dibuat. Fungsi *feature vector* yaitu mengambil vektor representasi dari sebuah kata.

3. Learning Algorithm

Learning algorithm yang digunakan pada masalah ini adalah simplified *Convolutional Neural Network* (CNN) dan *Bidirectional Long Short Term Memory* (BiLSTM).

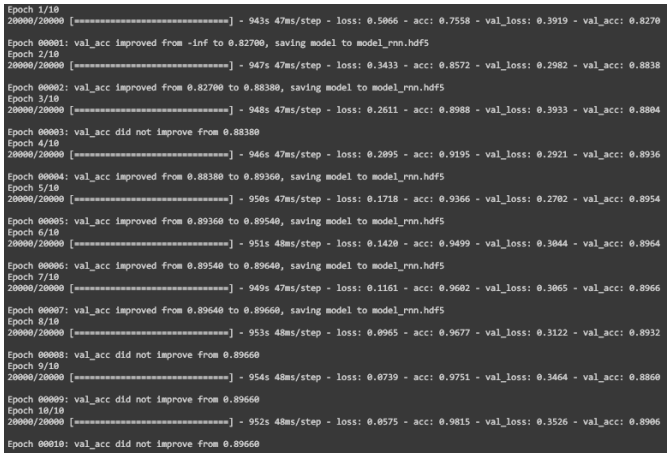
CNN adalah kelas jaringan saraf tiruan dimana koneksi antara *nodes* tidak membentuk suatu siklus, sedangkan BiLSTM membentuk koneksi graf arah bersama sekuens antara *nodes* yang ada. BiLTSM dilatih untuk mengenal pattern berdasarkan waktu sedang CNN dilatih untuk mengenal pattern berdasarkan ruang.

4. Predictive Model

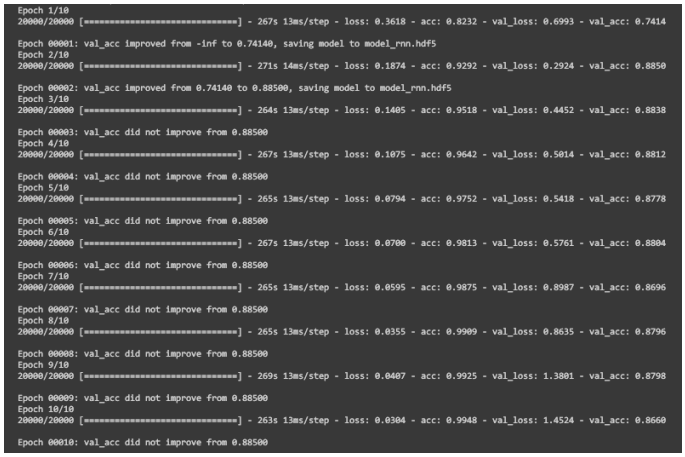
Model yang dibuat terdiri dari empat layer utama, yaitu *input layer, embedding layer, dropout layer,* dan *dense layer*. Embedding layer menggunakan word2vec yang sudah ada yaitu GloVe. Nilai *dropout* yang digunakan pada model yaitu 0,5.

Model CNN yang digunakan yaitu model satu dimensi. Tambahan layer pada model CNN yaitu menggunakan layer *flatten* dan tiga layer *max pooling*.

B. Output Program



Gambar 1. BiLSTM

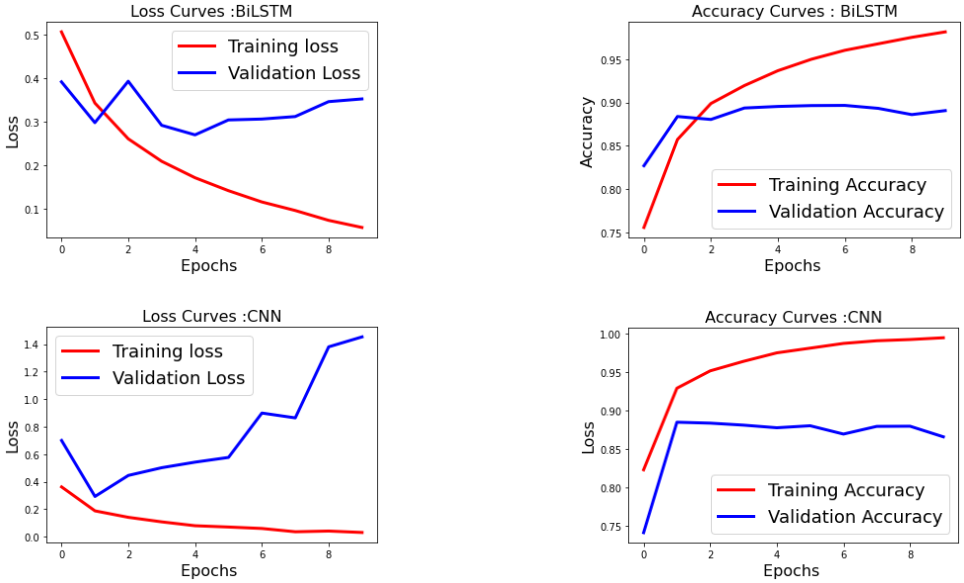


Gambar 2. CNN

Pada BiLSTM, hasil setelah 10 iterasi yaitu memiliki nilai *training accuracy* terbesar dan *loss* terkecil yaitu 98% dan 0,0575. *Validation accuracy* terbersar dan *loss* terkecil yaitu 89% dan 0,2702.

Pada CNN, hasil setelah 10 iterasi yaitu memiliki nilai *training accuracy* terbesar dan *loss* terkecil yaitu 99% dan 0,0304. *Validation accuracy* terbersar dan *loss* terkecil yaitu 88% dan 1,4524.

C. Analisis



Dari hasil grafik diatas, perbedaan terbesar terjadi kepada *validation loss* model CNN. Semakin banyak iterasi, nilai *validation loss* semakin besar. Nilai *validation loss* lebih besar dari *accuracy loss* terjadi karena jumlah data yang lebih sedikit. Untuk mengatasi nilai *loss* yang terlalu besar, dapat dilakukan pengaturan *dropout rate*, mengurangi kompleksitas dari model dan menentukan *learning rate* dan *decay rate*.