Klasifikasi Sentimen Twitter dengan Neural Network

Kelompok 1 - DSC Wave 2
DSC220300071 - Algaza Geofarry Susanto
DSC220900031 - Keitaro Mirakel Wongso
Binar Academy

Pendahuluan

Pesatnya kemajuan teknologi telah memunculkan banyak website dan aplikasi yang dijadikan sarana khalayak ramai untuk membicarakan suatu topik dan mengekspresikan suatu opini. Informasi yang didapatkan dari website dan aplikasi tersebut sangat berguna untuk digali karena dapat menunjukan suatu sentimen tertentu terhadap isu tertentu. Salah satu manfaat dari sentimen dari para pengguna adalah untuk mendapatkan umpan balik atas suatu hal, sehingga dapat membantu pengambilan keputusan di kemudian hari.

Twitter merupakan salah satu media sosial populer di Indonesia. Berdasarkan We Are Social, sebuah agensi media sosial yang berpusat di inggris, pengguna twitter di Indonesia mencapai 18,45 juta pada tahun 2022. Sebuah pesan pada twitter atau lebih dikenal sebagai 'tweet' memiliki berbagai macam variasi, salah satunya adalah dalam bentuk teks. Sebuah tweet dapat diklasifikasikan berdasarkan sentimennya untuk dapat menganalisis sikap dan perilaku dari pengguna terhadap suatu isu tertentu.

Penelitian ini bertujuan untuk membentuk sebuah model *artificial neural network* dan *Long Short Term Memory* (LSTM) Neural Network yang dapat melakukan klasifikasi teks tweet berdasarkan tiga sentimen, yaitu negatif, netral, dan positif.

Metode Penelitian

Analisis Deskriptif

Analisis deskriptif dilakukan untuk mengetahui deskripsi dari data. Pola dan karakteristik penyebaran kata dan huruf juga dapat ditemukan dengan analisis deskriptif.

Cleansing :

Cleansing dilakukan dengan menggunakan **regrex** untuk menghilangkan simbol, url, dan raw string yang tidak memiliki makna. **Normalisasi kata yang tidak baku** juga dilakukan untuk meningkatkan performa dari proses training model. **Stopwords** yang merupakan kata yang tidak memiliki arti khusus dan kurang signifikan dalam suatu kalimat juga dihilangkan sebelum proses training dilakukan.

Tokenisasi :

Tokenisasi dilakukan untuk melakukan vektorisasi pada teks yang sudah terkumpul dalam corpus. Kata-kata pada korpus akan dipecah menjadi beberapa bagian yang masing-masing memiliki nilai atau token tersendiri. Pada penelitian ini, tokenisasi dilakukan pada python dengan dua cara, yaitu Bag of Words (CountVectorizer) dan Tokenizer

Neural Network :

Neural network yang digunakan pada penelitian ini adalah Artificial Neural Network dan Long Short Term Memory Neural Network

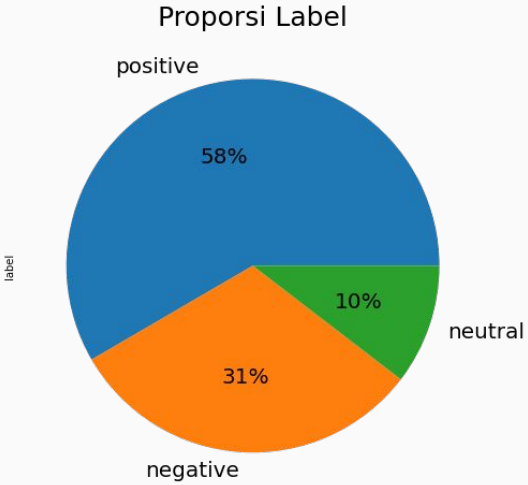
Dataset Training

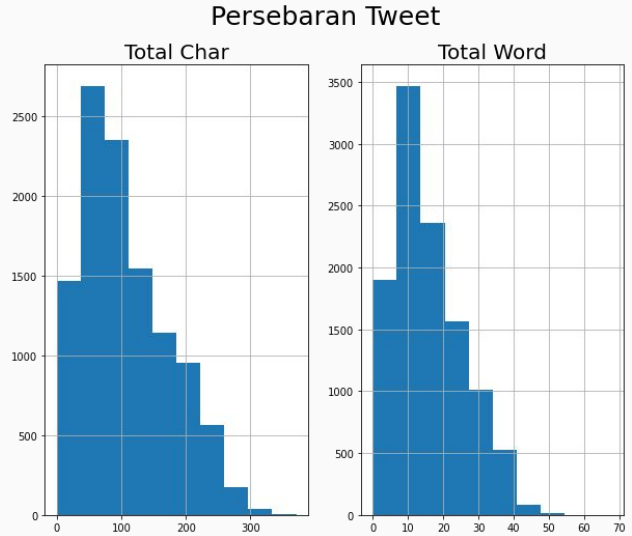
Dataset ini digunakan untuk proses training dari model NN dan LSTM. Dataset ini terdiri dari teks beserta labelnya (negative, neutral, dan positive). Model akan mempelajari cara mengklasifikasi teks tweet berdasarkan tweet yang sudah di pre-processing

index	text	label
0	warung ini dimiliki oleh pengusaha pabrik tahu yang sudah puluhan tahun terkenal membuat tahu putih di bandung . tahu berkualitas , dipadu keahlian memasak , dipadu kretivitas , jadilah warung yang menyajikan menu utama berbahan tahu , ditambah menu umum lain seperti ayam . semuanya selera indonesia . harga cukup terjangkau . jangan lewatkan tahu bletoka nya , tidak kalah dengan yang asli dari tegal !	positive
1	mohon ulama lurus dan k212 mmbri hujjah partai apa yang harus diwlh agar suara islam tidak pecah-pecah	neutral
2	lokasi strategis di jalan sumatera bandung . tempat nya nyaman terutama sofa di lantai 2 . paella nya enak , sangat pas dimakan dengan minum bir dingin . appetiser nya juga enak-enak .	positive
3	betapa bahagia nya diri ini saat unboxing paket dan barang nya bagus ! menetapkan beli lagi !	positive
4	duh . jadi mahasiswa jangan sombong dong . kasih kartu kuning segala . belajar dulu yang baik , tidak usahlah ikut-ikut politik . nanti sudah selesai kuliah nya mau ikut politik juga tidak telat . dasar mahasiswa .	negative

Dataset Training

Proporsi dari label dataset ini didominasi oleh teks dengan sentimen positif, sementara persebaran kata dan karakter dapat dilihat seperti di histogram di bawah





Dataset Cleansing - Stopwords

Dataset ini digunakan untuk mendapatkan kumpulan kata stopwords yang akan di hapus dari data yang kita gunakan. Hal ini dilakukan untuk meminimalisir informasi yang kurang penting agar nantinya program hanya akan terfokus pada kata-kata yang lebih penting.

index	0
0	adalah
1	adapun
2	agaknya
3	akan
4	akhir
5	akhirnya

Dataset Cleansing - Slang Words

Dataset ini digunakan untuk mendapatkan kumpulan kata slang/tidak baku yang akan di hapus dari data yang kita gunakan. Hal ini dilakukan untuk meminimalisir informasi yang kurang penting agar nantinya program hanya akan terfokus pada kata-kata yang lebih penting.

index	slang	normal
0	@	di
1	abis	habis
2	ad	ada
3	adlh	adalah
4	afaik	as far as i know
5	ahaha	haha

Dataset Validation/Testing

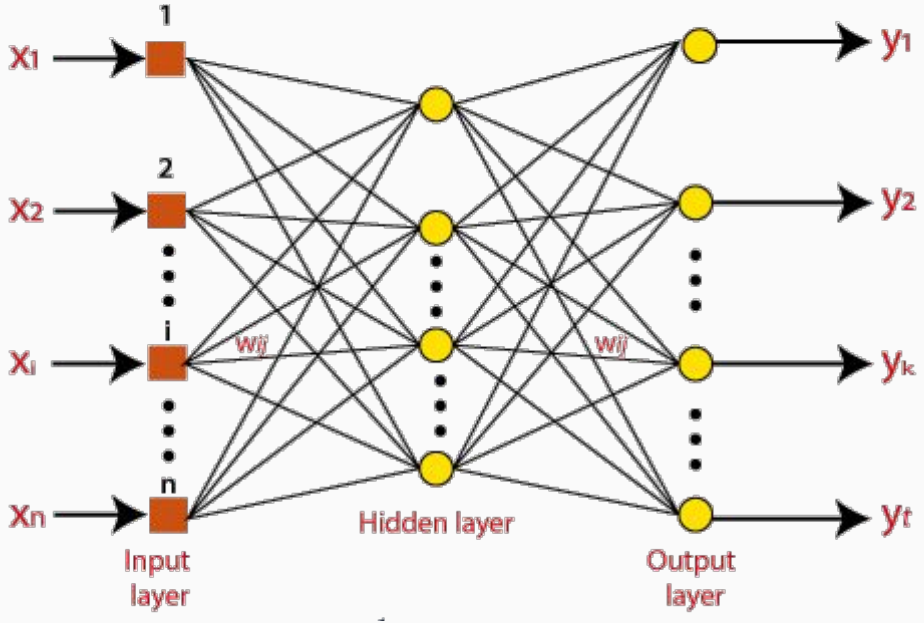
Dataset ini digunakan untuk memvalidasi model NN dan LSTM yang telah kita buat. Dimana komputer akan mempelajari tiap teks dan memberitahu sentiment dari teks tersebut

index	text
0	- disaat semua cowok berusaha melacak perhatian gue. loe lantas remehkan perhatian yg gue kasih khusus ke elo. basic elo cowok bego ! ! !'
1	RT USER: USER siapa yang telat ngasih tau elu?edan sarap gue bergaul dengan cigax jifla calis sama siapa noh licew juga'
2	41. Kadang aku berfikir, kenapa aku tetap percaya pada Tuhan padahal aku selalu jatuh berkali-kali. Kadang aku merasa Tuhan itu ninggalkan aku sendirian. Ketika orangtuaku berencana berpisah, ketika kakakku lebih memilih jadi Kristen. Ketika aku anak ter
3	USER USER AKU ITU AKU\n\nKU TAU MATAMU SIPIT TAPI DILIAT DARI MANA ITU AKU'
4	USER USER Kaum cebong kapir udah keliatan dongoknya dari awal tambah dongok lagi hahahah'

Artificial Neural Network (ANN)

Artificial Neural Network (ANN) yang digunakan merupakan neural network dengan input layer, satu hidden layer berdimensi (100 x 1) dan output layer yang menyesuaikan jumlah sentimen (negative, neutral, positive).

Fungsi aktivasi yang dipilih adalah Rectified Linear Unit (relu) dan metode optimasi yang dipilih adalah ‘adam’.



Sumber : https://www.javatpoint.com/artificial-neural-network

Python

```
model_neural = MLPClassifier(hidden_layer_sizes=(100,),max_iter=1000,activation='relu',solver='adam',learning_rate_init=0.001,early_stopping=True,n_iter_no_change=20, verbose = 1)
```

Artificial Neural Networks (Testing Metrics)

Testing selesai

	precision	recall	f1-score	support
negative	0.78	0.82	0.80	684
neutral	0.85	0.65	0.74	220
positive	0.91	0.92	0.91	1283
accuracy			0.86	2187
macro avg	0.85	0.80	0.82	2187
weighted avg	0.86	0.86	0.86	2187

Jika dilakukan K-fold Cross Validation dengan membagi data menjadi 5 bagian, maka didapatkan rata-rata akurasi senilai = **0.8496282825696717**

Long short-term memory

Merupakan pengembangan dari metode RNN, dimana metode ini mengatasi kelemahan RNN yang tidak dapat memprediksi kata yang disimpan dalam memori jangka panjang.

Model yang kami buat menggunakan satu embedding layer dengan dimensi 256, satu layer LSTM dengan Unit/neuron 32, dan output layer berdimensi tiga dengan fungsi aktivasi softmax.

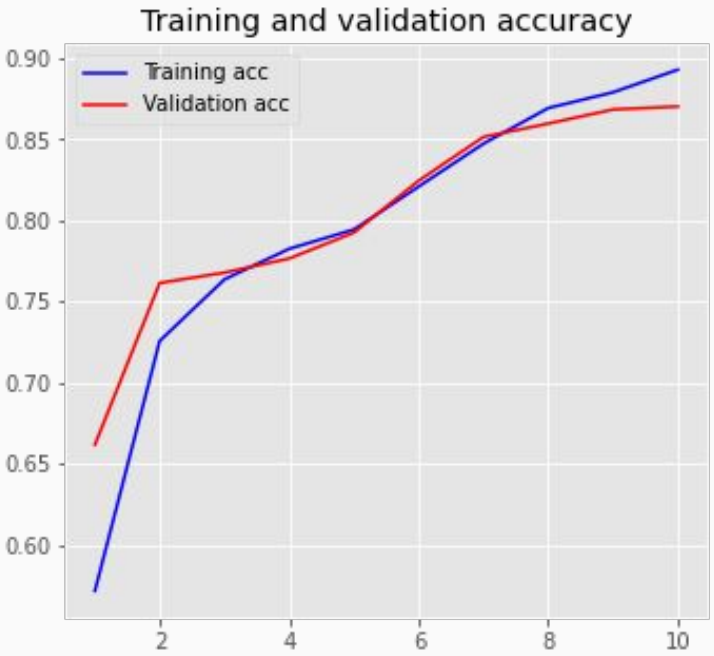
Python

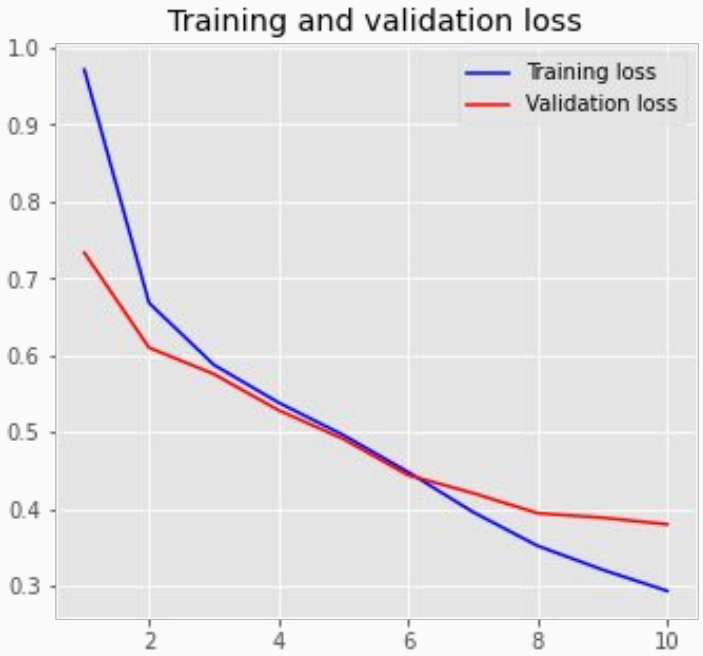
```
embed_dim = 256

model = Sequential()
model.add(Embedding(max_features, embed_dim, input_length=
X.shape[1]))
model.add(SpatialDropout1D(0.76))
model.add(LSTM(32, dropout=0.77, recurrent_dropout=0.76))
model.add(Dense(3, activation='softmax'))

model.compile(optimizer='adam',
loss='categorical_crossentropy', metrics=['accuracy'])
history = model.fit(X_train, y_train, epochs=10,
batch_size=256, validation_data=(X_test, y_test), verbose=1,
callbacks=None)
```

LSTM Learning Rate





LSTM (Testing Metrics)

Testing selesai

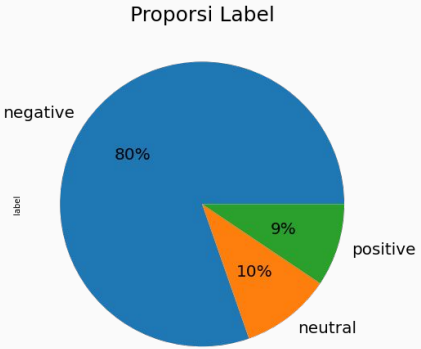
	precision	recall	f1-score	support
0	0.79	0.81	0.80	668
1	0.82	0.70	0.76	231
2	0.91	0.92	0.91	1288
accuracy			0.86	2187
macro avg	0.84	0.81	0.82	2187
weighted avg	0.86	0.86	0.86	2187

Jika dilakukan K-fold Cross Validation dengan membagi data menjadi 5 bagian, maka didapatkan rata-rata akurasi senilai = **0.8682213077274806**

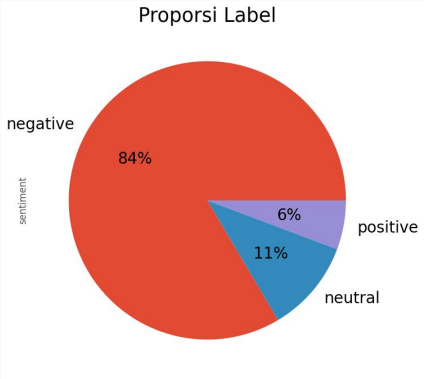
Hasil dan Kesimpulan

Rata-rata metrics akurasi kedua model, yaitu ANN dan LSTM, masing-masing menunjukan nilai 84,9% dan 86,8%. Nilai tersebut sudah dapat terbilang cukup bagus, sehingga model yang telah dibuat dapat diaplikasikan pada dataset testing yang menghasilkan klasifikasi sentimen dengan proporsi sebagai berikut :

ANN



LSTM



Proporsi sentimen menunjukan mayoritas data validasi memiliki sentimen negatif. Hal tersebut wajar dikarenakan data validasi yang digunakan merupakan data tweet yang sudah dioptimasi untuk pendeteksian hate speech dan abusive tweet.