IMPLEMENTASI ALGORITMA PEMBELAJARAN BACK PROPAGATION UNTUK KLASIFIKASI TWEET RESOLUSI

LINGGAR JUWITA HANDAYANI
NRP 5112 100 031

Dosen Pembimbing
Diana Purwitasari, S.Kom., M.Sc.
Dini Adni Navastara, S.Kom., M.Sc.

No	Tweet
1	*#NewYearsResolution Stop doubting myself*
2	*#NewYearsResolution to eat buy and eat environmentally responsible. buy for local businesses and eat a plant based diet. #eatclean2015*
3	*Adventures. I'm ready for more of those. #NewYearsResolution*
4	*#NewYearsResolution will be moving to @ExploreGeorgia #Altanta going to school and becoming a hair dresser or a actor!*
5	*I'm going to stop being honest and start telling people whatever they want to hear. So much easier. #NewYearsResolution*
6	*@jeffreymarshnow My New Years Resolution: Be brave and speak up about issues that matter! #NewYearsResolution*
7	*"New Years Resolution* *Save Money don't Spend Money"*
8	*New Years resolution: more brunch.*

LATAR BELAKANG

- Pengguna social media, dalam kasus ini Twitter, sering berbagi pengalaman keseharian.
- Jumlah *tweet* meningkat pesat menjelang tahun baru. Dan t*weet* resolusi mendominasi pada awal tahun.
- Awal tahun merupakan waktu yang bagus untuk menyusun strategi bisnis, baik penyedia barang dan jasa.
- Dari *tweet-tweet* resolusi, dilakukan klasifikasi untuk mengetahui keinginan pengguna, sebagai konsumen.

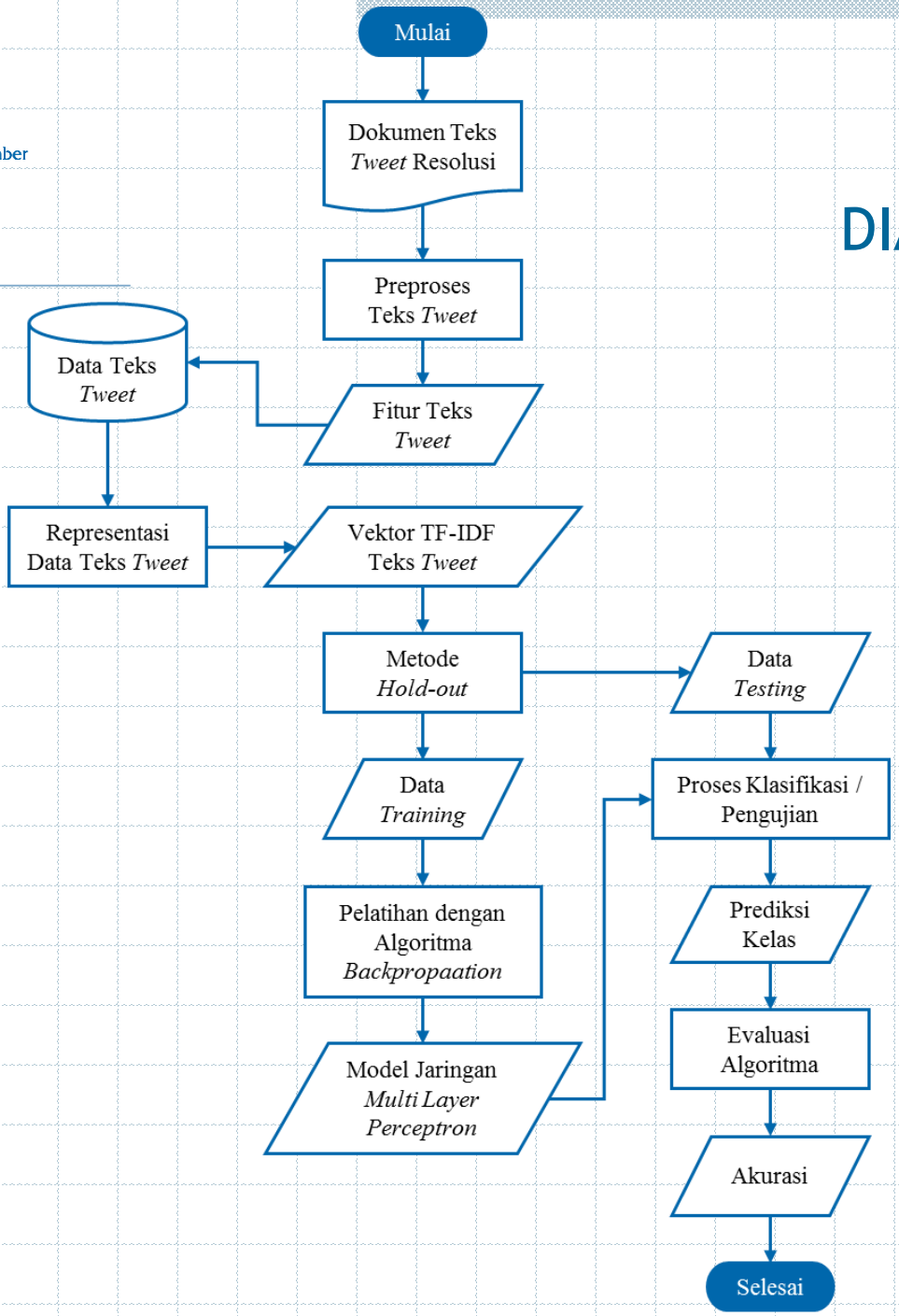
RUMUSAN MASALAH

- Bagaimana melakukan klasifikasi teks tweet ke dalam kategori yang telah ditentukan?
- Bagaimana mengekstraksi fitur teks berdasarkan kemunculan kata dalam dokumen?
- Bagaimana melakukan prediksi kategori teks tweet dengan model jaringan yang telah ada?

BATASAN MASALAH

- Sistem dibangun berbasis desktop, dengan basis data dan bahasa pemrograman C# serta kerangka kerja .NET.
- Data tweet resolusi yang digunakan adalah dataset tweet resolusi Tahun Baru 2015.
- Data tweet resolusi diambil berdasarkan penggunaan #newyearsresolution pada teks tweet.
- Data tweet resolusi yang akan diproses adalah teks Bahasa Inggris.
- Data masukan tweet resolusi telah dikompilasi dalam bentuk dokumen excel spreadsheet, dengan tipe berkas Microsoft Excel (ekstensi .xls atau .xlxs).
- Model Jaringan Saraf Tiruan yang digunakan adalah jenis Multi Layer Perceptron dengan satu hidden layer.
- Kelas tweet terdiri dari enam kelas, di antaranya, Career & Education, Finance, Health & Fitness, Personal Growth, Recreation & Leisure, dan Relationship.

DIAGRAM ALIR SISTEM



TAHAP PRAPROSES TEKS

No	Input	Case Folding	Tokenizing		Filtering
1	*Self improvement!.. Mentally, physically, and financially. #NewYearsResolution*	*self improvement!.. mentally, physically, and financially. #newyearsresolution*	*self*		*self*
			improvement		*improvement*
			mentally		*mentally*
			physically		*physically*
			and		*financially*
			financially		
			newyearsresolution		
2	*social media purging > #NewYearsResolution*	*social media purging > #newyearsresolution*	*social*		*social*
			media		*media*
			purging		*purging*
			gt		
			newyearsresolution		
3	*to actually work out and not be lazy about it #NewYearsResolution*	*to actually work out and not be lazy about it #newyearsresolution*	*to*	*not*	*work*
			actually	*be*	*lazy*
			work	*lazy*	
			out	*about*	
			and	*it*	
			newyearsresolution		

PEMBOBOTAN FITUR TEKS

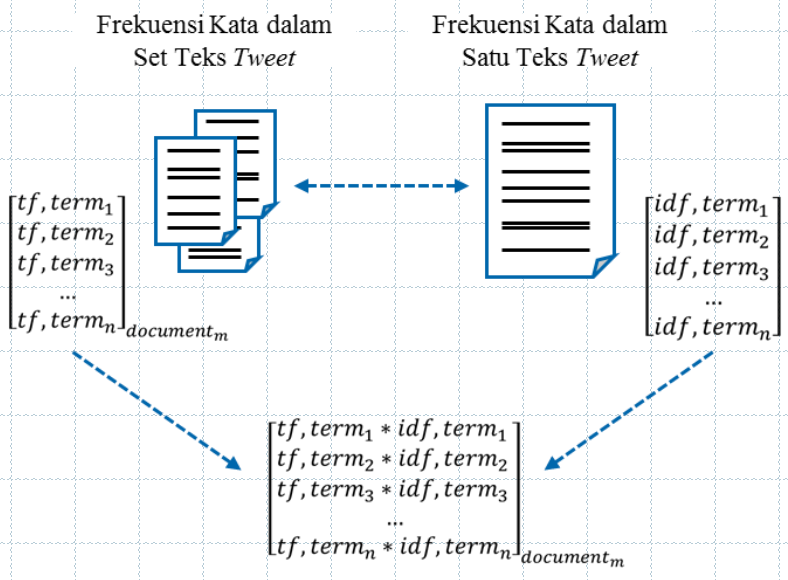
$$tf_i(d_j) = \frac{freq_i(d_j)}{\sum_{i=1}^{k} freq_i(d_j)} \quad (2.1)$$

$$idf_i = \log \frac{|D|}{|\{d: t_i \in d\}|} \quad (2.2)$$

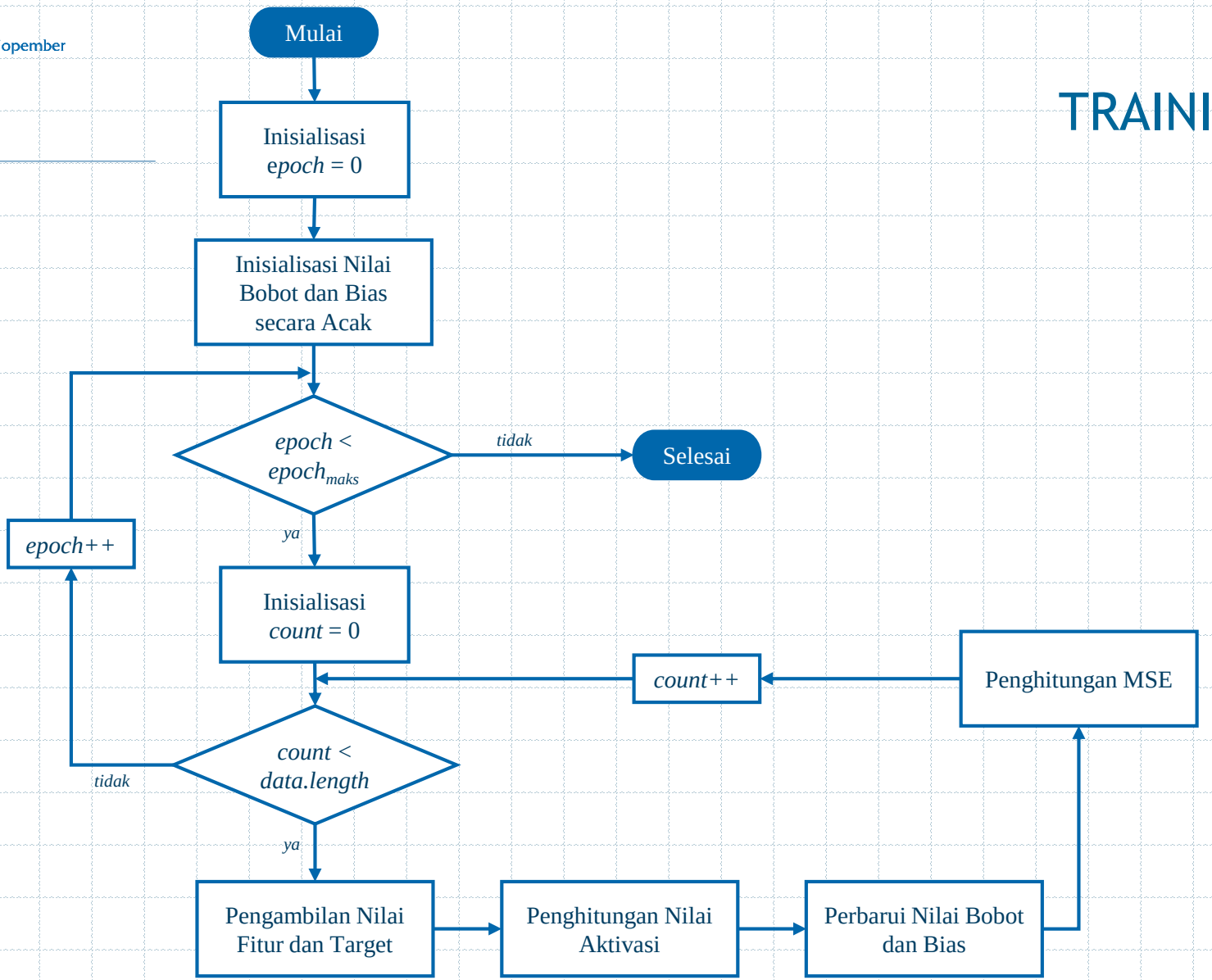
$$(tf - idf)_{ij} = tf_i(d_j) \,.\, idf_i \quad (2.3)$$

Di mana,

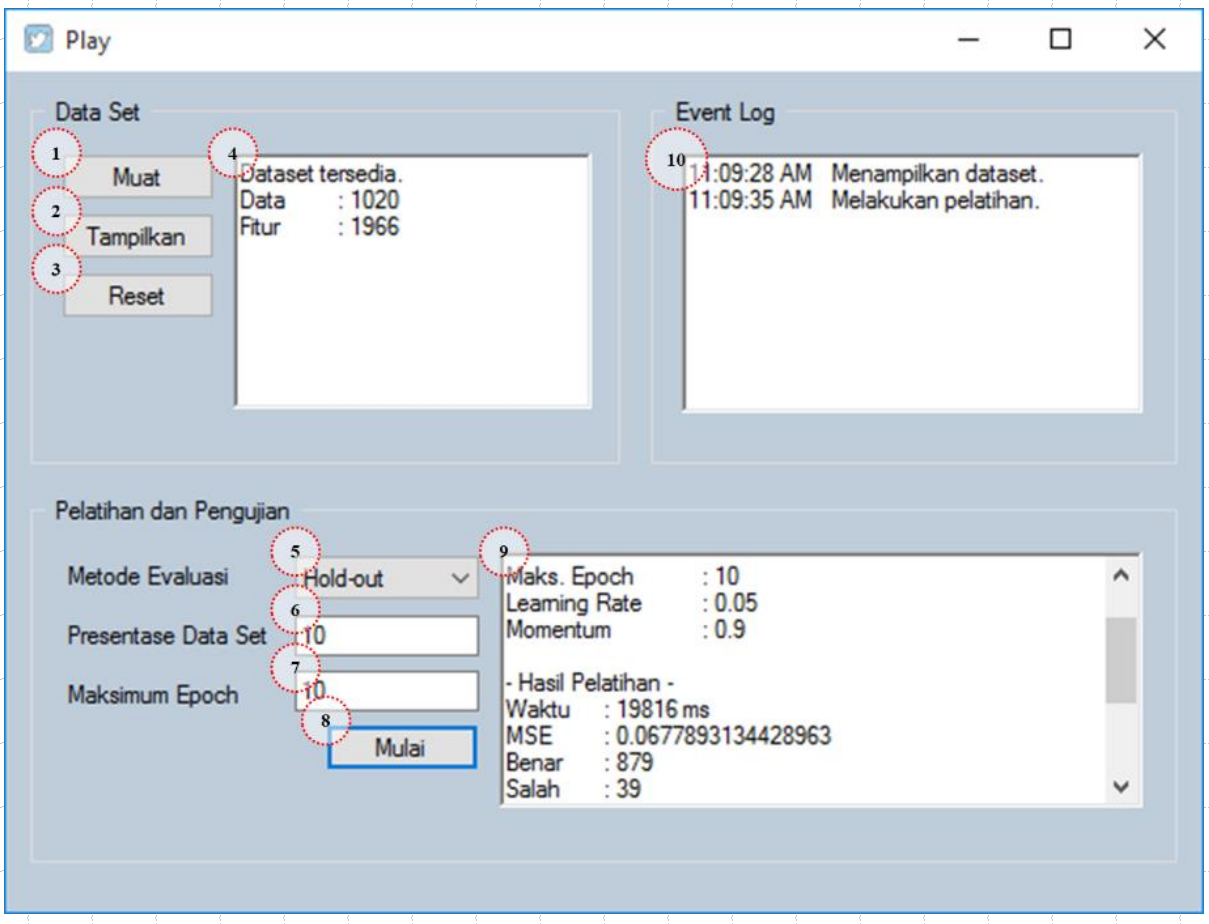
- $freq_i(d_j)$ adalah frekuensi *term* ke- i dalam dokumen ke- j.
- $\sum_{i=1}^{k} freq_i(d_j)$ adalah jumlah *term* pada dokumen ke- j.
- $|D|$ adalah jumlah dokumen dalam *corpus*.
- $|\{d: t_i \in d\}|$ adalah dokumen yang mengandung *term* ke- i.



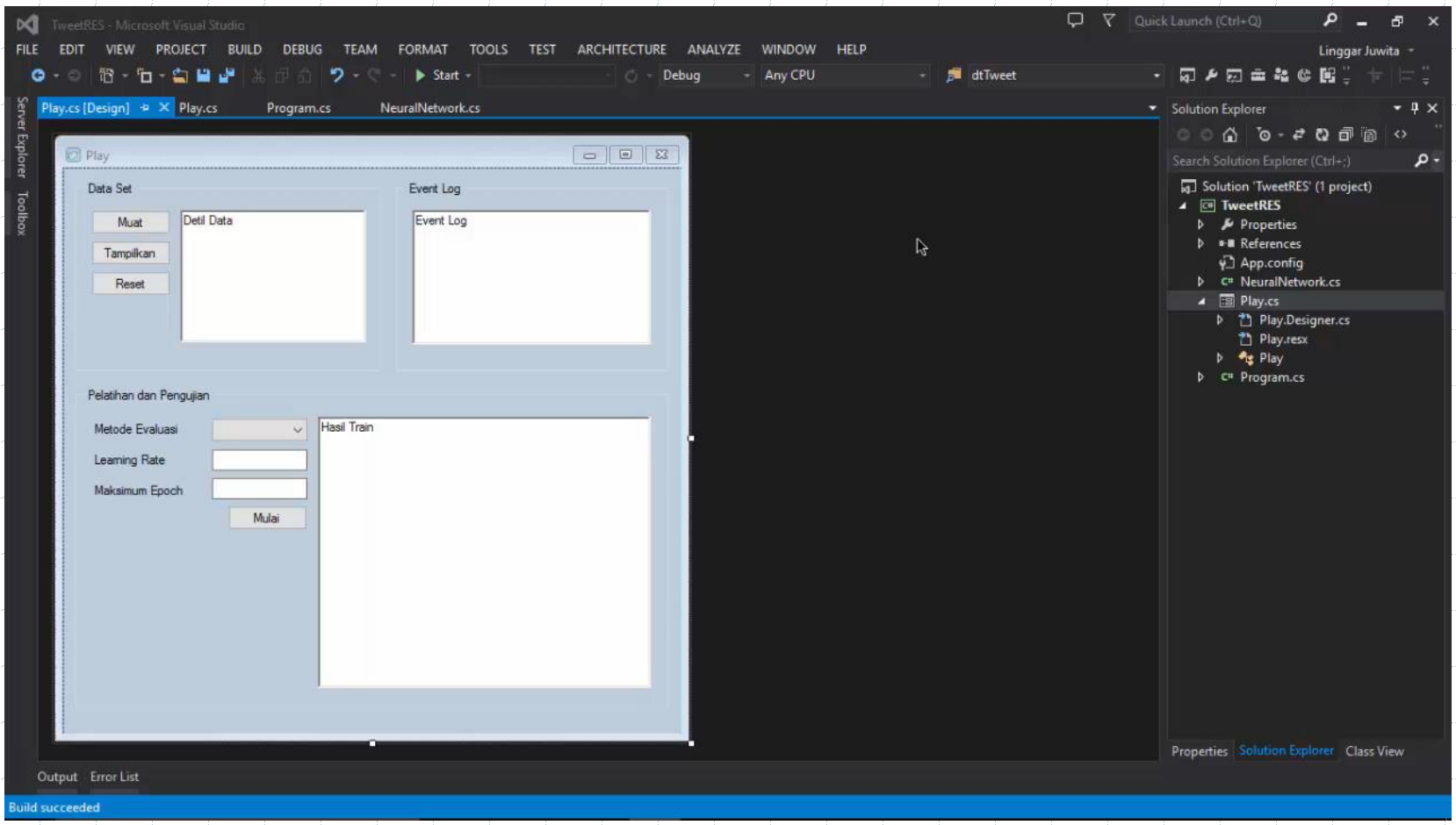
TRAINING


Mulai
Inisialisasi
epoch = 0
Inisialisasi Nilai
Bobot dan Bias
secara Acak
epoch <
epochmaks
tidak
Selesai
ya
epoch++
Inisialisasi
count = 0
count++
Penghitungan MSE
count <
data.length
tidak
ya
Pengambilan Nilai
Fitur dan Target
Penghitungan Nilai
Aktivasi
Perbarui Nilai Bobot
dan Bias

TAMPILAN ANTARMUKA PENGGUNA



ANTARMUKA PENGGUNA



SKENARIO UJI COBA

- Hasil akurasi berdasarkan variasi nilai learning rate.

Tujuan skenario ini adalah untuk mengetahui pengaruh nilai learning rate terhadap nilai akurasi.

- Hasil akurasi berdasarkan variasi jumlah neuron pada hidden layer.

Tujuan skenario ini adalah untuk mengetahui pengaruh jumlah neuron pada hidden layer terhadap nilai akurasi.

- Hasil akurasi berdasarkan variasi maksimum nilai epoch.

Tujuan skenario ini adalah untuk mengetahui pengaruh nilai epoch terhadap nilai akurasi.

- Hasil akurasi berdasarkan variasi persentase data pada metode hold-out.

Tujuan skenario ini adalah untuk mengetahui pengaruh perbandingan data training dan data testing terhadap nilai akurasi.

- Hasil akurasi menggunakan kombinasi neuron, epoch, dan learning rate dari skenario uji coba ke-1, ke-2, dan ke-3.

Tujuan skenario ini adalah untuk mengetahui seberapa besar perubahan akurasi terhadap perubahan nilai learning rate menjadi 0,10 dan 0,05; serta pembagian data set dengan persentase 10% dan 20%.

DATA UJI COBA

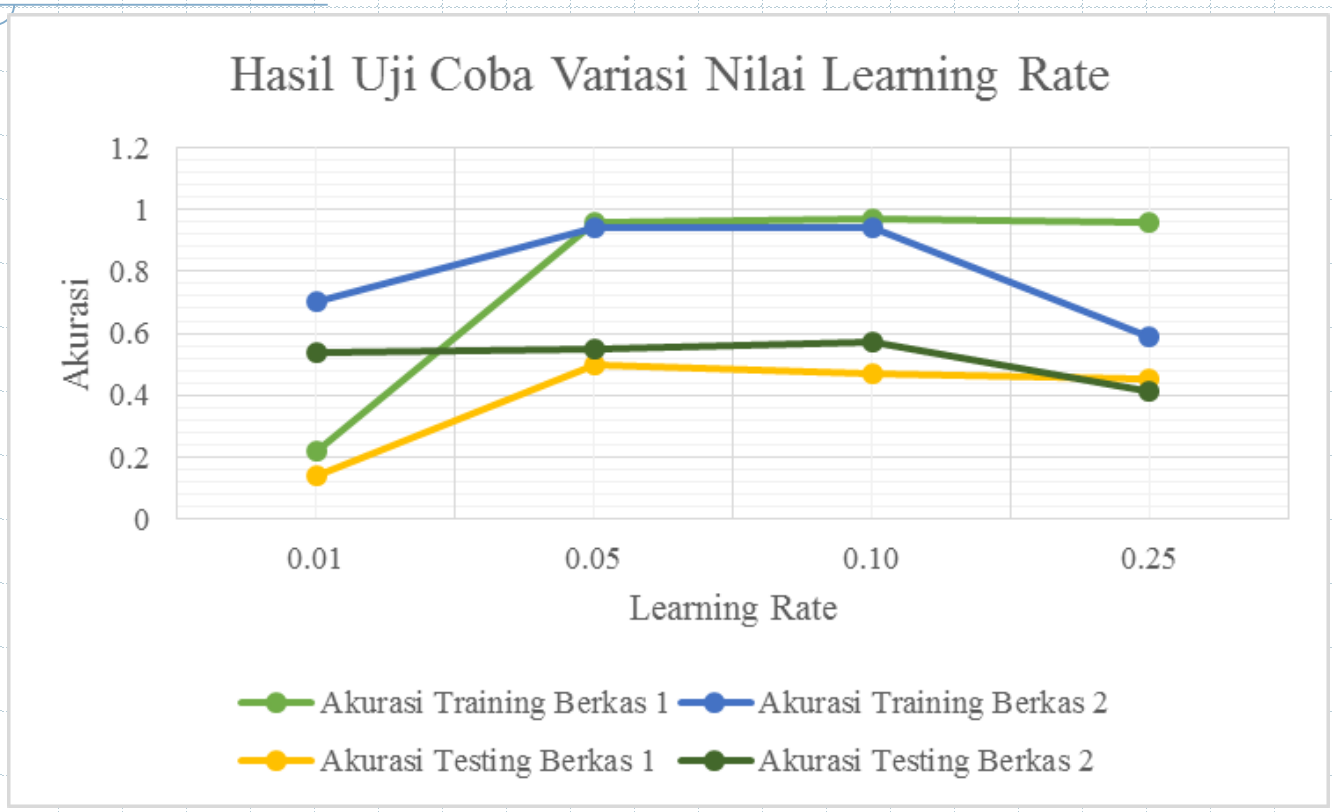
No	Tweet	Topik	Kelas
1	*#NewYearsResolution Stop doubting myself*	*Be more positive*	*Personal Growth*
2	*#NewYearsResolution to eat buy and eat environmentally responsible. buy for local businesses and eat a plant based diet. #eatclean2015*	*Eat healthier*	*Health & Fitness*
3	*Adventures. I'm ready for more of those. #NewYearsResolution*	*Take a trip*	*Recreation & Leisure*
4	*#NewYearsResolution will be moving to @ExploreGeorgia #Altanta going to school and becoming a hair dresser or a actor!*	*Get dream job*	*Career & Education*
5	*I'm going to stop being honest and start telling people whatever they want to hear. So much easier. #NewYearsResolution*	*Be better at keeping in touch with loved ones or friends*	*Relationship*
6	*@jeffreymarshnow My New Years Resolution: Be brave and speak up about issues that matter! #NewYearsResolution*	*Be more confident*	*Personal Growth*
7	*"New Years Resolution Save Money don't Spend Money"*	*Save money*	*Finance*
8	*New Years resolution: more brunch.*	*Eat more*	*Health & Fitness*

Uji Coba dan Analisa Nilai *Learning Rate*

No	Data	Fitur	Learning Rate	MSE	Akurasi (%)		Lama Train (s)
					Train	Test	
1	1056	1990	0,01	0,82	0,22	0,14	78
2			0,05	0,06	0,96	0,50	80
3			**0,10**	**0,02**	**0,97**	**0,47**	**77**
4			0,25	0,03	0,95	0,45	97
5	3914	4551	0,01	0,35	0,70	0,54	3328
6			0,05	0,07	0,94	0,55	3497
7			**0,10**	**0,06**	**0,94**	**0,57**	**2959**
8			0,25	0,62	0,59	0,41	3073

- Nilai MSE paling rendah dimiliki oleh *learning rate* 0,05.
- Perubahan *learning rate* dari 0,01 hingga 0,10 menunjukkan peningkatan akurasi. Namun pada *learning rate* 0,25, akurasi menurun.
- Sistem klasifikasi memiliki nilai akurasi pada nilai *learning rate* 0,10.

Uji Coba dan Analisa Nilai *Learning Rate*

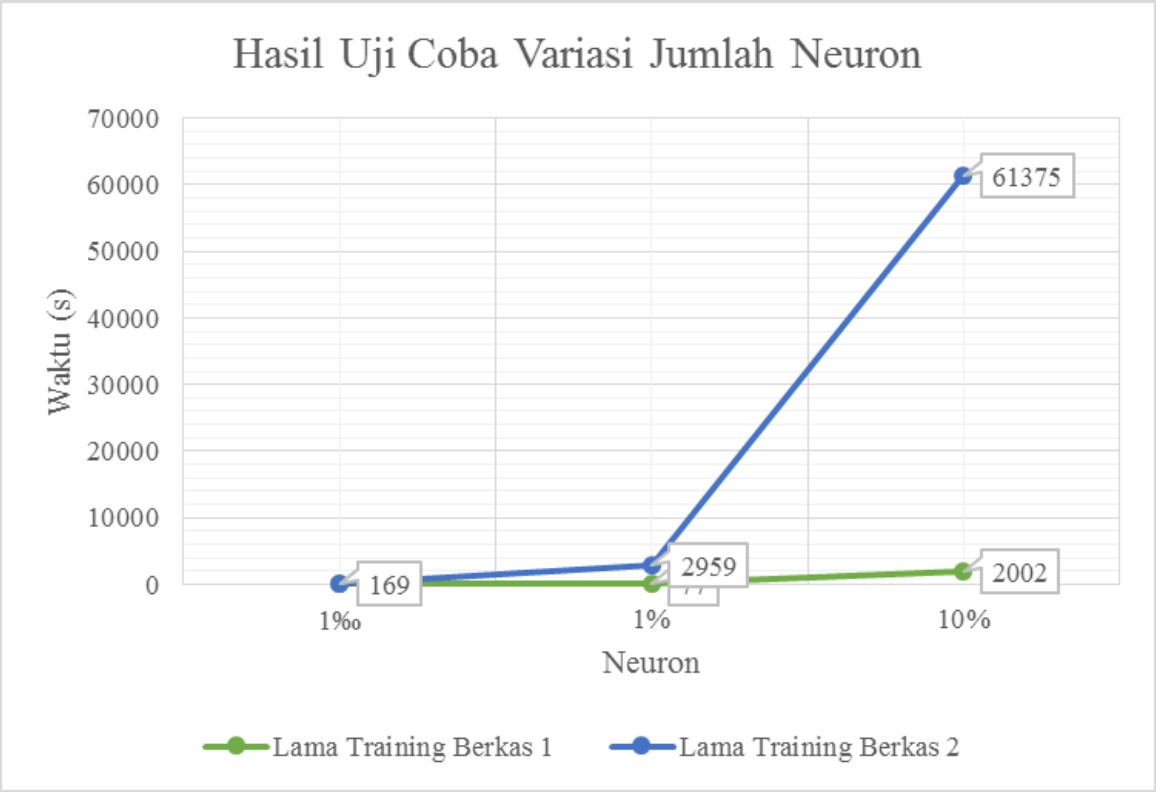


Uji Coba dan Analisa Jumlah *Neuron*

No	Data	Fitur	Neuron	MSE	Akurasi (%)		Lama Train (s)
					Train	Test	
1	1056	1990	1‰	0,58	0,54	0,34	4
2			**1%**	**0,02**	**0,97**	**0,47**	**77**
3			10%	0,01	0,95	0,50	2002
4	3914	4551	1‰	0,08	0,93	0,57	169
5			**1%**	**0,06**	**0,94**	**0,57**	**2959**
6			10%	0,06	0,94	0,57	61375

- *Hidden layer* dengan *neuron* sejumlah 1‰ dari *neuron* pada *input layer* belum cukup untuk menghasilkan akurasi yang tinggi.
- Mulai uji coba dengan *neuron* 1%, nilai akurasi mulai stabil dan cukup tinggi.

Uji Coba dan Analisa Jumlah *Neuron*



- Lama waktu *training* bertambah seiring jumlah *neuron* pada *hidden layer*. 1% dianggap terbaik dengan mempertimbangkan akurasi.

Uji Coba dan Analisa Maksimum *Epoch*

No	Data	Fitur	Epoch	MSE	Akurasi (%)		Lama Train (s)
					Train	Test	
1	1056	1990	1	0,82	0,17	0,11	6
2			5	0,14	0,92	0,45	27
3			**10**	**0,02**	**0,97**	**0,47**	**77**
4			50	0,01	0,97	0,46	418
5			100	0,01	0,97	0,42	1276
6	3914	4551	1	0,51	0,61	0,52	406
7			5	0,11	0,92	0,54	1624
8			**10**	**0,06**	**0,94**	**0,57**	**2959**
9			50	0,06	0,94	0,57	16063
10			100	0,06	0,94	0,55	49035

- Maksimum *epoch* tidak memengaruhi akurasi.
- Namun, dengan *epoch* terlalu kecil, akurasi juga akan berpengaruh. Karena sistem tidak memiliki cukup kesempatan untuk mempelajari data.

Uji Coba dan Analisa Maksimum *Epoch*

- Variasi maksimum *epoch* mempengaruhi lama waktu *training*. Semakin banyak *epoch*, semakin lama waktu *training*.

Uji Coba dan Analisa Persentase Data Set

No	Data	Fitur	Persentase	Jumlah Data	
				Training	Testing
1	1056	1990	10%	951	105
2			20%	845	211
3			50%	528	528
4			80%	212	844
5	3914	4551	10%	3523	391
6			20%	3132	782
7			50%	1957	1957
8			80%	783	3131

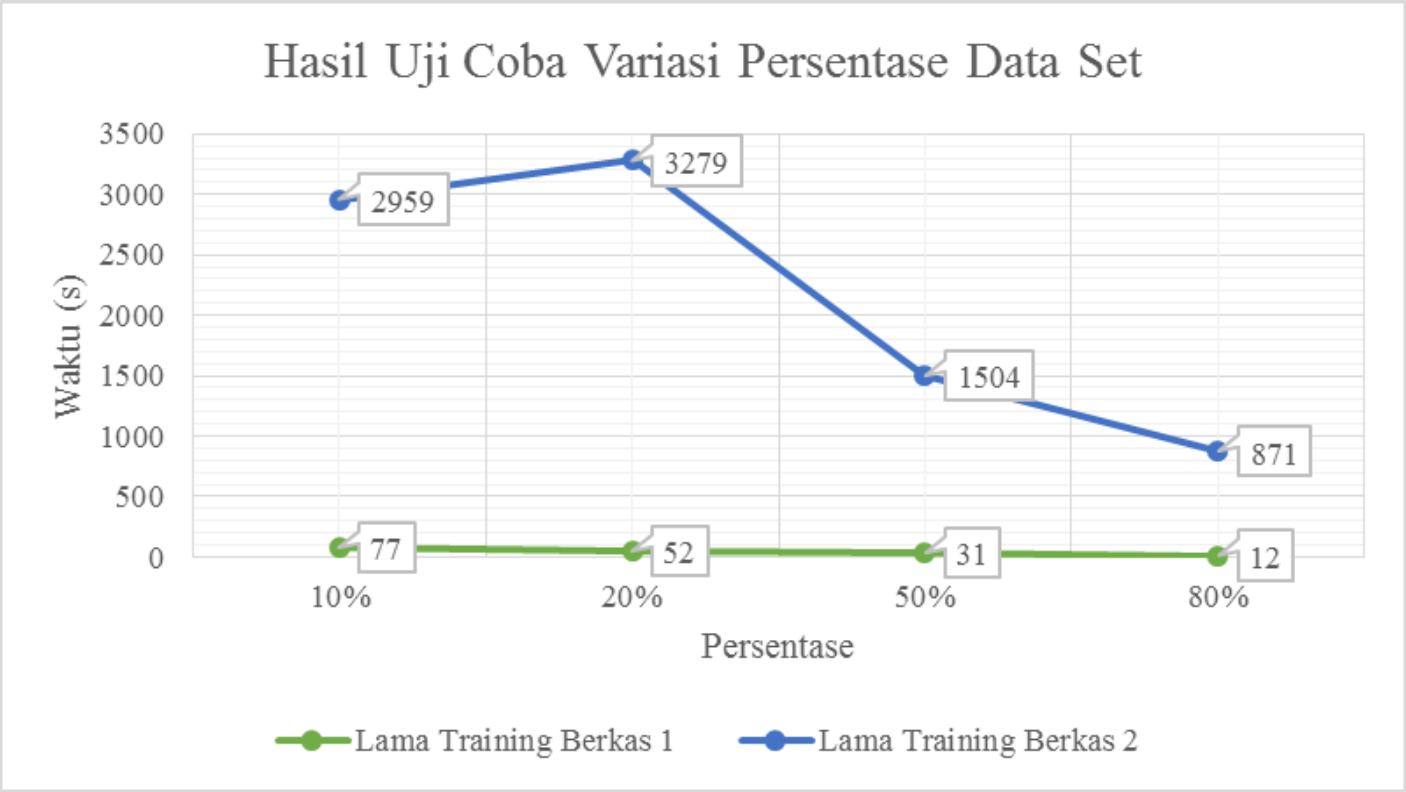
- 10% artinya data *testing* diambil 10% dari data set, dan 90% sisanya digunakan sebagai data *training*.

Uji Coba dan Analisa Persentase Data Set

No	Data	Fitur	Persen	MSE	Akurasi (%)		Lama Train (s)
					Train	Test	
1	1056	1990	10%	0,02	0,97	0,47	77
2			20%	0,01	0,98	0,51	52
3			50%	0,02	0,97	0,48	31
4			80%	0,35	0,70	0,32	12
5	3914	4551	10%	0,06	0,94	0,57	2959
6			20%	0,06	0,94	0,58	3279
7			50%	0,04	0,94	0,55	1504
8			80%	0,03	0,92	0,50	871

- Diperlukan data lebih banyak pada data *training* dibandingkan dengan data *testing* untuk mendapat nilai akurasi yang baik.
- Namun, berkas ke-2 dengan 80% mendapat akurasi baik karena 20% data sebagai data *training* dari 3914 sudah mewakili cukup banyak.

Uji Coba dan Analisa Maksimum *Epoch*



- Semakin besar persentase data *training*, sistem membutuhkan waktu lebih lama untuk *training*.

Uji Coba dan Analisa Variasi Terbaik 1056 Data

No	Variabel		MSE	Akurasi (%)		Lama Train (s)
				Train	Test	
1	Learning Rate Neuron Persentase	: 0,05 : 1% : 10%	0,07	0,96	0,50	80
2	Learning Rate Neuron Persentase	: 0,10 : 1% : 10%	0,02	0,97	0,47	87
3	Learning Rate Neuron Persentase	: 0,05 : 1% : 20%	**0,01**	0,95	**0,54**	**52**
4	Learning Rate Neuron Persentase	: 0,10 : 1% : 20%	**0,01**	**0,98**	0,52	**52**

- Variasi ke-3 dan ke-4 memiliki kemiripan. Namun, perbedaan pada akurasi *training* dan *testing*.

Uji Coba dan Analisa Variasi Terbaik 3914 Data

No	Variabel		MSE	Akurasi (%)		Lama Train (s)
				Train	Test	
1	Learning Rate Neuron Persentase	: 0,05 : 1% : 10%	0,07	**0,94**	0,55	3497
2	Learning Rate Neuron Persentase	: 0,10 : 1% : 10%	**0,06**	**0,94**	0,57	**2959**
3	Learning Rate Neuron Persentase	: 0,05 : 1% : 20%	**0,06**	**0,94**	**0,58**	4461
4	Learning Rate Neuron Persentase	: 0,10 : 1% : 20%	**0,06**	**0,94**	**0,58**	3280

- Dengan pertimbangan data sebelumnya, maka variasi terbaik diambil variasi ke-4.

Tabel Kebenaran Klasifikasi 1056 Data

Kelas		Prediksi						Total
		1	2	3	4	5	6	
Target	1	138	0	0	0	0	0	138
	2	1	138	0	0	0	0	139
	3	0	0	139	0	0	0	139
	4	0	1	0	131	0	0	132
	5	18	0	0	0	136	0	154
	6	0	0	0	1	0	142	143
Total		157	139	139	132	136	142	845

Data *Training*

Kelas		Prediksi						Total
		1	2	3	4	5	6	
Target	1	19	0	3	5	4	1	32
	2	1	25	1	1	0	1	29
	3	3	0	18	2	4	2	29
	4	2	5	4	11	3	6	31
	5	11	3	3	11	19	2	49
	6	4	0	4	11	4	16	39
Total		40	33	33	41	34	28	209

Data *Testing*

- Pada tahap *training*, 18 data dari kelas *Recreation & Leisure* diprediksi sebagai kelas *Career & Education*.

Tabel Kebenaran Klasifikasi 3914 Data

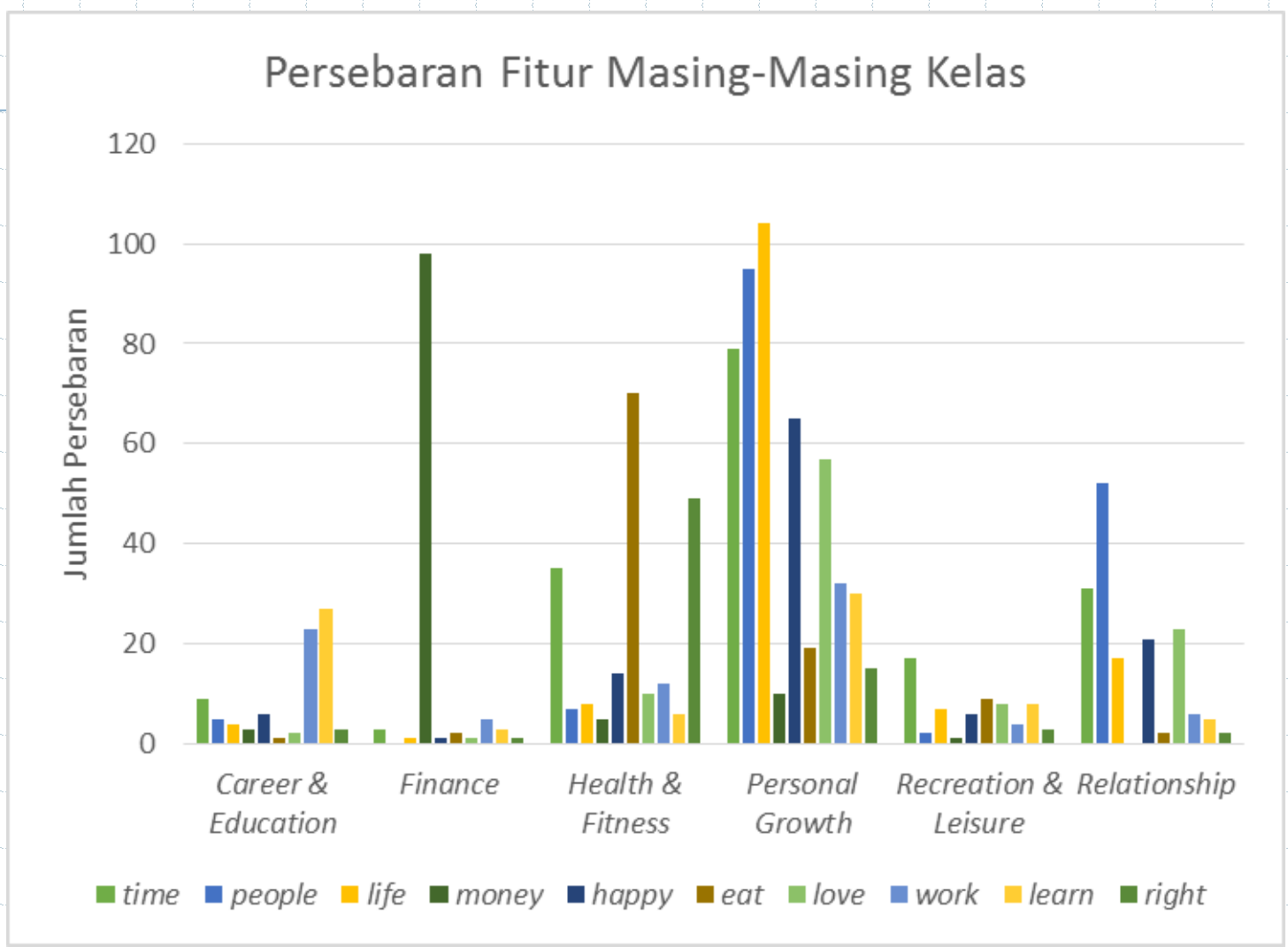
Kelas		Prediksi						Total
		1	2	3	4	5	6	
Target	1	149	0	1	0	0	2	152
	2	0	131	1	3	0	0	135
	3	1	0	636	10	5	0	652
	4	106	2	7	1375	17	20	1527
	5	1	0	0	7	327	3	338
	6	0	0	0	5	0	323	328
Total		257	133	645	1400	349	348	3132

Data *Training*

Kelas		Prediksi						Total
		1	2	3	4	5	6	
Target	1	17	1	7	9	5	0	39
	2	2	27	2	5	0	0	36
	3	6	2	104	21	12	6	151
	4	33	7	52	254	44	43	433
	5	8	1	2	17	29	6	63
	6	2	1	6	15	12	24	60
Total		68	39	173	321	102	79	782

Data *Testing*

- Pada tahap *training*, 106 data dari kelas *Personal Growth* diprediksi sebagai kelas *Career & Education*.
- Jumlah data pada kelas *Personal Growth* yang dominan, menyebabkan kecenderungan prediksi condong ke kelas tersebut.


Persebaran Fitur Masing-Masing Kelas
Jumlah Persebaran
120
100
80
60
40
20
0
Career & Education
Finance
Health & Fitness
Personal Growth
Recreation & Leisure
Relationship
time
people
life
money
happy
eat
love
work
learn
right

Kesimpulan

- *Learning rate* mempengaruhi nilai akurasi.
- *Neuron* pada *hidden layer* yang terlalu sedikit mempengaruhi kecilnya nilai akurasi.
- Maksimum *epoch* tidak mempengaruhi nilai akurasi. Namun, *epoch* yang terlalu sedikit mengurangi nilai akurasi karena sistem memerlukan kesempatan lebih lama untuk mempelajari sistem.
- Untuk sistem *training* klasifikasi perlu dipertimbangkan jumlah data untuk masing-masing kelas, karena adanya dominasi data pada kelas tertentu mempengaruhi hasil klasifikasi. Dan, menyebabkan banyak kesalahan klasifikasi data.