

Identifikasi Bahasa Artikel Publikasi Ilmiah Menggunakan Jaringan Syaraf Tiruan Berbasis Unigram

Adi Ariyo Munandar, Imam Much Ibnu Subroto, Dedy Kurniadi

Jurusan Teknik Informatika, Universitas Islam Sultan Agung

Correspondence Author: imam@unissula.ac.id

Abstract

Bahasa merupakan alat komunikasi untuk bisa saling berhubungan satu sama lain. Dengan bahasa manusia dapat membuat sebuah karya, salah satunya dalam bentuk tulisan. Tulisan yang dibuat berupa penelitian yang dimuat dalam bahasa publikasi. Bahasa publikasi dibedakan menjadi dua macam diantaranya adalah bahasa publikasi nasional dan internasional. Publikasi nasional yaitu publikasi yang dibuat oleh penulis berdasarkan asal negara pembuat. Publikasi internasional adalah publikasi yang dibuat penulis dengan menggunakan bahasa yang banyak diakui oleh banyak negara. Dan ditulis dengan menggunakan bahasa resmi perserikatan bangsa-bangsa yaitu Arab, Inggris, Perancis, Rusia, Spanyol, dan Tiongkok. Oleh karena itu pentingnya sebuah dokumen untuk mengklasifikasikan kedalam bahasa publikasi nasional atau internasional. Untuk itu diperlukan sebuah sistem yang dapat membedakan bahasa. Salah satunya dengan membuat sebuah sistem Identifikasi bahasa publikasi jaringan syaraf tiruan berbasis unigram. Sistem tersebut mampu membedakan bahasa penulisan. Bahasa penulisan akan di ekstrak dengan menggunakan unigram, kemudian diterapkan metode jaringan syaraf tiruan untuk mengambil sebuah keputusan. Jaringan syaraf tiruan merupakan salah satu metode kecerdasan buatan. Prosesnya terdiri dari preprocessing, training dan testing dan evaluasi hasil. Sehingga hasilnya, sistem akan mampu mengenali bahasa penulisan tersebut. Dokumen berbahasa indonesia akan menghasilkan angka diatas sama dengan 0.5 dan dokumen berbahasa inggris menghasilkan angka dibawah 0.5. Hasil evaluasi precision bahasa indonesia 98.4% dan bahasa inggris 98%. Untuk recall bahasa indonesia 98% dan bahasa inggris 98.4%.

Keyword: Bahasa, Jaringan Syaraf Tiruan, Unigram

1. PENDAHULUAN

Bahasa adalah sistem lambang bunyi berartikulasi yang bersifat sewenang-wenang dan konvensional yang dipakai sebagai alat komunikasi untuk melahirkan perasaan dan pikiran. Bahasa publikasi dibedakan menjadi dua macam diantaranya adalah bahasa publikasi nasional dan internasional. Oleh karena itu pentingnya sebuah dokumen untuk mengklasifikasikan kedalam bahasa publikasi nasional atau internasional. Akan tetapi sistem tidak dapat membedakan bahasa tersebut.

Dari permasalahan diatas maka diperlukan sebuah sistem yang dapat mengatasi permasalahan tersebut. Salah satunya dengan membuat sebuah sistem identifikasi bahasa publikasi dengan menggunakan jaringan syaraf tiruan berbasis unigram. Sistem tersebut akan melakukan pekerjaannya untuk mengidentifikasi dokumen yang ada. Identifikasi dokumen dilakukan dengan menggunakan algoritma unigram yang akan menghitung setiap frekuensi yang ada. Kemudian diterapkan metode jaringan syaraf tiruan yang akan melakukan proses identifikasi berdasarkan data dokumen. Sehingga sistem akan mampu mengenali bahasa penulisan yang digunakan.

Penelitian dengan *N-Gram* telah dilakukan oleh Alifian Sukma, Bagus Puji Santoso, Dian Ramadhan, Ni Made Ayu Karina Wiraswari, Tiara Ratna Sari dengan judul Klasifikasi Dokumen Bahasa Jawa menggunakan Metode *N-Gram*. Penelitian ini untuk membedakan dokumen bahasa jawa ngoko, madya dan kromo. Ada 30 dokumen yang diuji, terdiri dari 10 dokumen jawa ngoko, 10 dokumen jawa madya dan 10 dokumen jawa kromo. Setiap dokumen diuji dengan menggunakan algoritma *N-Gram*. Dan untuk mendapatkan hasil yang akurat, Evaluasi keakuratan dilakukan menggunakan parameter *precision*. Hasilnya, dokumen tersebut diklasifikasikan sesuai jenis dokumen dan diurutkan berdasarkan frekuensi dan jarak [1].

Beberapa penelitian mengenai identifikasi bahasa telah dilakukan oleh Amir Hamzah, dengan judul Deteksi Bahasa Untuk Dokumen Teks Berbahasa Indonesia. Penelitian dilakukan dengan menggunakan dokumen berbahasa indonesia dan bahasa inggris. Peneliti menerapkan metode *N-gram*, untuk menghitung frekuensi. Ada 3 frekuensi yang didapat yaitu dengan *Unigram*, *bigram* dan *trigram* dan dihitung berdasarkan probabilitas. Kemudian 3 frekuensi tersebut dibandingkan. Hasil yang diperoleh bahwa Kemampuan identifikasi *trigram* dan *bigram* lebih baik dari *Unigram*. Nilai *threshold* untuk *Unigram* adalah 0.5, untuk *bigram* 0.4 dan untuk *trigram* adalah 0.05 [2].

Selain itu, penelitian dengan menggunakan Jaringan Syaraf Tiruan telah dilakukan oleh Esa Prakasa, Edi Kurniawan, Purnomo Husnul Khotimah, Taufiq Wirahman dengan judul Klasifikasi Bahasa Dokumen Dengan Menggunakan Jaringan Syaraf Tiruan. Penelitian dilakukan dengan menggunakan beberapa variabel yaitu: distribusi vokal per kalimat, distribusi konsonan per kalimat, rata-rata jumlah vokal per kata, dan rata-rata jumlah konsonan per kata. Keempat ciri ini diharapkan menjadi variabel pembeda yang cukup signifikan dalam mengenali

jenis bahasa. Algoritma jaringan saraf tiruan diterapkan untuk mengolah data keempat variabel tersebut, sehingga diperoleh keluaran yang mampu mengenali 6 jenis bahasa, yaitu: Indonesia, Malaysia, Inggris, Jerman, Italia, dan Portugis[3].

Jaringan Syaraf Tiruan (JST) merupakan salah satu representasi buatan dari otak manusia yang selalu mencoba untuk mensimulasikan proses pembelajaran pada otak manusia tersebut. Istilah buatan digunakan karena jaringan syaraf ini diimplementasikan dengan menggunakan program komputer yang mampu menyelesaikan sejumlah proses perhitungan selama proses pembelajaran[4].

Pemodelan ini didasari oleh kemampuan otak manusia dalam mengorganisasikan sel-sel penyusunnya yang disebut *neuron*, sehingga mampu melaksanakan tugas-tugas tertentu, khususnya pengenalan pola dengan efektivitas yang sangat tinggi. Sebuah *neuron* bisa memiliki banyak masukan bagi *neuron-neuron* yang lain. Aliran sinyal masukan ini dikalikan dengan suatu penimbang (bobot sinapsis) dan kemudian dilakukan penjumlahan terhadap semua masukan yang telah dikalikan dengan bobot. Hasil penjumlahan tersebut, kemudian diaktivasi menggunakan fungsi aktivasi untuk menghasilkan kriteria atau keluaran tertentu[5].

2. METODE PENELITIAN

2.1 Tahap Pengumpulan Dokumen

Dokumen yang dibutuhkan, berupa sebuah *file* dokumen yang ber-*ekstensi* Pdf. Dokumen harus menggunakan penulisan berbahasa Indonesia dan berbahasa Inggris. Dokumen yang akan digunakan sebanyak 500 dokumen, yang terdiri dari 200 data *training* dan 300 data *testing*. Dokumen didapatkan dari berbagai sumber seperti jurnal online, jurnal penelitian dosen, Google Cendekia, dan lain-lain.

2.2 Text Preprocessing

a. Text Cleaning

Text Cleaning adalah proses pembersihan sebuah dokumen agar dapat diolah menjadi sebuah data. Dokumen di *Konversi* dari yang semula berbentuk *file PDF* diubah ke *TXT*. Kemudian dokumen di *cleaning* dengan cara, menghilangkan semua tanda baca, menghilangkan angka dan mengubah semua huruf besar (A-Z) menjadi kecil (a-z).

b. Unigram

N-gram adalah sebuah metode pemotongan atau pemisahan *string* di dalam kalimat atau kata. *N-gram* diaplikasikan untuk mengambil potongan *string* atau karakter dengan jumlah (*n*) tertentu dari sebuah kata yang ada secara berkesinambungan dari awal hingga akhir [6]. *Uni* yang berarti satu. Sehingga *unigram* mempunyai arti memisahkan satu potongan karakter dalam setiap teks atau kalimat.

Proses pemisahan dilakukan dengan menggunakan rumus *unigram*. Rumus *unigram* adalah $n=1$. Yang berarti dalam penelitian ini memisahkan potongan-potongan huruf dari kata atau kalimat.

Contoh :

Teks :

STEMMINGWORDS

Uni-gram : S, T, E, M, M, I, N, G, W, O, R, D, S

c. Token

Tokenization disini adalah memisahkan kata, simbol, frase, dan entitas penting lainnya dari sebuah *text* untuk kemudian di analisa. *Token* tersedia dalam bentuk huruf. Huruf yang menjadi token akan mencari berdasarkan unigram $n=1$.

Tabel 1 Token Alfabet

NO	HURUF	NO	HURUF
1	a	14	n
2	b	15	o
3	c	16	p
4	d	17	q
5	e	18	r
6	f	19	s
7	g	20	t
8	h	21	u
9	i	22	v
10	j	23	w
11	k	24	x
12	l	25	y
13	m	26	z

Pada tabel 1 *Token Alfabet* adalah *token* yang digunakan untuk mencari dan memisahkan potong-potongan karakter yang ada. Setiap dokumen akan dicari berdasarkan token tersebut. Setiap dokumen, akan dicari berupa huruf a,b,c,d sampai dengan z. Sehingga setiap karakter akan memisahkan diri berdasarkan token tersebut.

d. *Vector Space Model*

Vector Space Model sering digunakan untuk mempresentasikan sebuah dokumen dalam ruang *vector*. VSM merupakan model *Information Retrieval* yang mempresentasikan dokumen dan *query* sebagai *vector* pada ruang *multidimensi*. Kesamaan suatu dokumen dengan *query* dapat diukur dengan *vector* dokumen dan *vector query* [7].

Hasil dari *tokenisasi*, kemudian direpresentasikan kedalam bentuk *vektor space model*. Yaitu berupa jumlah frekuensi setiap token huruf yang ada.

Huruf sebanyak 26 yang terdiri a-z akan menjadi *token*. Kemudian setiap *token* akan dipisah dan dikelompokkan berdasarkan *token* tersebut. Sehingga setiap huruf menghasilkan jumlah *frekuensi* Dan hasil tersebut dibuat kedalam bentuk *vektor space model*.

e. *Normalisasi*

Normalisasi merupakan proses perbaikan data yang dilakukan untuk menghindari jumlah perhitungan dimensi kata yang melebar. Perhitungan dimensi kata akan melebar jika kata yang salah eja atau disingkat tidak diubah karena kata tersebut sebenarnya memiliki kontribusi dalam merepresentasikan dokumen tetapi akan dianggap sebagai entitas yang berbeda proses penyusunan matriks.

Normalisasi pada tahap ini dilakukan dengan cara membagi total *frekuensi token* dengan total *frekuensi* semua huruf. Hal ini dilakukan untuk menghasilkan jumlah setiap *token* yang ada, angkanya tidak boleh lebih dari angka 1.

Hasil normalisasi berupa angka setiap *tokennya* akan menjadi variabel *inputan* untuk diolah kedalam jaringan syaraf tiruan.

rumus normalisasi :

$$\text{Normalisasi} = \frac{\text{Total frekuensi huruf setiap token}}{\text{Total frekuensi huruf}} \quad (1)$$

2.3 Training dan testing

3.3.1 Mempersiapkan Data

Data adalah kumpulan informasi yang diperoleh dari suatu pengamatan, dapat berupa angka, lambang atau sifat. Data disini didapat dari hasil *text mining* yaitu pada tahap *preprocessing*. Dimana data-data tersebut sudah siap untuk diolah kedalam simulasi jaringan syaraf tiruan.

Data dalam penelitian ini dibagi menjadi 3, yaitu:

a. *Data Training*

Data training adalah data yang digunakan untuk proses pelatihan. Pada penelitian ini disiapkan data training sebanyak 200 sampel data training. 200 sampel tersebut dibagi menjadi 2 yaitu 100 data *training* dengan menggunakan bahasa indonesia dan 100 data training dengan menggunakan bahasa inggris. Data *training* digunakan untuk mendapatkan bobot hasil *training* yang sesuai.

b. *Data Testing*

Data *testing* adalah data yang digunakan untuk proses pengujian. Proses pengujian dilakukan dengan menggunakan bobot yang berasal dari hasil data *training*. Data *testing* menggunakan dokumen sebanyak 300 data. Yang terdiri dari 150 data berbahasa indonesia dan 150 data berbahasa inggris.

c. *Data Target*

Data target berisi hasil atau capaian data yang diinginkan. Pada penelitian data target dibagi menjadi 2 kriteria, yaitu data dengan target 1 untuk dokumen berbahasa indonesia dan data dengan target 0 untuk dokumen berbahasa inggris. pada data target diterapkan *threshold* pada angka 0.5. Dengan arti jika target lebih besar sama dengan 0.5 keatas akan mengikuti target 1 sedangkan 0.5 kebawah akan mengikuti target 0.

Data target dipersiapkan sebanyak jumlah data *training* dan data *testing* yang dipersiapkan untuk keperluan proses pembelajaran pada bobot yang didapatkan nantinya. Pembuatan data target harusurut dan sesuai dengan kriteria pada dokumen data *training* dan *testing*.

3.3.2 Tahap Pembelajaran dengan Matlab

Proses pembelajaran dalam penelitian ini dilakukan dengan menggunakan bantuan simulasi matlab. Matlab digunakan karena didalam matlab sudah tersedia simulasi dan *tool* tentang jaringan syaraf tiruan.

Berikut adalah proses pembelajaran yang dilakukan didalam matlab untuk mendapatkan hasil yang sesuai:

. Untuk membuka aplikasi matlab, cukup dengan melakukan mengklik 2 kali *icon* aplikasi matlab. Setelah itu akan muncul tampilan taskbar pada matlab. Setelah tampilan awal pada matlab tampil. Kemudian menulis beberapa baris perintah di *comman window*. Berikut beberapa baris perintah yang digunakan;

```

>> P=xlsread('DATA_Training');
>> T=xlsread('DATA_Target_Training');
>> Test=xlsread('Data_Testing');
>> ouputActual=xlsread('Data_Target_Test');
>> %plotpv(P,T);
>> net = newff(P,T,1);
>> [net,tr] = train(net,P,T);

```

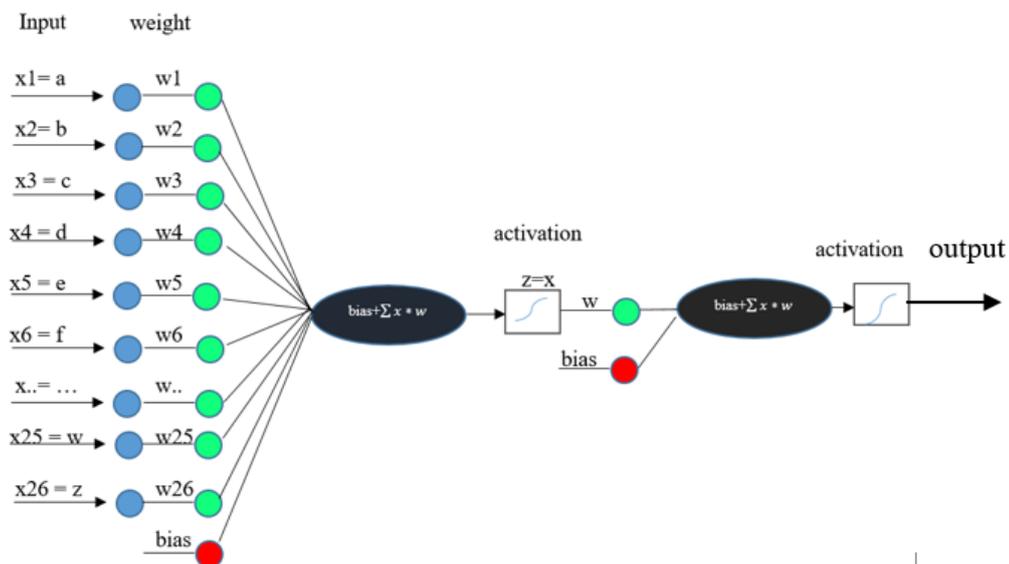
2.4 Bobot dan bias dari Matlab

Setelah proses *training* selesai dan hasil sesuai yang diinginkan. Langkah selanjutnya yaitu melihat bobot dan bias yang digunakan. Bobot dan bias didapat dari proses *training* yang telah dilakukan. Hal ini dilakukan, untuk menerapkan bobot dan bias kedalam sistem yang akan dibuat nantinya. Pada dasarnya bobot dan bias yang didapat akan menjadi sebuah pemicu kecerdasan dalam proses pembelajaran. Untuk mendapatkan bobot dan bias, dapat dilakukan dengan membedah proses *training* tersebut. Proses tersebut dilakukan dengan menuliskan kembali beberapa baris perintah. Yang ditulis di *command window* pada tampilan *taskbar* matlab.

2.5 Jaringan Syaraf Tiruan

Jaringan Syaraf Tiruan (JST) merupakan salah satu representasi buatan dari otak manusia yang selalu mencoba untuk mensimulasikan proses pembelajaran pada otak manusia tersebut. Istilah buatan digunakan karena jaringan syaraf ini diimplementasikan dengan menggunakan program komputer yang mampu menyelesaikan sejumlah proses perhitungan selama proses pembelajaran[4].

Gambar 1 adalah langkah-langkah detail dari algoritma jaringan syaraf tiruan yang diterapkan kedalam sistem.



Gambar 1 Arsitektur JST

Perhitungan akan dilakukan dengan melihat dan menerapkan alur sesuai dengan arsitektur jaringan syaraf tiruan sesuai pada tampilan gambar 1 Arsitektur JST.

Algoritma :

```

langkah 1:
    Proses hidden layer adalah proses awal dimana sebanyak 26 input
    Hidden Layer = (bias Input + (Input*Bobot Input))
langkah 2: Aktivasi layer= 1 / (1+exp(-2*Hidden Layer))-1
langkah 3: Output Layer = (bias Output + (Aktivasi layer*Bobot Output))
langkah 4: Output = 1 / (1+exp(-2*Output Layer))-1
Langkah 5;
    Output>= 0.5 , Hasilnya Bahasa INGGRIS
    Output< 0.5 , Hasilnya Bahasa INDONESIA

```

2.6 Evaluasi menggunakan Akurasi, *Precision* dan *Recall*

Tujuan uji *Recall* dan *Precision* adalah untuk mendapatkan informasi hasil pencarian yang didapatkan oleh *information retrieval*. Hasil pencarian *information retrieval* bisa dinilai tingkat *recall* dan *precision* nya. *Precision* dapat dianggap sebagai ukuran ketepatan atau ketelitian, sedangkan *recall* adalah kesempurnaan. Nilai *precision* adalah proporsi dokumen yang terambil oleh sistem adalah *relevan*. Nilai *recall* adalah proporsi dokumen relevan yang terambil oleh sistem [8]

Kemudian untuk mengukur kinerja sistem, Pada penelitian identifikasi jaringan syaraf tiruan berbasis unigram hanya menghasilkan 2 keluaran, yaitu bahasa inggris dan indonesia. Sehingga data dapat dikelompokkan disajikan dalam bentuk tabel *confusion matrix*.

Tabel 2 *confusion matrix*

JUMLAH TOTAL (N)	HASIL PREDIKSI	
	INGGRIS	INDONESIA
INGGRIS	<i>True Negatif (TN)</i>	<i>False Positif (FP)</i>
INDONESIA	<i>False Negatif (FN)</i>	<i>True Positif (TP)</i>

Pada Tabel 2 *confusion matrix* merupakan tabel yang digunakan untuk penyajian data pada proses penelitian ini. Pada tabel tersebut, data hasil penelitian akan dipisahkan dan dikelompokkan.

Sehingga dari data diatas dapat menghasilkan rumus ;

$$AKURASI = \frac{TP+TN}{TP+TN+FP+FN} * 100\% \quad (2)$$

$$Precision = \frac{TP}{FP+TP} * 100\% \quad (3)$$

$$RECALL = \frac{TP}{FN+TP} * 100\% \quad (4)$$

Keterangan :

TP = *True Positif*

TN = *True Negatif*

FP = *False Positif*

FN = *False Negatif*

3. HASIL DAN ANALISA

3.1 Hasil *Preprocessing*

Berikut hasil *text preprocessing* dari salah satu dokumen yang digunakan dalam proses penelitian identifikasi bahasa publikasi menggunakan jaringan syaraf tiruan berbasis unigram.

Tabel 3 Hasil *Preprocessing*

ABJAD	Frekuensi	NORMALISASI
a	1975	0.1757
b	270	0.024
c	63	0.0056
d	494	0.0439

e	931	0.0828
f	104	0.0093
g	409	0.0364
h	211	0.0188
i	1187	0.1056
j	159	0.0141
k	472	0.042
l	528	0.047
m	428	0.0381
n	965	0.0858
o	228	0.0203
p	326	0.029
q	2	0.0002
r	623	0.0554
s	625	0.0556
t	533	0.0474
u	450	0.04
v	39	0.0035
w	61	0.0054
x	5	0.0004
y	144	0.0128
z	10	0.0009

Tabel 3 Hasil *Preprocessing* adalah tabel hasil dari *Preprocessing* pada dokumen “Studi Tentang Regulasi Diri Dalam Belajar, Efikasi Diri Dan Prestasi Belajar Matematika.Pdf”. Data ini didapatkan setelah melalui proses *Preprocessing* yaitu, proses *text cleaning*, unigram, *token*, *vektor space model* dan normalisasi.

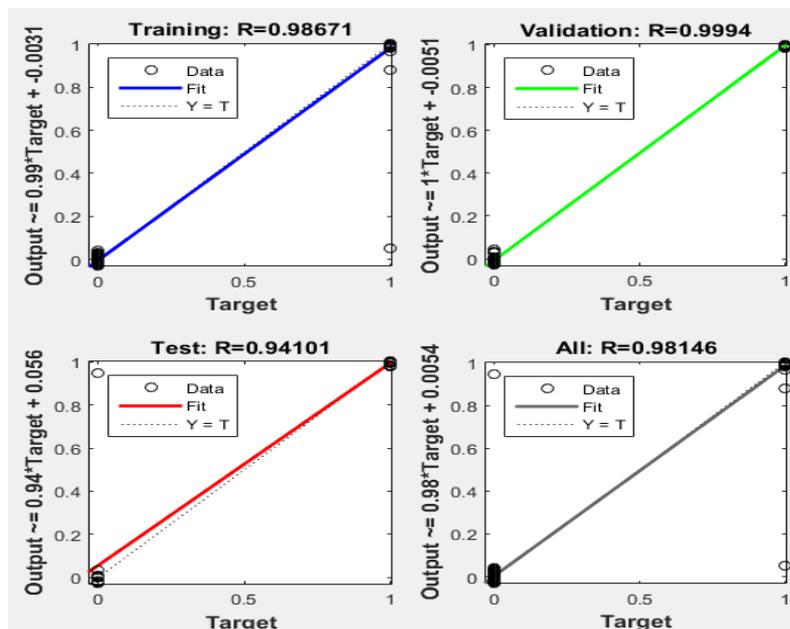
Setiap dokumen yang telah melalui proses *Preprocessing*, Hasilnya menjadi *variabel inputan*. Dimana pada dokumen tersebut hasil normalisasi dari *token a-z*. Yang terdapat pada kolom normalisasi akan menjadi *variabel input* pada dokumen tersebut.

Inputan tersebut yang akan melakukan proses selanjutnya, yaitu proses perhitungan dan pengambilan keputusan dengan menggunakan metode jaringan syaraf tiruan. Untuk menentukan bahasa penulisan yang digunakan dokumen tersebut

3.2 Simulasi Training dan Testing dengan Matlab

1. Hasil Simulasi

Setelah proses *training* selesai, hasil yang didapat disimpulkan dengan melihat proses *regresion*. Hal ini dilakukan untuk memastikan bahwa data *training* dan *validasi* berada dalam satu garis lurus. Yang menandakan bahwa proses *training* berjalan dengan sempurna dan hasil yang diinginkan.



Gambar 2 Evaluasi *Training* dengan ROC

Gambar 2 Evaluasi *Training dengan ROC* merupakan tampilan hasil *training* yang telah dilakukan dalam bentuk diagram. Hasil tampilan diatas dapat diartikan bahwa *training* dan target hasilnya sangat bagus dengan gradien mendekati nilai satu baik untuk training, validation, maupun testing dengan rata-rata total 0,98146.

2. Bobot dan bias yang dihasilkan dari proses simulasi matlab

Berikut baris perintah dan hasil, berupa nilai bobot dan bias dari hasil training:

```
>> net.IW{1,1}
Fungsi untuk memanggil bobot pada hidden layer
Berikut bobot yang dihasilkan:
-19.6533,7.5436 , 8.5946 , 3.2458 , 7.6040 , 7.4160 , -4.8679 , 3.9757,
-5.2880 , -7.8170 , -19.9077,2.8692 , -1.2688 , 6.3428 , 14.3673, -10.7552 ,
-2.3825 , 8.2678 , 4.4990 , 4.7891 , -18.1358 , 3.0688, 3.7537 , -1.7298 ,
-4.8742 , 5.0658
>> net.b{1}
Fungsi untuk memanggil bias pada hidden layer.
Berikut bias yang dihasilkan:
-0.6086
>> net.LW{2,1}
Fungsi untuk memanggil bobot pada output layer.
Berikut bobot yang dihasilkan:
-0.0532
>> net.b{2}
Fungsi untuk memanggil bias pada output layer.
-0.9465
```

3.3 Implementasi Jaringan Syaraf Tiruan

Pada tahap implementasi jaringan syaraf tiruan, hasil akhir yang didapatkan dengan nilai dibawah 0.5 akan menghasilkan Bahasa Inggris dan diatas sama dengan 0.5 akan menghasilkan bahasa indonesia. Berikut contoh hasil implementasi pada dokumen dengan judul “Studi Tentang Regulasi Diri Dalam Belajar, Efikasi Diri Dan Prestasi Belajar Matematika.Pdf”. Berikut adalah contoh perhitungannya;

Tahap 1: Input*Bobot

Tabel 4 Bobot Input

INPUT(I)	BOBOT(W)	HASIL=I*W
0.1757	-19.6533	-3.45308481
0.024	7.5436	0.1810464
0.0056	8.5946	0.04812976
0.0439	3.2458	0.14249062
0.0828	7.604	0.6296112
0.0093	7.416	0.0689688
0.0364	-4.8679	-0.17719156
0.0188	3.9757	0.07474316
0.1056	-5.288	-0.5584128
0.0141	-7.817	-0.1102197
0.042	-19.9077	-0.8361234
0.047	2.8692	0.1348524
0.0381	-1.2688	-0.04834128
0.0858	6.3428	0.54421224
0.0203	14.3673	0.29165619
0.029	-10.7552	-0.3119008
0.0002	-2.3825	-0.0004765
0.0554	8.2678	0.45803612
0.0556	4.499	0.2501444
0.0474	4.7891	0.22700334
0.04	-18.1358	-0.725432
0.0035	3.0688	0.0107408
0.0054	3.7537	0.02026998
0.0004	-1.7298	-0.00069192

0.0128	-4.8742	-0.06238976
0.0009	5.0658	0.00455922

Tahap 2: *hidden layer*

$$\begin{aligned} \text{Hidden Layer} &= (-0.6086 + (-3.1977999)) \\ &= -3.8063999 \end{aligned}$$

Tahap 3: *Aktivasi layer*

$$\begin{aligned} \text{Aktivasi layer} &= 1 / (1 + \exp(-2 * -3.8063999)) - 1 \\ &= -0.99950615742446 \end{aligned}$$

Tahap 4: *Output Layer*

$$\begin{aligned} \text{Output Layer} &= (-0.9465 + (-0.99950615742446 * -0.0532)) \\ &= 0.89283257800225 \end{aligned}$$

Tahap 5: *Output*

$$\begin{aligned} \text{Output} &= 1 / (1 + \exp(-2 * 0.89283257800225)) - 1 \\ &= 0.71278997851463 \end{aligned}$$

Tahap 6; Hasil Jaringan Syaraf Tiruan

$$0.71278997851463 \geq 0.5, \text{ Hasilnya Bahasa INDONESIA}$$

2.1 Evaluasi Hasil

1. Confusion Matrix

Tabel 5 Confusion Matrix

JUMLAH TOTAL (N)	HASIL PREDIKSI	
	INGGRIS	INDONESIA
INGGRIS	245 (TN)	4 (FP)
INDONESIA	5 (FN)	246 (TP)

2. Akurasi

$$\text{AKURASI} = \frac{246+245}{246+245+4+5} * 100\% = 98.2\%$$

Jadi, berdasarkan hasil penelitian akurasi dari implementasi sistem identifikasi bahasa menggunakan jaringan syaraf tiruan berbasis *unigram* adalah sebesar 98.2 %

3. Precision

$$\begin{aligned} \text{Precision INDONESIA} &= \frac{246}{4+246} * 100\% = 98.4\% \\ \text{precision INGGRIS} &= \frac{245}{5+245} * 100\% = 98\% \end{aligned}$$

Jadi, berdasarkan hasil penelitian *precision* dari implementasi sistem identifikasi bahasa menggunakan jaringan syaraf tiruan berbasis *unigram* adalah bahasa indonesia 98.4 % dan bahasa inggris 98%

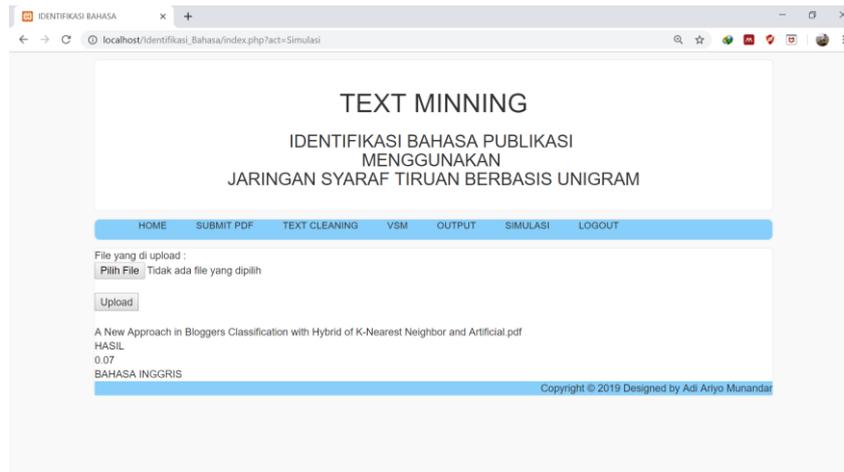
4. Recall

$$\begin{aligned} \text{RECALL INDONESIA} &= \frac{246}{5+246} * 100\% = 98\% \\ \text{RECALL INGGRIS} &= \frac{245}{4+245} * 100\% = 98.4\% \end{aligned}$$

Jadi, berdasarkan hasil penelitian *recall* dari implementasi sistem identifikasi bahasa menggunakan jaringan syaraf tiruan berbasis *unigram* adalah bahasa indonesia 98 % dan bahasa inggris 98.4%.

User Interface Hasil Sistem Jaringan Syaraf Tiruan berbasis Unigram

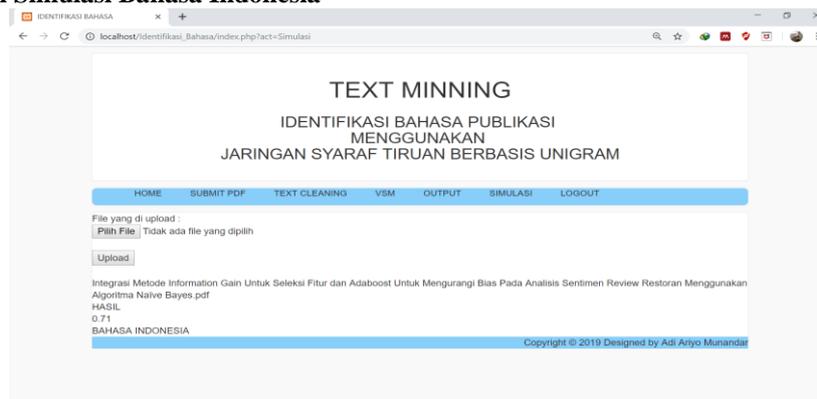
1. Tampilan Hasil Simulasi Bahasa Inggris



Gambar 3 Hasil Simulasi Bahasa Inggris

Gambar 3 Hasil Simulasi Bahasa Inggris merupakan tampilan dari hasil simulasi ketika dokumen sudah berhasil *diupload*. Hasil akan muncul dibawah *button upload* yang berisi nama dokumen, nilai dan hasil. Hasil terdiri dari bahasa indonesia dan bahasa inggris tergantung dari nilai yang dihasilkan. Dimana jika nilai $0.5 \geq$ akan menghasilkan Bahasa Indonesia dan $0.5 <$ Bahasa Inggris. Pada Gambar terlihat bahwa hasil 0.07 yang berarti dokumen menggunakan Bahasa Inggris.

2. Tampilan Hasil Simulasi Bahasa Indonesia



Gambar 4 Hasil Simulasi Bahasa Indonesia

Gambar 4 Hasil Simulasi Bahasa Indonesia merupakan tampilan dari hasil simulasi ketika dokumen sudah berhasil *diupload*. Hasil akan muncul dibawah *button upload* yang berisi nama dokumen, nilai dan hasil. Hasil terdiri dari bahasa indonesia dan bahasa inggris tergantung dari nilai yang dihasilkan. Dimana jika nilai $0.5 \geq$ akan menghasilkan Bahasa Indonesia dan $0.5 <$ Bahasa Inggris. Pada Gambar terlihat bahwa hasil 0.71 yang berarti dokumen menggunakan Bahasa Indonesia

4. KESIMPULAN

Proses penelitian dan pembuatan sistem serta uji coba simulasi berjalan dengan baik. Dokumen yang menghasilkan angka lebih besar sama dengan 0.5 akan menghasilkan dokumen berbahasa indonesia. Dokumen yang menghasilkan angka lebih kecil dari 0.5 akan menghasilkan dokumen berbahasa inggris. Penelitian menghasilkan akurasi dari implementasi sistem identifikasi bahasa menggunakan jaringan syaraf tiruan berbasis *unigram* adalah sebesar 98.2 %. *Precision* untuk bahasa indonesia 98.4 % dan bahasa inggris 98%. *Recall* untuk bahasa indonesia 98 % dan bahasa inggris 98.4%.

DAFTAR PUSTAKA

- [1] A. Sukma, B. P. Santoso, D. Ramadhan, N. M. A. K. Wiraswari, and T. R. Sari, "Klasifikasi Dokumen Bahasa Jawa Menggunakan Metode N-Gram," pp. 1–8, 2006.
- [2] A. Hamzah, "Deteksi Bahasa Untuk Dokumen Teks Berbahasa Indonesia," *Semin. Nas. Inform.*, vol. 1, no. 1, Jul. 2015.
- [3] T. W. et al. Esa Prakasa, Edi Kurniawan, Purnomo Husnul Khotimah, "Klasifikasi Bahasa Dokumen

- Dengan Menggunakan Jaringan Saraf Tiruan,” *Pros. Konf. Nas. Teknol. Inf. Komun. untuk Indones.*, pp. 477–481, 2006.
- [4] S. Kusumaningtyas and R. A. Asmara, “Identifikasi Kematangan Buah Tomat Berdasarkan Warna Menggunakan Metode Jaringan Syaraf Tiruan (Jst),” *J. Inform. Polinema*, vol. 2, no. 2, pp. 72–75, 2016.
- [5] Suyanto, *Artificial Intelligence*, Revisi ked. yogyakarta: Informatika Bandung, 2014.
- [6] S. et al. Nadya Permadi, “Stemming Words Dengan N-Gram Dan Lexeme Based Untuk Teks Berbahasa Korea,” *e-Proceeding Eng.*, vol. 2, 2015.
- [7] Amburika, “Vector Space Model,” *Pros. SNST ke-7 Tahun 2016*, pp. 10–27, 2016.
- [8] F. Amin, “Sistem Temu Kembali Informasi dengan Metode Vector Space Model,” *J. Sist. Inf. Bisnis*, vol. 2, no. 2, pp. 78–83, 2016.