

ARTIKEL

**IMPLEMENTASI METODE *LATENT DIRICHLET ALLOCATION*
UNTUK MENENTUKAN TOPIK TEKS BERITA**



Oleh:

DIMAS ARYANTO SAPUTRO

14.1.03.02.0087

Dibimbing oleh :

- 1. Intan Nur Farida, M.Kom.**
- 2. Daniel Swanjaya, M.Kom.**

**PROGRAM STUDI TEKNIK INFOTMATIKA
FAKULTAS TEKNIK
UNIVERSITAS NUSANTARA PGRI KEDIRI
TAHUN 2019**



SURAT PERNYATAAN
ARTIKEL SKRIPSI TAHUN 2019




Yang bertanda tangan dibawah ini:

Nama Lengkap : Dimas Aryanto Saputro
NPM : 14.1.03.02.0087
Telepon/HP : 085745155882
Alamat Surel (Email) : dimassen7@gmail.com
Judul Artikel : Implementasi Metode *Latent Dirichlet Allocation* Untuk Menentukan Topik Teks Berita
Fakultas – Program Studi : Teknik - Teknik Informatika
Nama Perguruan Tinggi : Universitas Nusantara PGRI Kediri
Alamat Perguruan Tinggi : Jl. K.H. Achmad Dahlan No. 76 Mojoroto Kota

Dengan ini menyatakan bahwa:

- a. artikel yang saya tulis merupakan karya saya pribadi (bersama tim penulis) dan bebas plagiarisme;
- b. artikel telah diteliti dan disetujui untuk diterbitkan oleh Dosen Pembimbing I dan II.

Demikian surat pernyataan ini saya buat dengan sesungguhnya. Apabila dikemudian hari ditemukan ketidaksesuaian data dengan pernyataan ini dan atau ada tuntutan dari pihak lain, saya bersedia bertanggungjawab dan diproses sesuai dengan ketentuan yang berlaku.

Mengetahui		Kediri, 14 Februari 2019
Pembimbing I  <u>Intan Nur Farida, M.Kom.</u> NIDN.0704108701	Pembimbing II  <u>Daniel Swanjaya, M.Kom.</u> NIDN.0723098303	Penulis,  <u>Dimas Aryanto Saputro</u> NPM. 14.1.03.02.0087

IMPLEMENTASI METODE *LATENT DIRICHLET ALLOCATION* UNTUK MENENTUKAN TOPIK TEKS BERITA

Dimas Aryanto Saputro

14.1.03.02.0087

Fakultas Teknik – Teknik Informatika

Dimassen7@gmail.com

Intan Nur Farida, M.Kom.¹, Daniel Swanjaya, M.Kom.²

UNIVERSITAS NUSANTARA PGRI KEDIRI

ABSTRAK

Dengan cara manual pemilihan berita memerlukan banyak waktu untuk memeriksa semua berita yang ada.

Berdasarkan pernyataan di atas akan dilakukan proses Implementasi Metode *Latent Dirichlet Allocation* Untuk Menentukan Topik Teks Berita. Dari proses *Latent Dirichlet Allocation* akan diperoleh dokumen dengan topik yang memiliki bobot.

Dengan perhitungan pada sistem yang menggunakan metode *latent dirichlet allocation*, maka dihasilkan urutan rekomendasi berita yang dijadikan berita utama berdasarkan bobotnya. Pada Sistem ini dari 4 percobaan dengan dokumen yang sama menghasilkan bobot yang berbeda namun mempunyai nilai yang konsisten pada dokumen ke-5 dengan menghasilkan bobot 15.1032, 14.0063, 14.8223 dan 14.8223, dan bobot tersebut lebih besar dibandingkan dengan dokumen lainnya.

KATA KUNCI :Jaringan Syaraf Tiruan, Berita, Latent Dirichlet Allocation.

I. LATAR BELAKANG

Dalam kamus besar bahasa Indonesia, arti kata berita adalah cerita atau keterangan mengenai kejadian atau peristiwa yang hangat dan berita juga berarti laporan dan juga bisa berarti pemberitahuan.

Ada berbagai macam jenis berita, beberapa contoh jenis berita yaitu berita pendidikan, berita olahraga, berita politik, berita criminal dan berita umum. Berita dapat disampaikan dalam bentuk media cetak, lisan, radio, siaran TV dan media online. Dalam penelitian ini berita yang dipakai adalah berita dari media cetak, yaitu Koran.

Dalam sebuah perusahaan media cetak seperti Koran harian, tiap harinya akan mempublikasikan berita yang sudah dipilih dan terdapat berita utama atau headline tiap harinya.

Terdapat beberapa tahapan dalam pembuatan koran, berawal dari wartawan yang mencari berita, kemudian berita akan dikirim ke bagian tim redaksi. Tim redaksi bertugas sebagai penyunting berita, memperbaiki kesalahan ejaan, diksi, tanda baca, serta keterpaduan antar paragraf, dari naskah berita yang telah ditulis dan dikirimkan oleh wartawan. Kemudian seorang redaktur akan memilih berita yang

layak untuk dipublikasikan dan memilih beberapa berita untuk dijadikan headline. Cara pemilihan berita dan headline berita masih menggunakan cara yang manual, yaitu dengan memeriksa semua berita dalam bentuk teks yang masuk, lalu memilih berita yang sesuai untuk dipublikasi dan memilih salah satu atau lebih berita untuk dijadikan headline. Kemudian berita yang telah lolos penyuntingan akan diproses oleh layouter yang bertugas menata berita ke dalam layout Koran, barulah Koran siap dicetak dan di edarkan.

Cara pemilihan berita dan headline berita masih menggunakan cara yang manual, yaitu dengan memeriksa semua berita dalam bentuk teks yang masuk, lalu memilih berita yang sesuai untuk dipublikasi dan memilih salah satu atau lebih berita untuk dijadikan headline. Dengan cara manual tersebut memerlukan banyak waktu untuk memeriksa semua berita yang ada.

Berdasarkan latar belakang diatas, penulis mengajukan penelitian berjudul *Implementasi Metode Latent Dirichlet Allocation Untuk Menentukan Topik Pada Teks Berita*. Diharapkan dengan adanya penelitian ini mampu mempermudah

perusahaan media cetak Koran harian dalam pemilihan *headline* berita, yaitu dengan cara pemilihan teks berita secara otomatis

II. METODE

2.1 Jaringan Syaraf Tiruan

Jaringan syaraf tiruan pertama kali didesain oleh Waren McCulloch dan Walter Pits (1943), namun teknologi yang tersedia pada saat itu belum memungkinkan mereka berbuat lebih jauh (sutikno 2016). Cara kerja JST seperti cara kerja manusia yaitu belajar melalui contoh. Sebuah JST konfigurasi untuk aplikasi tertentu, seperti pengenalan pola atau klasifikasi data melalui proses pembelajaran (Sunarya 2013).

2.2 Teks Mining

Text mining adalah satu langkah dari analisis teks yang dilakukan secara otomatis oleh komputer untuk menggali informasi yang berkualitas dari suatu rangkaian teks yang terangkum dalam sebuah dokumen (Han & Kamber, 2006).

2.3 Latent Dirichlet Allocation

Topik dapat ditentukan dengan memperhatikan beberapa kata-kata penting yang ada pada dokumen dan tidak terletak secara berurutan, kumpulan kata yang tidak berurutan ini disebut *bag-of-words*. Usaha untuk menentukan topik pada dokumen disebut ekstraksi topik, beberapa metode untuk mengekstrak topik adalah *Latent Semantic Analysis* (LSA), *Probabilistic Latent Semantic Analysis* (pLSA) dan *Latent Dirichlet Allocation* (LDA). Pada metode LSA saat mengekstraksi topik memperhatikan adanya sinonim (arti kata sama) dan polisemi (kata yang mempunyai banyak arti), hal ini menjadi kelemahan dari LSA karena harus membangun kamus sinonim dan polisemi terlebih dahulu. Metode pLSA adalah perkembangan dari LSA, dengan menambahkan model probabilistik, tetapi untuk menggunakan pLSA kita harus menyusun urutan dokumen dengan benar, apabila tertukar akan memberikan hasil yang berbeda. LDA menyempurnakan

pLSA dengan membuang ketergantungan pada urutan dokumen (Swanjaya, 2014).

Algoritma *Latent Dirichlet*

Allocation :

1. Jumlah Dokumen = M.

$$Document = \begin{bmatrix} d_1 \\ d_2 \\ d_3 \\ \dots \\ d_n \end{bmatrix}$$

2. Vector Vocab, Ekstrak *distinct* kata pada semua dokumen.

$$Vocab = \begin{bmatrix} w_1 \\ w_2 \\ w_3 \\ \dots \\ w_n \end{bmatrix}$$

3. Matriks Word per Dokumen, tidak diunikkan.

$$Word = \begin{bmatrix} w_{d_1,1} & w_{d_1,2} & \dots & w_{d_1,n} \\ w_{d_2,1} & w_{d_2,2} & \dots & w_{d_2,n} \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ w_{d_m,1} & w_{d_m,2} & \dots & w_{d_m,n} \end{bmatrix}$$

Dimana :

- d_m : Dokumen ke-m
- n : Nomor urut *term* pada dokumen
- $w_{d_m,n}$: *Term* ke-n pada dokumen ke-m

4. Tentukan jumlah distribusi Topik = K.

$$Topics = \begin{bmatrix} z_1 \\ z_2 \\ z_3 \\ \dots \\ z_n \end{bmatrix}$$

5. *Gibbs Sampling*.

- a. Untuk setiap dokumen d, secara random beri label tiap kata pada dokumen d dengan salah satu topik K.

$$Z = \begin{bmatrix} z_{w_{d_1,1}} & z_{w_{d_1,2}} & \dots & z_{w_{d_1,n}} \\ z_{w_{d_2,1}} & z_{w_{d_2,2}} & \dots & z_{w_{d_2,n}} \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ z_{w_{d_m,1}} & z_{w_{d_m,2}} & \dots & z_{w_{d_m,n}} \end{bmatrix}$$

Dimana :

- d_m : Dokumen ke-m
- n : Indeks *term* pada dokumen d_m
- $w_{d_m,n}$: *Term* ke-n pada dokumen ke-m
- $z_{w_{d_m,n}}$: Topik dari *Term* ke-n pada dokumen ke-m
- b. Distribusi kata, w_i , pada tiap topik, z_i , NW

NW=

$$NW = \begin{bmatrix} NW_{w_1,z_1} & NW_{w_1,z_2} & \dots & NW_{w_1,z_k} \\ NW_{w_2,z_1} & NW_{w_2,z_2} & \dots & NW_{w_2,z_k} \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ NW_{w_n,z_1} & NW_{w_n,z_2} & \dots & NW_{w_n,z_k} \end{bmatrix}$$

Dimana :

- w_n : *Term* ke-n pada Vocab
- z_k : Topik ke-k

NW_{w_n, z_k} : Banyak *term* ke-n yang berlabel topik ke-k

c. Distribusi topik, Z_i , pada tiap dokumen, di, ND

$$ND = \begin{pmatrix} ND_{d_1, z_1} & ND_{d_1, z_2} & \dots & ND_{d_1, z_k} \\ ND_{d_2, z_1} & ND_{d_2, z_2} & \dots & ND_{d_2, z_k} \\ \vdots & \vdots & \ddots & \vdots \\ ND_{d_n, z_1} & ND_{d_n, z_2} & \dots & ND_{d_n, z_k} \end{pmatrix}$$

Dimana :

d_m : Dokumen ke-m

Z_k : Topik ke-k

ND_{d_n, z_k} : Banyak label topik ke-k pada Dokumen ke-m

d. Jumlah distribusi NW dan ND

$$NWSum_{z_i} = \sum_{j=1}^m NW_{w_j, z_i} \dots \dots \dots \quad (2.1)$$

$$NDSum_{d_i} = \sum_{j=1}^k ND_{d_i, z_j} \dots \dots \dots \quad (2.2)$$

e. Untuk setiap kata, $w_{d_i, j}$, pada Matrik Word per Dokumen.

1. Topik = $Z_{w_{d_i, j}}$
2. $NW_{w_{d_i, j}, topik} = NW_{w_{d_i, j}, topik} - 1$
3. $ND_{d_1, topik} = ND_{d_1, topik} - 1$
4. $NWSum_{topik} = NWSum_{topik} - 1$

$$5. \quad NWSum_{d_1} = NWSum_{d_1} - 1$$

6. *Multinomial sampling* menggunakan metode komulatif.

$$s_x = \frac{NW_{w_{d_i, j}, x} + \beta}{NWSum_x + n \times \beta} \times \frac{ND_{d_1, x} + \alpha}{NDSum_{d_1} + K \times \alpha} \dots \dots \dots \quad (2.3)$$

$$p_x = \begin{cases} s_x; & x = 1 \\ \sum_{x=2}^k s_x - 1; & x > 1 \end{cases}$$

dimana :

d_i : Dokumen ke-i

j : Indeks *term* pada dokumen d_i

x : Indeks sampling topik

β : Parameter multinomial

α : Parameter dirichlet

K : Banyak topik

n : Banyak *term* pada Vocab

$w_{d_i, j}$: *Term* ke-j pada dokumen d_i

$NW_{w_{d_i, j}, x}$: Banyak *term* ke-n yang berlabel x

$ND_{d_i, x}$: Banyak label x pada dokumen

$NWSum_x$: Jumlah distribusi *term* yang berlabel

s_x : Multinomial sampling topik ke-x

P_x : Komulatif Multinomial sampling topik ke-x

7. Skala sample, $u = Random \times P_{k-1}$
8. Topik baru untuk kata, $w_{d_i,j}$, adalah indeks terkecil dari P_x yang lebih kecil dari u .
9. $NW_{w_{d_i,j}, topikBaru} = NW_{w_{d_i,j}, topikBaru} + 1$
10. $ND_{d_i, topikBaru} = ND_{d_i, topikBaru} + 1$
11. $NWSum_{topikBaru} = NWSum_{topikBaru} + 1$
12. $NWSum_{d_1} = NWSum_{d_1} + 1$

6. *Update Parameter*

$$\theta_{d_1,j} = \frac{ND_{d_i,x} + \alpha}{NDSum_{d_1} + K \times \alpha} \dots \dots \dots$$

(2.4)

$$\theta_{d_1,j} = \frac{NW_{w_{d_i,j},x} + \beta}{NWSum_x + n \times \beta} \dots \dots \dots$$

(2.5)

7. Hasilnya adalah :
 - a. Matriks Phi, Φ , probabilitas *term* pada Vocab terhadap Topik, dan Matriks Theta, θ , probabilitas Topik terhadap dokumen. Pertama-tama adalah mengolah data teks berita menjadi data indeks yang berupa angka. Data teks yang sudah melalui tahap *stemming* Kemudian akan dijadikan data indeks sesuai dengan data

indeks yang ada pada database dan dijumlahkan kata yang sama yang ada pada teks tersebut. Hasil dari data indeks dapat dilihat pada tabel 2.1 dan tabel 2.2 .

Tabel 2.1. Data Indeks 1

9176;6
238;2
15322;2
7737;2
6986;2
396;2
13821;1
6940;3
17801;4
12231;2
315;1
12060;1
10059;1
1280;1
12189;1
9948;1
1479;1
16390;1
14418;1
15727;1
6974;1
9122;1
9106;1
13439;1
1815;1
4731;1
11090;1
8646;1

Tabel 2.2. Data Indeks 2

Word	Dokumen l (jumlah)
9176	6
238	2
15322	2
7737	2
6986	2
239	2
13821	1
6940	3
17801	4
12231	2
315	1
12060	1
10059	1
1280	1
12189	1

9948	1
1479	1
16390	1
14418	1
15727	1
6974	1
9122	1
9106	1
13439	1
1815	1
4731	1
11090	1
8646	1

Kemudian secara random beri label tiap kata pada dokumen dengan salah satu topic. Hasil dari pelabelan dapat dilihat pada tabel 2.3.

Tabel 2.3 warna keanggotaan topik.

9176;6	238;2	15322;2	7737;2
6986;2	396;2	13821;1	6940;3
17801;4	12231;2	315;1	12060;1
10059;1	1280;1	12189;1	9948;1
1479;1	16390;1	14418;1	15727;1
6974;1	9122;1	9106;1	13439;1
1815;1	4731;1	11090;1	8646;1

Misal pada sebuah Dokumen seperti pada tabel 2.3 setelah diekstrak topiknya didapat distribusi topik seperti pada Tabel 2.4, dimana warna pada tabel 2.3 menunjukkan keanggotaan kata pada topik.

Tabel 2.4. Hasil pelabelan.

Topik 1 (Merah)	Topik 2 (Biru)	Topik 3 (Hijau)
9176 ;6	17801 ;4	6986 ;2
15322 ;2	238 ;2	13821;1
7737 ;2	10059 ;1	12231 ;2
396 ;2	9948 ;1	1280 ;1
6940 ;4	1479 ;1	14418 ;1
315 ;1	15727 ;1	9106 ;1
12060 ;1	9122 ;1	11090 ;1
12189 ;1	4731 ;1	8646 ;1
16390 ;1	13439 ;1	1815 ;1

Berdasarkan indeks dominan tersebut dapat disimpulkan bahwa Topik 1 adalah tentang indeks 9176, Topik 2 tentang indeks 17801, Topik 3 tentang kata dengan indeks 6986. Tetapi tidak ada aturan baku dalam penentuan label dari tiap topik. Berdasarkan frekuensi tiap topik pada sebuah Dokumen didapatkan probabilitas topik dari Dokumen tersebut, dan topik yang memiliki probabilitas tertinggi dianggap sebagai *main topic* dari Dokumen tersebut.

Tabel 2.5. Probabilitas Topik.

Topik	Dokumen
	1
1	60%

2	25%
3	15%

Dari tabel 2.5 dapat dilihat bahwa probabilitas topik 1 adalah probabilitas yang paling besar.

Kemudian setelah proses LDA diterapkan pada semua dokumen, maka semua dokumen yang telah diproses dengan LDA akan diurutkan berdasarkan probabilitas topik yang paling tinggi.

Gambar 2.6. Urutan Data

Dokumen	Topik	Probabilitas
1	1	60%
3	2	25%
2	3	15%

Dari tabel diatas dapat dijelaskan bahwa Dokumen 1 dengan topik 1 berada pada urutan pertama dengan probabilitas 65%, kemudian diurutkan kedua ada Dokumen 3 dengan topic 2 dan probabilitas 25% dan diurutkan terakhir ada Dokumen 2 dengan topik 3 dengan probabilitas 15%.

III. HASIL DAN KESIMPULAN

3.1 Implementasi sistem

3.1.1 Tampilan Masuk Sistem

Tampilan *masuk sistem* adalah form pertama yang muncul ketika aplikasi di jalankan.

Tampilan form masuk sistem dapat dilihat pada gambar 3.2.



Gambar 3.2 Tampilan Masuk Sistem

Pada gambar 3.2 menampilkan *field* untuk mengisi *username* dan *password* yang nantinya untuk masuk ke form wartawan ataupun redaktur.

3.2.2 Tampilan Beranda

Tampilan beranda adalah form pertama yang muncul setelah masuk aplikasi. Tampilan form beranda dapat dilihat pada gambar 3.3.



Gambar 3.2 Tampilan Beranda

Beranda

Pada gambar 3.2 menampilkan informasi tentang Visi Misi perusahaan.

3.2.3 Tampilan Input Berita

Tampilan input berita adalah form yang muncul setelah memilih menu input

berita pada halaman wartawan. Tampilan form input berita dapat dilihat pada gambar 3.4.



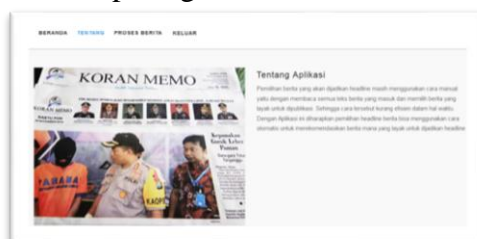
Gambar 3.4 Tampilan Input Berita

Pada gambar 3.4 menampilkan *field* untuk mengisi judul berita, isi berita dan jenis berita yang nantinya akan dikirim kedalam database setelah menekan tombol kirim.

Pengacuan Pustaka.

3.2.4 Tampilan Tentang

Tampilan tentang adalah form yang akan keluar ketika memilih menu tentang setelah masuk sistem. Tampilan tentang terdapat pada halaman wartawan maupun halaman redaktur. Tampilan halaman tentang dapat dilihat pada gambar 3.4.



Tentang

Pada gambar 3.4 menampilkan informasi sekilas tentang aplikasi.

3.2.5 Tampilan Redaktur

Tampilan redaktur adalah form yang dapat diakses dari form masuk sistem. Form ini diakses dengan cara mengisi username redaktur dan mengisi *password*. Tampilan halaman proses berita dapat dilihat pada gambar 3.5



Gambar 3.5 Tampilan Proses Berita

Pada gambar 3.5 menampilkan informasi yang berisi tentang bobot dokumen, jenis dokumen, hasil proses aplikasi dan *delete* dokumen.

4. SIMPULAN

Dari hasil penelitian yang berjudul “Implementasi Metode Lathent Dirichlet Allocation Untuk Menentukan Topik Teks Berita” ini dapat disimpulkan sebagai berikut:

1. Metode *lathent dirichlet allocation* dapat digunakan untuk menentukan topik berita.
2. Uji coba manual penentuan headline berita memiliki hasil yang sama dengan perhitungan pada sistem yang menggunakan metode *lathent dirichlet allocation*, maka dihasilkan urutan rekomendasi berita yang dijadikan headline sesuai dengan bobotnya

5.SARAN

Dari hasil penelitian yang berjudul “Implementasi Metode *Lathent Dirichlet Allocation* Untuk Menentukan Topik Teks Berita” ini didapat saran untuk penelitian selanjutnya sebagai berikut :

1. Penelitian dilakukan dengan objek atau studi kasus yang berbeda dari teks berita.
2. Penelitian selanjutnya dapat menggunakan kecerdasan buatan lainnya dengan metode yang berbeda agar dapat membandingkan tingkat akurasi dengan metode *latent dirichlet allocation*.
3. Menambah fungsi muat data secara dinamis agar memudahkan pengguna dan untuk mengirim dokumen.

DAFTAR PUSTAKA

- [1] Berry, M.W. 2010. *Text Mining Application and Theory*. WILEY : United Kingdom
- [2] Daniel,S.2014.*Multi-Document Summarization Based On Sentence Clustering Improved Using Topic Words*. JUTI: Jurnal Ilmiah Teknologi Informasi - Volume 12, Number 2, July 2014.
- [3] Dragut, E. 2009. *Stop Word and Related Problems in Web Interface Integration*.
- [4] Fadillah, Z. 2003. *A Study of Stemming Effects on Information Retrieval in Bahasa Indonesia*. Institute for Logic, Language and Computation Universiteit van Amsterdam The Netherland.
- [5] Feldman,R. 2007. *The Text Mining Handbook : Advance Approaches in Analyzing Unstructured Data*. Cambridge University Press : New York.
- [6] Kadir, Abdul. 2008. *Tuntunan Praktis: Belajar Database Menggunakan MySQL*.
- [7] Mustaqhfiri, M. 2014. *Peringkasan teks otomatis berita berbahasa indonesia menggunakan metode maximum marginal relevance* Jurusan

- Teknik Informatika, Fakultas Sains dan Teknologi Universitas Islam Negeri Maulana Malik Ibrahim Malang.
- [8] Nugroho, Bunafit. 2005. *Database Relasional dengan MySQL*. Yogyakarta: Andi.
- [9] Peranganing, Kasiman. 2006. *Aplikasi WEB dengan PHP dan MySQL*. Yogyakarta: Andi.
- [10] Prihatini. M.N. 2017. *Metode Latent Dirichlet Allocation untuk ekstraksi Topik dokumen*. Politeknik Negeri Bali
- [11] Sidik, Betha. 2014. *Pemrograman Web dengan PHP*. Bandung: Informatika.
- [12] Sukamto, Rosa A., Salahuddin, M. 2016. *Rekayasa Perangkat Lunak Terstruktur Dan Berorientasi Objek*. Bandung: Informatika.
- [13] Weiss, S.M. 2005. *Text Mining : Predictive Methods fo Analyzing Unstructured Information*. Springer : New York.
- [14] Zulhanif. 2016. *Pemodelan topik dengan Latent Dirichlet Allocation*. Departemen Staistika FMIPA UNPAD
- [15] Fitriyadi, K, Sutikno. 2016. *Pengenalan Jenis Golongan Darah Menggunakan Jaringan Syaraf Tiruan Perceptron*, Jurnal Masyarakat Informatika, Vol 7, No 1, hh 700-710
- [16] Sunarya, I M G. 2013. *Sistem Biometrika Tandatangan MEnggunakan Jaringan Syaraf Tiruan Model Perceptron*, Jurnal Masyarakat Informatika, Vol 7, No 1, hh 1-10.