KLASIFIKASI WEBSITE MENGGUNAKAN ALGORITMA MULTILAYER PERCEPTRON

Nyoman Purnama, I Ketut Gede Darma Putra*, Putu Agung Bayupati*
Teknik Elektro, * Teknologi Informasi, Fakultas Teknik, Universitas Udayana
Kampus Bukit Jimbaran, Bali, 80361
pur182@yahoo.com.

Abstrak

Sistem klasifikasi merupakan proses temu balik informasi yang sangat bergantung dari elemen-elemen penyusunnya. Sistem ini banyak digunakan untuk mengatasi permasalahan segmentasi data. Klasifikasi dapat digunakan pada *website* sebagai metode untuk mengelompokkan website. *Website* merupakan salah satu data yang memiliki informasi yang beraneka-ragam, sehingga pengelompokan data ini penting untuk diteliti. Sistem klasifikasi dimulai dengan melakukan proses pengumpulan informasi dari halaman *website (parsing)* dan untuk setiap hasil *parsing* dilakukan proses penghapusan kata henti, *stemming*, *feature selection* dengan tf-idf. Hasil dari proses ini berupa fitur yang menjadi inputan algoritma *Multilayer Perceptron*. Dalam algoritma ini terjadi proses pembelajaran terhadap pola input masukan dan pembuatan bobot pelatihan. Bobot ini akan digunakan pada proses klasifikasi. Hasil dari penelitian menunjukkan bahwa algoritma *Multilayer Perceptron* dapat menghasilkan klasifikasi *website* dengan akurasi yang bagus. Hal ini dibuktikan dengan beberapa tahapan penelitian yang berbeda dan didapatkan nilai akurasi rata-rata diatas 70%.

Kata Kunci : *Multilayer perceptron, website, pelatihan, bobot, stemming, klasifikasi, feature selection*

1. PENDAHULUAN

Klasifikasi *website* memegang peran yang cukup vital dalam berbagai tugas manajemen dan proses untuk mendapatkan informasi [1]. Klasifikasi penting dalam hal : penelusuran *website* oleh mesin pencarian, membantu pengembangan direktori berbasis *website* dan juga untuk meningkatkan kualitas dari pencarian *website*. Pada pengembangan *website* berbasis direktori, kegiatan pemeliharaan direktori dilakukan secara manual. Oleh karena itu sebuah sistem yang dapat mengerjakan proses klasifikasi secara otomatis, merupakan sebuah pengembangan yang penting [2]. Dengan adanya klasifikasi, memungkinkan proses penelusuran (*crawling*) menjadi lebih fokus dan sesuai dengan apa yang diharapkan pengguna.

Perkembangan teknologi Internet dan kebutuhan akan layanan informasi yang kian tinggi, memunculkan berbagai *website* baru. Keberadaan *website-website* ini berguna untuk mendapatkan informasi yang diinginkan dengan lebih cepat dan mudah. Namun yang sering menjadi kendala adalah bagaimana informasi bisa didapatkan dengan cepat dan relevan sesuai keinginan pengguna. Proses temu balik informasi (*Information retrieval*) dari sebuah halaman *website* dan pengkategoriannya, menjadi suatu hal yang amat berguna bagi pengguna Internet. Dari sini kemudian istilah *web mining* menjadi cukup populer dikalangan peneliti maupun praktisi [3].

Web mining merupakan aplikasi dari *data mining* yang berfungsi untuk mengekstrak pengetahuan /informasi dari halaman web, seperti dokumen web, *hyperlink* antar dokumen, log dari *website* dan lain-lain. *Web mining* [4] memiliki tiga topik yang menarik yakni pengelompokkan secara alami *(Clustering)*, klasifikasi dan analisa sekuensial. Area *clustering* dan klasifikasi memiliki peran yang hampir sama, yakni untuk mengelompokan *website* sesuai kategorinya. Namun dalam proses pengelompokannya klasifikasi bersifat pembelajaran terbimbing (*supervised*) sementara *clustering* bersifat tak terbimbing (*unsupervised*).

Klasifikasi dokumen *website* berbeda dengan klasifikasi teks biasa [5]. Dimana w*ebsite* terdiri dari struktur yang kompleks, yang didalamnya terdapat tag-tag HTML, tautan dan *javascript*. Tautan yang terdapat didalam *website* memungkinkan untuk mendapatkan informasi yang lebih banyak, dalam proses klasifikasi. Dari informasi yang didapatkan, kemudian semua proses pada teks *mining* [6] seperti pencarian kata dasar sebuah kata (*stemming*), penghapusan kata henti (*stop words*) dan *Feature Selection* dapat dilakukan. Kata-kata hasil *stemming* dan *stop words* kemudian dinamakan dengan "fitur". *Feature Selection* berfungsi untuk menghilangkan *noise* yang terdapat pada fitur yang diolah. Sehingga hasil klasifikasi yang didapatkan akan lebih akurat dan efisien, walaupun terjadi pengurangan fitur yang digunakan [7]. Ada beberapa teknik *feature selection* yakni *document frequency tresholding* (df), gabungan antara *term frequency* dan *inverse document frequency* (tf-idf), *information gain* (ig) dan *mutual information* (mi). Pada penelitian ini digunakan metode *term frequency* dan *inverse document frequency* (tf-idf) karena berdasarkan beberapa penelitian metode ini lebih handal untuk mengurangi dimensi/ukuran atribut dari sebuah dokumen sehingga menghasilkan hasil klasifikasi yang lebih akurat [8]. Tugas dasar dari klasifikasi *website* adalah untuk mengetahui topik utama dari sebuah *website*. Topik ini dapat diketahui secara umum melalui halaman depan atau *homepage* dari sebuah *website* [5]. Halaman depan merupakan jalan masuk utama

untuk mengetahui keseluruhan isi *website*. Selain itu kegunaan dari klasifikasi *website*, adalah untuk mengelompokkan *website* secara otomatis. Salah satu contoh pengelompokan *website* berdasarkan subjek yakni direktori "*all business directory" (http://www.allbusinessdirectory.biz*). Direktori ini merupakan direktori klasifikasi *website* yang banyak digunakan bagi para *webmaster* untuk meningkatkan trafik. *Website* di-*input*-kan secara manual oleh pemilik *website* dan diberikan kategori sesuai dengan tema dari *website* itu. Ada 18 kategori dalam direktori *all business directory*. Masing-masing kategori memiliki subkategori, yang membentuk hirarki dari kategori *website*. Semua *website* dalam direktori ini berbahasa Inggris. Tidak seperti direktori lainnya yang menyediakan kategori khusus untuk bahasa yang berbeda. Jumlah *website* dalam masing-masing kategori pun berbeda-beda.

Ada beberapa metode klasifikasi yang umum digunakan pada proses klasifikasi berbasis teks. Diantaranya yakni *C45, Naïve bayes, K-Nearest Neighbor, K-Means, SVM, Neural Networks*, Algoritma Genetika dan *Multi layer Perceptron*. Pada penelitian ini digunakan algoritma *Multi layer Perceptron* (MLP). MLP merupakan algoritma yang mengadopsi cara kerja saraf pada makhluk hidup. Algoritma ini handal karena dalam proses pembelajarannya dilakukan secara terarah. Pembelajaran dilakukan dengan jalan memperbaharui bobot balik *(backpropagation*). Dengan bobot yang optimal maka klasifikasi yang dihasilkan lebih baik. MLP sebagai *classifier* dibantu dengan metode *feature selection* yang tepat dapat meningkatkan tingkat akurasi dari klasifikasi itu sendiri, hal ini dikarenakan adanya proses perbaikan bobot untuk menjadi lebih baik.

Pada hasil eksperimen [8] untuk klasifikasi teks dengan menggunakan metode MLP *backpropagation* didapatkan bahwa pembelajaran menggunakan *backpropagation* menghasilkan performa yang bagus. Pada penelitian ini juga dilakukan pengujian dalam mengurangi dimensi dari input pembelajaran dengan metode *feature selection* tf-idf dan didapatkan hasil sebesar 78,8 %. Hasil eksperimen lainnya dilakukan oleh G. Dhaneswara & V.S. Moertini, 2004 dimana pemilihan konfigurasi jaringan (jumlah lapis tersembunyi, neuron, momentum dan *learning rate*) amat diperlukan dalam proses pelatihan dimana konfigurasinya bisa berbeda-beda dari satu set data pelatihan yang lain, sehingga diperlukan eksperimen dalam mencarinya. Disini juga dijelaskan bahwa penggunaan jaringan *backpropagation* memiliki masalah pada lamanya waktu pelatihan. Penelitian selanjutnya mengenai *backpropagation* untuk prediksi penyakit paru, dilakukan oleh Novi Indah Pradasari dkk. didapatkan nilai akurasi yang cukup tinggi sebesar 91,66% dengan konfigurasi jumlah data latih 96 buah, 24 data uji dan 2 buah *hidden layer*. Berdasarkan beberapa penelitian yang ada, maka pada penelitian ini telah dilakukan percobaan penggunaan algoritma MLP dengan *backpropagation* untuk mengklasifikasikan keanekaragaman dimensi input dalam halaman *website* yang ada di Internet, dengan perubahan beberapa konfigurasi jaringan.

2. METODOLOGI

A. Gambaran Umum Sistem

Pada penelitian ini proses yang dilakukan dibagi menjadi 2 tahapan utama yakni proses pelatihan dan proses klasifikasi.

Proses pelatihan adalah proses pembelajaran yang dikerjakan secara *offline*. Data penelitian yang digunakan adalah berupa tautan dari halaman utama *website*. Tautan *website* dikumpulkan terlebih dahulu dari direktori *"All Business Directory (ABD)"* dengan alamat *http://www.allbusinessdirectory.biz*. Tautan yang telah dikumpulkan akan diseleksi dan diunduh pada bagian halaman utamanya.

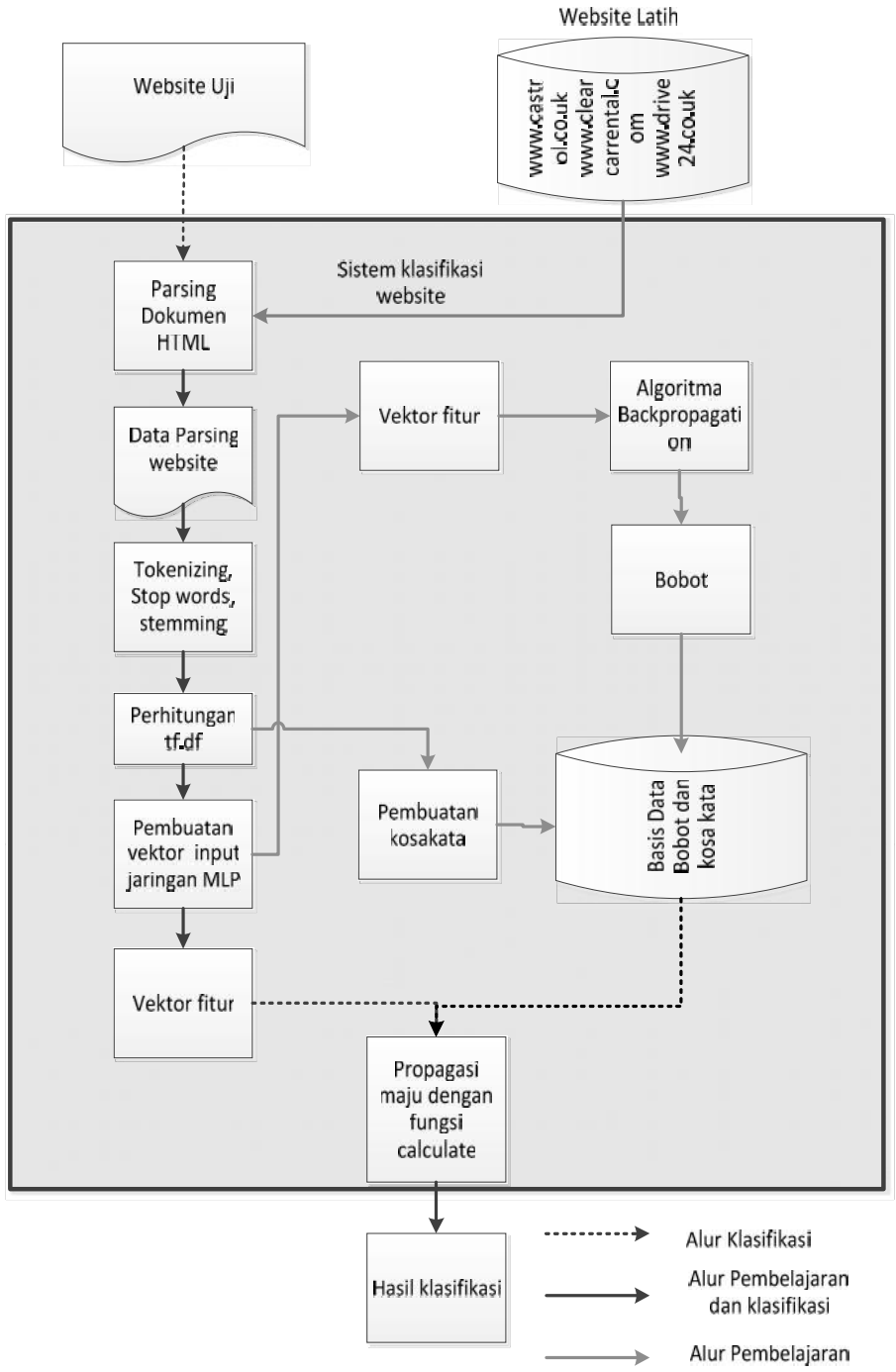
Sedangkan proses klasifikasi adalah proses untuk mengkategorikan sebuah alamat *website* atau *Universal Resource Locator* (URL) sesuai kategorinya secara otomatis. URL yang digunakan dalam proses klasifikasi adalah URL yang terdapat pada direktori ABD, selain yang digunakan pada proses pembelajaran. Proses klasifikasi dilakukan secara *online* dengan mengunduh langsung *website* yang akan diuji. Adapun gambaran umum rancangan diperlihatkan pada gambar 1.

Direktori *Allbussinessdirectory* (ABD) memiliki 18 kategori. Pada masing-masing kategori terdapat kategori turunan dan seterusnya, ke 18 kategori itu yakni *Automotive, business & economy, careers & jobs, computers, education & , entertainment & media* dengan*, health & beauty care, industry, Internet & www, law, real estate, science, shopping & services, small business, society, sports, telecommunications, travel & recreation.* Pada penelitian ini dipilih 6 kategori saja yakni *Automotive, computers, education & , health & beauty care, sports, travel & recreation.* Semua *website* dalam ABD adalah berbahasa Inggris. URL dimasukkan kedalam direktori ABD oleh pengguna dan diberikan kategori secara manual. Dengan adanya program berbasis JST ini diharapkan dapat mengklasifikasikan *website* secara otomatis berdasarkan pengetahuan yang dimilikinya.

B. Data penelitian

Proses untuk mendapatkan informasi pada dunia Internet (*web mining*) berbeda dengan proses untuk mendapatkan informasi yang berbasis teks biasa (*text mining*). Keanekaragaman fitur yang ada pada *website* memberikan tantangan tersendiri jika dibandingkan dengan *text mining*. Dimana suatu halaman web tidak hanya berisikan data berupa teks saja, melainkan juga berupa informasi HTML tag seperti <TITLE>, <BODY>, <H1>, <META> dan lain-lain.

..



Gambar 1. Gambaran Umum Sistem

..

C. Data penelitian

Proses untuk mendapatkan informasi pada dunia Internet (*web mining*) berbeda dengan proses untuk mendapatkan informasi yang berbasis teks biasa (*text mining*). Keanekaragaman fitur yang ada pada *website* memberikan tantangan tersendiri jika dibandingkan dengan *text mining*. Dimana suatu halaman web tidak hanya berisikan data berupa teks saja, melainkan juga berupa informasi HTML tag seperti <TITLE>, <BODY>, <H1>, <META> dan lain-lain.

Tidak semua bagian dari halaman web perlu digunakan dalam penelitian ini, beberapa tag dalam kode HTML suatu *website* tidak terlalu penting, maka isi dari tag tersebut dapat diabaikan. Tag yang umumnya berisikan informasi yang dapat digunakan

dalam proses pelatihan dan pengujian yakni tag <TITLE> dan tag <BODY> (tag lainnya bisa dilihat pada tabel di bawah Elemen/tag <TITLE> berisikan poin penting dari sebuah halaman *website*. Maka dari itu teks yang berada pada tag <TITLE> akan digunakan sebagai acuan nantinya. Sementara tag <BODY> merupakan penjabaran dari poin-poin yang terdapat pada tag <TITLE> atau <META>. Tag <BODY> juga memilki beberapa tag lain didalamnya yang dapat diabaikan dalam penelitian ini seperti : tag <LINK>, <IMG>, <SCRIPT> dll.

Jumlah *website* yang terdaftar dalam direktori ABD, pada saat pembuatan penelitian ini sebesar 20,022 *website*. Jumlah *website* yang akan digunakan pada proses pelatihan akan divariasikan mulai dari 10, 20 dan 30. Sehingga nantinya dapat dilihat nilai kesalahan yang terjadi dengan penambahan jumlah dokumen latih. Begitu juga dengan data uji klasifikasi akan divariasikan jumlahnya yakni 100, 150, 200 dan 250. Pada proses pelatihan dokumen akan diambil secara acak pada masing-masing kategori.

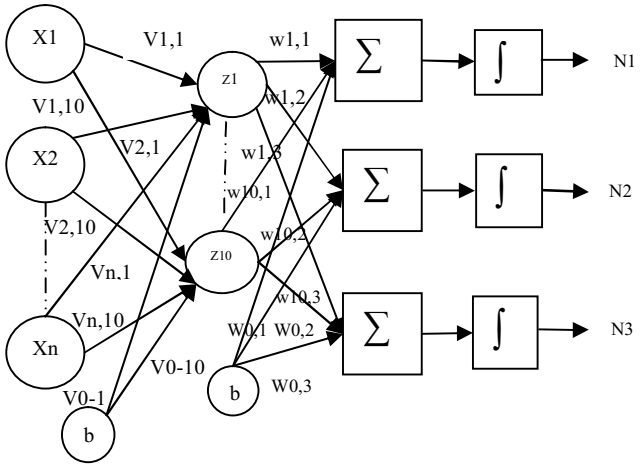
Satu persatu tautan yang ditampilkan pada halaman direktori ini akan diseleksi dengan kriteria berikut:

a. Tautan datang dari halaman pertama (*home page*)
b. *Homepage* yang tidak terdapat elemen flash atau aplikasi lainnya
c. *Homepage* berbahasa Inggris dengan jumlah kata yang cukup digunakan dalam proses klasifikasi. Dalam penelitian ini digunakan jumlah kata mulai dari 100 kata.

Tautan tidak datang dari *website* berupa portal seperti *yahoo.com* dll.

D. Arsitektur jaringan Multi Layer Perceptron

Arsitektur jaringan Neural network yang akan dibuat adalah 3 *layer feed forward network*. Jaringan Neural 3 layer *feed forward network*, terdiri dari layer input, layer *hidden* dan layer output. Nilai masukan jaringan neural berupa vektor, dimana masing masing dokumen memiliki jumlah vektor yang berbeda. Fungsi aktivasi yang akan digunakan pada bagian input ke *hidden layer* dan dari *hidden layer* ke output layer yakni fungsi *aktivasi hyperbolic tangent.* Fungsi ini menurut beberapa sumber di Internet merupakan fungsi aktivasi yang dapat mempercepat konvergensi dalam jaringan *Neural network.* Selain itu memiliki iterasi lebih sedikit dalam proses pembelajaran. Nilai yang akan dihasilkan dari fungsi aktivasi *hyperbolic tangent* yaitu -1 dan 1. Sehingga untuk mengklasifikasikan 6 kategori diperlukan 3 output untuk membedakannya. Ketiga output tadi akan menghasilkan kombinasi yang merepresentasikan masing-masing kategori. Mengenai nilai kombinasi angka yang setara dengan masing-masing kategori, tidak terdapat keterikatan. Pada penelitian ini kategori *Automotive* dilambangkan dengan kombinasi output *[1,1,1], computers* dengan kombinasi *output [1,1,-1], education & training* [*1,-1,-1], health & beauty care* [*-1,1,1], *sports* dengan kombinasi *[-1,1,-1], travel & recreation* dengan kombinasi *[-1,-1,1]*



Gambar 2. Arsitektur jaringan Neural Network yang digunakan

Gambar 2 di atas adalah arsitektur jaringan MLP yang digunakan pada penelitian ini. Disini terlihat nilai v1,1 sampai vn,10 yang merupakan bobot pada lapisan tersembunyi. Dengan angka sebelum tanda koma (“,“) menunjukkan neuron yang akan menerima bobot Dan nilai setelah tanda koma (“,“) menunjukkan masukan nilai bobot pada layer tersembunyi. Vij adalah bobot yang dikalikan unit masukan (xi). Nilai x1-xn merupakan inputan dari jaringan MLP dimana fitur yang akan digunakan merupakan hasil seleksi fitur yang memiliki nilai tf-idf tertinggi, berdasarkan tabel pengetahuan yang telah dibuat sebelumnya. Lapisan tersembunyi (*hidden layer*) dilambangkan dengan huruf z1 sampai z10. Jumlah unit pada layer ini akan dirubah nantinya, sampai menghasilkan keluaran pembelajaran yang sesuai.

Penentuan nilai parameter algoritma MLP nantinya dapat menentukan keberhasilan dari pembelajaran. Pada penelitian ini digunakan inisialisasi nilai parameter MLP sebagai berikut **:** epoh =1000, *learning rate* =0,5, momentum = 0,6 dan *Minimum Square Error*(MSE) =0,01 serta bobot awal secara random dengan rentang nilai antara -0,25 sampai 0,25.

Proses pelatihan harus dilakukan sebelum bisa digunakan untuk pengujian. Selama proses pelatihan, bobot koneksi dari jaringan MLP akan di perbaiki terus sehingga kesalahan pembelajaran berada dibawah ambang toleransi. Dari proses pelatihan akan didapatkan nilai bobot akhir yang akan digunakan pada proses pengujianProses yang dilakukan pada bagian pengujian hampir sama dengan proses pembelajaran, hanya pada proses pengujian tidak dilakukan pembelajaran lagi karena bobot yang digunakan adalah bobot hasil pelatihan. Dimana data yang dimasukkan berupa *link website* yang belum diketahui kategorinya. Kemudian proses yang sama dilakukan seperti pada proses

pembelajaran yakni *parsing, tokenizing, stopping* dan *stemming*. Namun pada proses klasifikasi dilakukan penyaringan kata hasil *stemming*, dengan daftar pengetahuan kata yang didapatkan pada proses pembelajaran sebelumnya. Kosakata/fitur yang tidak termasuk didalam daftar pengetahuan kata, akan dihilangkan. Hasil dari penyaringan ini kemudian dihitung bobotnya, dan menjadi inputan jaringan *Multi Layer Perceptron.*

Hasil output proses klasifikasi dengan fungsi aktivasi *hyperbolic tangent* yang memiliki nilai berupa desimal. Ketiga nilai output dari unit keluaran jaringan syaraf tiruan ini akan dirubah menjadi bentuk 1 dan -1. Dimana nilai unit output yang berupa nilai desimal dibulatkan ke bilangan terdekatnya. Nilai unit > 0 akan dikonversikan dengan nilai 1 dan nilai unit output < 0 dengan nilai -1. Jika proses pengujian berhasil maka akan dilakukan proses uji jaringan sampai ditemukan nilai peramalan yang akurasinya tinggi dan kesalahannya rendah. Jika uji jaringan tidak berhasil maka proses pelatihan diulangi lagi dengan nilai parameter yang berbeda.

3. HASIL DAN PEMBAHASAN

A. Implementasi metode MLP

Proses pelatihan dengan metode MLP dilakukan untuk mendapatkan bobot pembelajaran dari data yang dimasukkan. Berdasarkan nilai tf-idf masing masing kata dalam tabel pengetahuan pada masing masing dokumen, akan dicari pola pembelajarannya. Sehingga menghasilkan kategori yang sesuai untuk *website* yang dilatih. Dalam proses pelatihan dapat diatur nilai epoh, mse, momentum dan *learning rate*. Perubahan bobot dilakukan melalui beberapa kali iterasi.

Data yang digunakan dalam pelatihan adalah data *website* yang telah diunduh secara manual dari direktori *website All Business Directory.* Data ini berupa file index.html (halaman depan dari *website*). Kumpulan file tadi dikelompokkan pada direktori dengan nama sesuai kategori masing-masing. *Website* dipilih sesuai dengan kriteria yang telah ditentukan sebelumnya. *Website* diambil secara acak pada masing-masing kategori. Hasil *parsing* dari masing masing *website* disimpan kedalam database *mysql* dengan nama tabel *tb_berita*. Sedangkan daftar fitur/kosakata yang didapatkan dari masing masing *website* disimpan kedalam database dengan nama tabel *tb_index*. Bagian dari *website* yang digunakan adalah gabungan dari semua isi dari tag HTML baik pada bagian *Title, body* dan META. Tidak ada *website* yang sama digunakan pada proses pengujian ataupun pelatihan.

Tabel pengetahuan adalah daftar kosakata yang dikumpulkan pada proses tf-idf. Dari setiap dokumen dianalisa kosakatanya dengan menentukan nilai tf-idf dari masing masing kata yang ditemukan dalam teks. Untuk mengurangi jumlah kosakata yang digunakan sebagai input jaringan syaraf tiruan maka dari semua kumpulan vektor dalam sebuah dokumen, diambil 30 kosakata dalam vektor dengan nilai tf-idf terbesar seperti yang telah diatur pada halaman *setting*. Semua kosakata yang dikumpulkan adalah unik. Tabel pengetahuan akan digunakan juga pada proses klasifikasi. Namun pada proses klasifikasi tidak ada proses penyusunan table pengetahuan. Kosakata hasil *stemming* pada *website* uji akan dibandingkan dengan kosakata pada tabel pengetahuan. Kosakata yang tidak terdapat pada tabel pengetahuan akan diabaikan. Dari kosakata yang tersisa akan dihitung nilai tf-idf nya dan menjadi inputan jaringan syaraf tiruan.

Proses pelatihan dengan metode MLP dilakukan untuk mendapatkan bobot pembelajaran dari data yang dimasukkan. Berdasarkan nilai tf-idf masing masing kata dalam tabel pengetahuan pada masing masing dokumen, akan dicari pola pembelajarannya. Sehingga menghasilkan kategori yang sesuai untuk *website* yang dilatih. Dalam proses pelatihan dapat diatur nilai epoh, mse, momentum dan *learning rate*. Perubahan bobot dilakukan melalui beberapa kali iterasi. Proses pelatihan diawali dengan proses index *website* pelatihan meliputi *tokenisasi, stop words, stemming* dan pembobotan dengan tf-idf. Setelah itu akan didapatkan tabel pengetahuan yang akan digunakan sebagai input pelatihan. Proses pembuatan input meliputi pengambilan nilai tf-idf pada dokumen yang akan dilatih, dengan kosakata yang sesuai dengan tabel pengetahuan. Jika dalam suatu dokumen tidak ditemukan kata pada tabel pengetahuan maka nilai input adalah 0, sebaliknya jika dalam *website* latih terdapat kata pada tabel pengetahuan maka nilainya sesuai dengan nilai bobot pada dokumen tersebut.

Pelatihan jaringan akan berhenti jika telah mendapatkan error yang lebih kecil dari target penelitian, yang dinamakan MSE (*Mean Squared Error*). Jika *error* tidak terpenuhi maka jaringan akan berhenti pada maksimum iterasi (epoh) yang dimasukkan. Setelah melewati proses pelatihan, maka bobot yang dihasilkan pada proses pelatihan akan menjadi bobot pada proses pengujian. Proses pengujian/klasifikasi memiliki urutan yang sama dengan proses pelatihan. Hanya saja disini tidak terdapat proses perbaikan bobot dengan jalan *backpropagation*. Proses diawali dengan *parsing* halaman HTML *website* yang diuji. Kemudian proses *tokenizing, stop words, stemming* dan perhitungan tf-idf juga dilakukan. Proses klasifikasi dilakukan dengan mengambil input kosakata dari tabel pengetahuan, kemudian dibandingkan dengan kosakata pada dokumen yang diuji. Jika didapatkan kosakata yang sama maka nilai tf-idf dari kosakata dokumen uji akan digunakan, jika tidak maka nilai kosakata pada tabel pengetahuan adalah 0. Proses pelatihan diperlihatkan pada gambar 3.

Gambar 3. Tampilan proses pelatihan

B. Pengujian

Analisis uji klasifikasi *website* dilakukan dengan membandingkan hasil klasifikasi yang diberikan oleh sistem dengan klasifikasi yang dilakukan secara manual. Data uji yang digunakan adalah data *website* yang ada dalam direktori "*All Business Directory*" selain yang digunakan pada proses pelatihan. Kategori yang difokuskan dalam penelitian ini sejumlah 6 kategori yakni *Automotive, sports, education and training, Health and beauty, computers dan travel & recreation*.

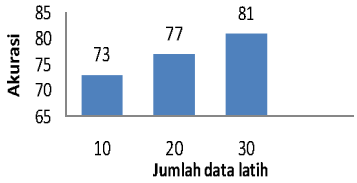
Hasil pengujian dibedakan menjadi 3 bagian yakni pengamatan terhadap perubahan jumlah data latih, pengamatan terhadap perubahan jumlah unit *hidden layer* dan perhitungan jumlah waktu yang diperlukan dengan berubahnya jumlah data latih, jumlah unit *hidden layer* dan jumlah data uji Hasil pengamatan kemudian dilakukan terhadap 250 data uji akan dibagi menjadi 4 bagian penelitian yang berbeda. Pengujian dilakukan dengan menggunakan metode akurasi, presisi dan *recall*.

Pada penelitian ini akan dilakukan pengamatan terhadap perubahan jumlah data latih yang digunakan pada proses pelatihan. Percobaan pertama data latih akan diambil sebanyak 10 buah *website*/kategori secara acak dari direktori ABD, sehingga terdapat total 60 buah *website* sebagai input pelatihan jaringan *Multi Layer*. Percobaan kedua akan diambil 20 buah *website* per kategori secara acak sebagai data latih. Begitu juga percobaan berikutnya diambil 30 *website* per kategori sehingga menghasilkan total 180 data latih dalam proses pelatihan. Dengan semakin banyaknya jumlah data latih, nilai akurasi yang diperoleh lebih baik. Pada pengujian dengan jumlah data latih perkategori 30 buah didapatkan nilai akurasi paling tinggi, yakni sebesar 81%. Oleh karena itu pada percobaan – percobaan selanjutnya digunakan data latih sebanyak 30 buah per kategori. Pengamatan untuk data latih 30 buah dirangkum dalam tabel 1 dan digambarkan dalam bentuk grafik pada gambar 4.

Berdasarkan pengujian ditunjukkan bahwa data dengan tema *computer* memiliki nilai akurasi paling besar dibanding lainnya. Data uji dengan *tema Health & beauty care* memiliki nilai akurasi yang cenderung tetap pada jumlah data latih yang berbeda. Sedangkan kategori *Sports* memiliki akurasi paling buruk diantara yang lainnya. Hal itu disebabkan karena kosakata pada *website computer* memiliki pembeda yang lebih besar dibanding kategori lainnya. Sementara kosakata pada kategori *sports* kemungkinan terdapat pada kategori lainnya. Tinggi rendahnya nilai akurasi disini tergantung dari data latih yang digunakan pada masing-masing kategori.

Tabel 1. Hasil Pengujian 1

No	Tema data uji	Akurasi	*Recall*	Presisi
1	Automotive	71%	76%	76%
2	Health & Beauty Care	65%	65%	100%
3	Education & Training	88%	88%	94%
4	Computers	94%	94%	88%
5	Sports	82%	82%	70%
6	Travel & Recreation	88%	88%	94%



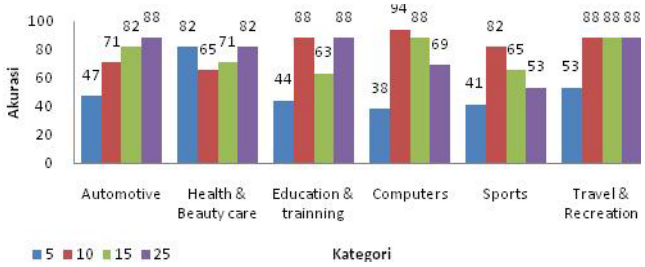
Gambar 4. Grafik Pengujian 1

Hasil pengujian selanjutnya Jumlah unit/neuron pada *hidden layer* akan dirubah-ubah sehingga menghasilkan tingkat akurasi yang baik Berdasarkan pengamatan dari peningkatan jumlah unit/neuron pada *hidden layer* didapatkan bahwa nilai akurasi tertinggi terjadi pada jumlah unit 10 buah. Sedangkan nilai akurasi terendah didapatkan pada jumlah unit 5 buah. Dari data tersebut di atas nilai *hidden layer* yang kecil mengakibatkan terjadinya *underfitting* atau jaringan menjadi terlalu sederhana, sehingga banyak informasi yang belum diproses. Oleh karena itu akurasi pada jumlah unit *hidden* 5 kurang baik. Sementara jumlah *hidden* unit di atas 10 dan seterusnya tidak begitu berpengaruh terhadap nilai akurasi yang dihasilkan. Namun jumlah unit hidden yang terlalu banyak dapat mengakibatkan *overfitting* yaitu kapasitas pengolah yang terlalu banyak sehingga ada *hidden* unit yang tidak ikut melatih dan menjadi *noise* (gangguan data). Sehingga akurasi yang dihasilkan akan menurun. Secara umum pada masing-masing kategori terjadi kenaikan akurasi sesuai dengan teori yang dijelaskan sebelumnya. Sementara untuk kategori *Sports* dan *computer* terjadi *overfitting* dimana kapasitas pengolahan lebih banyak dari informasi yang diproses, sehingga banyak *noise* yang terjadi. Dan hasil yang diperoleh untuk jumlah unit *hidden layer* 10 dirangkum dalam

tabel 2. Secara keseluruhan penelitian untuk jumlah unit hidden yang bervariasi digambarkan pada gambar 5.

Tabel 2. Hasil Pengujian 2

No	Tema data uji	Akurasi	*Recall*	Presisi
1	*Automotive*	71%	76%	76%
2	*Health & Beauty Care*	65%	65%	100%
3	*Education & Training*	88%	88%	94%
4	*Computers*	94%	94%	88%
5	*Sports*	82%	82%	70%
6	Travel & Recreation	88%	88%	94%



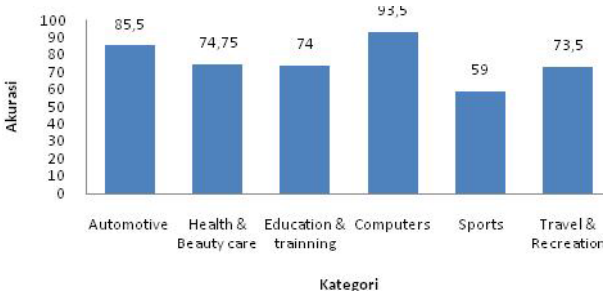
Gambar 5. Grafik Pengujian 2

Untuk mengetahui hubungan antara jumlah data terhadap nilai akurasi, *recall* dan presisi maka jumlah data ditambahkan secara bertahap, jumlah data yang akan digunakan yakni 100, 150, 200 dan 250 buah data. Berikut tabel 3 ditampilkan hasil pengujian terhadap jumlah data uji 100 buah.

Tabel 3. Hasil Pengujian 3

No	Tema data uji	Akurasi	*Recall*	Presisi
1	*Automotive*	71%	76%	76%
2	*Health & Beauty Care*	65%	65%	100%
3	*Education & Training*	88%	88%	94%
4	*Computers*	94%	94%	88%
5	*Sports*	82%	82%	70%
6	*Travel & Recreation*	88%	88%	94%

Berdasarkan pengujian yang telah dilakukan diatas, maka dapat digambarkan dalam bentuk grafik pada gambar 7. Secara keseluruhan dapat disimpulkan bahwa perubahan jumlah data uji mengakibatkan nilai akurasi yang tidak menentu. Hal ini dapat disimpulkan dimana pada saat data pengujian sebanyak 100 buah, data uji yang digunakan memiliki *noise* masih sedikit sehingga tidak terjadi *underfitting / overfitting*. Sementara dengan meningkatnya jumlah data uji mengakibatkan *noise* yang dihasilkan lebih banyak, sehingga diperlukan lebih banyak data latih kembali. Nilai akurasi yang dihasilkan dengan data uji diatas 100 buah berkisar 70%. Pada kategori *Automotive* dan *Computers* memiliki tingkat akurasi tinggi pada jumlah data uji 150 sebesar 92% dan 100% dengan nilai parameter yang digunakan adalah jumlah *hidden layer* 10 buah, mse 0,01, *learning rate* 0,5 dan momentum 0,6. Hal ini tergantung pada data latih yang digunakan dimana kosakata pada data latih dapat mewakili data uji yang digunakan dan *noise* yang terdapat pada kategori *automotive* dan *computers* sedikit.



Gambar 7. Grafik Pengujian 3

Pengujian selanjutnya dilakukan dengan mengukur waktu yang diperlukan untuk melakukan proses pelatihan. Pengukuran waktu dilakukan tanpa menghitung waktu untuk menampilkan halaman antar muka. Perhitungan waktu dimulai pada saat program mulai melakukan proses pelatihan. Pengujian pertama dilakukan dengan mengubah jumlah data latih kemudian diukur waktunya. Pengujian kedua dengan mengubah jumlah unit pada *hidden layer*, kemudian dihitung waktunya. Pengujian berikutnya dilakukan pada penambahan jumlah data uji.

Pengujian pertama akan menghitung waktu yang diperlukan untuk melakukan proses pelatihan terhadap perubahan jumlah data latih. Jumlah data latih yang diuji yakni 10, 20 dan 30. Parameter jaringan yang digunakan adalah MSE 0,01, jumlah unit 10, *learning rate* 0,5, momentum 0,6. Berikut pada tabel 4 merupakan pengujian terhadap perubahan jumlah data latih.

Tabel 4. Hasil Pengujian 4

No	Jumlah data latih	Waktu (detik)
1	10	49,375
2	20	207,796875
3	30	504,71875

Berdasarkan tabel di atas terlihat bahwa dengan semakin banyaknya data latih yang digunakan dalam proses pelatihan maka waktu yang dibutuhkan untuk melatih data tersebut semakin lama Hal ini disebabkan karena meningkatnya jumlah kosakata yang digunakan sebagai input jaringan MLP,

sehingga proses pembelajaran untuk mengenali pola input yang berupa vektor dilakukan lebih banyak oleh karena itu proses yang dilakukan lebih lama.

Pengujian kedua akan menghitung waktu yang diperlukan untuk melakukan proses pelatihan terhadap perubahan jumlah unit pada *hidden layer*. Jumlah unit *hidden* yang diuji yakni 5, 10, 15 dan 25. Parameter jaringan yang digunakan adalah MSE 0,01, jumlah data latih 30, *learning rate* 0,5, momentum 0,6. Berikut pada tabel 5 yang dihasilkan dari pengujian perubahan jumlah unit *hidden layer*.

Tabel 5. Hasil Pengujian 5

No	Jumlah unit	Waktu (detik)
1	5	504,703125
2	10	504,71875
3	15	1019,9375
4	25	1971,578125

Berdasarkan tabel di atas terlihat bahwa jumlah unit pada *hidden layer* berpengaruh pada waktu yang dibutuhkan untuk pelatihan jaringan *Multi Layer Perceptron*. Dimana semakin besar jumlah unit *hidden layer*, waktu pelatihan yang dibutuhkan lebih lama. Hal ini disebabkan dengan meningkatnya jumlah *hidden layer* kompleksitas jaringan menjadi lebih tinggi, sehingga proses yang dilakukan untuk pelatihan menjadi lebih lama.

Pengujian ketiga akan menghitung waktu yang diperlukan untuk melakukan proses pengujian pada sebuah *website* dengan alamat *"http://www.bleepingcomputer.com/"*. Parameter jaringan yang digunakan adalah MSE 0,01, jumlah data latih 30, *learning rate* 0,5, momentum 0,6 dan 10 unit *hidden layer*. Setelah dilakukan pengujian didapatkan rata-rata waktu proses pengujian untuk 1 buah *website* sebesar 204 detik atau sekitar 3,4 menit. Waktu yang dibutuhkan bervariasi tergantung jumlah kata dalam *website* yang diuji serta kemampuan komputer yang digunakan.

4. KESIMPULAN

Berdasarkan hasil uji coba yang telah dilakukan dapat diambil kesimpulan bahwa kelebihan dari penggunaan algoritma *multi layer perceptron* ini yaitu kemampuannya dalam melakukan proses klasifikasi cukup baik. Nilai akurasi tertinggi diperoleh sebesar 81 % untuk jumlah data latih 30 buah/kategori, jumlah data uji 100 buah, jumlah unit *hidden* 10, MSE 0,01, *learning rate* 0,5 dan momentum 0,6. Nilai akurasi yang dihasilkan dapat diatur sedemikian rupa dengan konfigurasi berbagai parameter jaringan sehingga menghasilkan akurasi yang optimal. Namun kekurangan yang paling terasa dari penggunaan algoritma ini yaitu lamanya waktu pelatihan, selain itu penggunaan data latih berupa *website* mengakibatkan proses perubahan data *website* menjadi vektor input cukup lama. Oleh karena itu penentuan parameter jaringan yang tepat sangat diperlukan untuk mempercepat proses pelatihan. Konfigurasi parameter jaringan ini bisa berbeda-beda dari satu set data dengan lainnya, sehingga diperlukan eksperimen lebih lanjut untuk mencari nilai terbaik.

5. DAFTAR PUSTAKA

[1] Xiaoguang Qi dan Brian D. Davison, "*Web Page Classification: Features and Algorithms", Department of Computer Science & Engineering Lehigh University, 2007.*

[2] Hans-Peter Kriegel dan Matthias Schubert, "*Classification of Websites as Sets of Feature Vectors", Institute for Computer Science,University of Munich, 2004.*

[3] Jaideep Srivastava, Prasanna Desikan, Vipin Kumar, " *Web Mining –Concepts, Applications & research Directions", 2001.*

[4] Anupan Joshi dan Pranam Kolari, "*Web mining : Research and practice*", University of Maryland, Baltimore County, 2011.

[5] Ajay S. Patil dan B.V. Pawar, " *Automated Classification of Web Sites using Naive Bayesian Algorithm*", Proc. IMECS Hongkong, 2012.

[6] Pornpon Thamrongrat dan Ladda Preechaveerakul dan Wiphada Wettayaprasit. " *A Novel Voting Algorithm of Multi-Class SVM for Web Page Classification", 2009.*

[7] Daniele Riboni, "*Feature Selection for Web Page Classification"*, Universita' degli Studi di Milano, Italy.

[8] Dominic Savio, Lam Lai Yin, "*Learned Text categorization by Backpropagation Neural Network*", The Hong Kong University of Science andTechnology, 1996.