Jurnal Sistem Komputer dan Informatika (JSON)
Volume 2, Nomor 2, Januari 2021
e-ISSN 2685-998X
DOI 10.30865/json.v2i2.2703

Perbandingan Klasifikasi Website Secara Otomatis Menggunakan Metode Multilayer Perceptron dan Naive Bayes

I Nyoman Purnama

Sistem Informasi, STMIK PRIMAKARA, Denpasar, Indonesia
Email: purnama@primakara.ac.id
Submitted: **22/12/2020**; Accepted: **08/01/2021**; Published: **24/01/2021**

Abstrak–World wide web telah menjadi sebuah gudang besar yang berisi banyak informasi. Karena jumlah situs web terus tumbuh secara eksponensial, kebutuhan klasifikasi situs web sangat diperlukan. Kemampuan manusia untuk melakukan klasifikasi manual semakin sulit. Klasifikasi Website dengan menggunakan teknik machine learning menjadi hal yang penting dilakukan, karena proses klasifikasi dilakukan secara otomatis. Sistem klasifikasi dimulai dengan proses pengumpulan informasi dari halaman depan situs web (parsing). Halaman beranda situs web adalah halaman khusus yang berfungsi sebagai titik masuk dengan menyediakan tautan ke seluruh situs web. Untuk setiap hasil parsing homepage terdapat proses penghapusan kata henti, stemming dan pemilihan fitur dengan tf-idf. Hasil dari proses ini adalah fitur yang menjadi masukan dari algoritma pembelajaran mesin. Dalam algoritma ini terdapat proses pembelajaran pola input dan pembuatan bobot. Bobot ini akan digunakan dalam proses klasifikasi. Pada penelitian ini dikembangkan proses pembelajaran dengan menggunakan algoritma multi layer perceptron dan naive bayes. Hasil klasifikasi dari setiap metode akan dibandingkan tingkat akurasinya. Berdasarkan hasil yang diperoleh, naive bayes memiliki tingkat akurasi yang lebih baik dari pada multi layer perceptron. Dengan akurasi untuk Naive Bayes sebesar 89% dan MLP memiliki akurasi 80%.

Kata Kunci: Website; Classification; Multilayer Perceptron; Naïve Bayes

Abstract–World wide web has become a big repository that contains lot of information. As the number of websites continuous growth exponentialy, the need of website classification gains attractions. The human ability to perform manual classifications is increasingly difficult. Website Classification using machine learning technique become more important to do, because the classification process is done automatically. The classification system begins with the process of collecting information from the home page of the web site(parsing). Home page of a web site is a distinguished page and it acts as an entry point by providing links to the rest of the web site. For each parsing result of the homepage there are process of removing the stop word, stemming and feature selection with tf-idf. The result of this process is a feature that becomes input of machine learning algorithm. In this algorithm there are learning process of input pattern and making of the weight. This weight will be used in the classification process. In this research, the learning process is developed by using multi layer perceptron and naive bayes algorithm. The classification results of each method will be compared. Based on the results obtained, naive bayes have a better accuracy rate than multilayer perceptron. With accuracy of 89% Naive Bayes and MLP has 80% accuracy.

Keywords: Website; Classification; Multilayer Perceptron; Naïve Bayes

1. PENDAHULUAN

World wide web menjadi sumber penting dalam pencarian informasi. Dengan semakin meningkatnya jumlah website, kemampuan untuk mendapatkan informasi yang spesifik sangat diperlukan. Menemukan kembali informasi merupakan salah satu kegunaan dari Information retrieval (Sistem temu kembali Informasi). Dimana dengan adanya system IR ini, informasi bisa lebih mudah diperoleh. Salah satu cara untuk mempersempit pencarian informasi yakni dengan jalan mengklasifikasikan informasi dalam suatu website terlebih dahulu. Klasifikasi website adalah proses untuk memberikan label sebuah website berdasarkan kategori yang sesuai, sehingga proses pencarian informasi bisa lebih cepat[1]. Klasifikasi website berbeda dengan klasifikasi teks biasa. Dimana website terdiri dari struktur yang cukup kompleks, yang didalamnya terdapat tag-tag HTML, tautan dan javascript[2]. Pada penelitian ini difokuskan pada web content mining dimana proses mining/pengumpulan informasi diambil dari isi websitenya. 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. Halaman depan merupakan jalan masuk utama untuk mengetahui keseluruhan isi website[3]. Selain itu kegunaan dari klasifikasi website, adalah untuk mengelompokkan website secara otomatis.

Ada beberapa algoritma yang umum digunakan pada proses klasifikasi. Diantaranya yakni C45, Naïve bayes, K-Nearest Neighbor, K-Means, SVM, Neural Networks, Algoritma Genetika dan Multi layer Perceptron. Pada penelitian ini dilakukan perbandingan algoritma klasifikasi Multi layer Perceptron (MLP) dan Naïve Bayes. 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[4]. Sementara Naïve Bayes merupakan metode klasifikasi sederhana yang menerapkan teorema bayes. Dengan menganggap semua fitur saling tidak berhubungan. Algoritma ini mengunakan teori probabilitas dalam memproses klasifikasinya[5]

Jurnal Sistem Komputer dan Informatika (JSON)
Volume 2, Nomor 2, Januari 2021
e-ISSN 2685-998X
DOI 10.30865/json.v2i2.2703

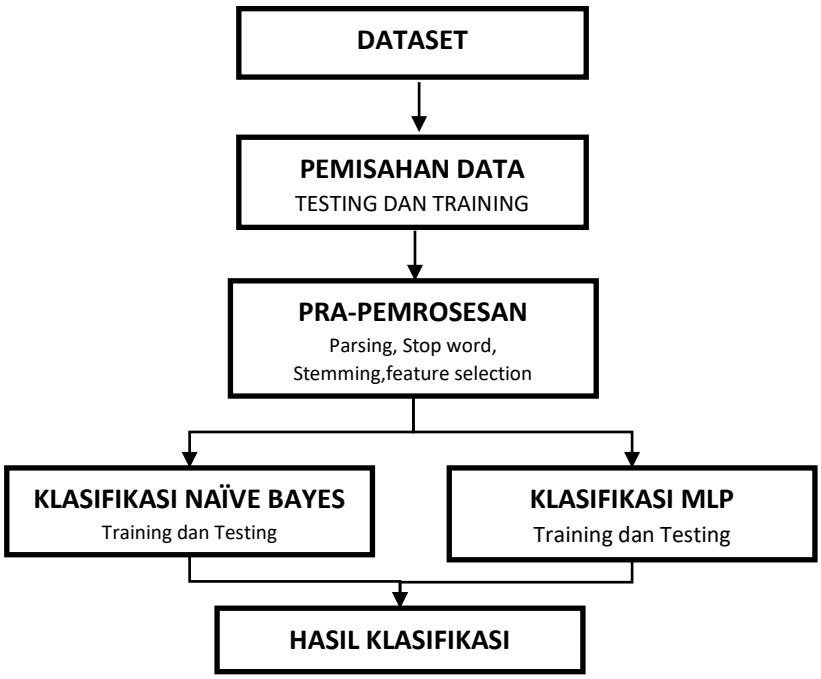
Hasil penelitian yang dilakukan oleh Savio untuk klasifikasi teks dengan menggunakan metode MLP backpropagation didapatkan bahwa pembelajaran menggunakan MLP menghasilkan performa yang bagus[6]. Pada penelitian ini juga dilakukan pengujian dalam mengurangi dimensi dari input pembelajaran dengan metode 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 untuk mencarinya. Penelitian lainnya menggunakan Naïve Bayes untuk klasifikasi web berbasis URL dan menghasilkan nilai Precission 0.7, Recall 0.88 dan F-measure 0.76[7]. Pada penelitian ini klasifikasi juga digunakan untuk memonitor lalu lintas yang tidak normal dalam sebuah orgnisasi.

Berdasarkan penelitian yang telah dilakukan sebelumnya dengan menggunakan MLP dan Naïve bayes, pada penelitian ini akan membandingkan kinerja dari kedua algoritma ini. Karena berdasarkan beberpa penelitian sebelumnya, masing-masing algoritma memiliki kelebihan dan kekurangan masing-masing. Kedua metode ini dibuat dalam aplikasi berbasis PHP dan mysql. Kemudian pengujian dilakukan dengan menggunakan metode precision, F-measure dan recall.

2. METODE PENELITIAN

2.1 Desain Penelitian

Dalam penelitian ini peneliti akan membandingkan 2 metode yang digunakan dalam proses klasifikasi yakni Naïve Bayes dan Multi Layer Perceptron. Alur penelitin klasifikasi website dapat digambarkan pada gambar 1, dimana semua proses dilakukan menggunakan bahasa pemrograman PHP.



Gambar 1. Desain Penelitian

Data penelitian yang digunakan adalah berupa tautan dari halaman utama website (homepage). 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. Data halaman utama website yang telah diunduh ini kemudian dibagi 2 menjadi data training dan data testing. Data training dan testing yang digunakan sebesar masing-masing 500 website. Dari masing-masing halaman website yang telah diunduh ini kemudian dilakukan proses parsing. Proses parsing akan mengambil teks utama atau bagian penting dari sebuah halaman website. Setelah dilakukan proses parsing kemudian dilakukan proses pembuangan kata stop words, stemming dan feature selection. Hasil akhir dari semu proses tadi, berupa indeks kata dengan bobotnya masing-masing. Setelah melalui proses pra pemrosesan kemudian terdapat 2 tahapan utama yakni proses pelatihan dan proses klasifikasi. Proses pelatihan adalah proses pembelajaran yang dikerjakan secara offline. Proses klasifikasi dilakukan secara online dengan mengunduh langsung website yang akan diuji. Proses klasifikasi adalah proses untuk mengkategorikan sebuah alamat

Jurnal Sistem Komputer dan Informatika (JSON)
Volume 2, Nomor 2, Januari 2021
e-ISSN 2685-998X
DOI 10.30865/json.v2i2.2703

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.

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.

2.2 Dataset

Proses untuk mendapatkan informasi pada 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, javascript 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 informasi yang berada dalam tag <TITLE> dan tag <BODY>. 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. Proses untuk ini dinamakan parsing.

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 100, 150 dan 200. 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 yang berupa website 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.
d. Tautan tidak datang dari website berupa portal seperti yahoo.com dll.

Data hasil tokenizing akan diolah pada proses berikutnya yakni pencarian kata dasar sebuah kata (stemming), penghapusan kata henti (stop words) dan Feature Selection dapat dilakukan. Kata-kata hasil stemming dan penghapusan kata 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(Selvakuberan et al., 2009). 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 (Savio, 2016). Hasil dari pra-proses ini kemudian akan digunakan pada algoritma klasifikasi Multi Layer perceptron dan Naïve Bayes sebagai inputan dalam proses pelatihan dan pengujian.

2.3 Metode Yang Digunakan

Metode yang digunakan dalam penelitian ini adalah algoritma klasifikasi Naïve Bayes dan Multi Layer Perceptron. Naïve bayes merupakan pengklasifikasian probavilistik sederhana yang menghitung sekumpulan probabilitas dengan menjumlahkan frekuensi dan kombinasi nilai dari dataset yang diberikan(Ibrahim et al., n.d.). Teorema bayes mengasumsikan semua atribut independen dan tidak saling bergantung yang berdampak pada nilai variable kelas. Keuntungan penggunaan Naïve Bayes adalah metode ini membutuhkan jumlah data pelatihan yang kecil untuk menentukan estimasi parameter yang diperlukan dalam proses pengklasifikasian. Berikut persamaan dari teorema Bayes:

Jurnal Sistem Komputer dan Informatika (JSON)
Volume 2, Nomor 2, Januari 2021
e-ISSN 2685-998X
DOI 10.30865/json.v2i2.2703

$$P(H \mid X) = \frac{P(X \mid H).P(H)}{P(X)} \quad (1)$$

P(H|X) yang dicari merupakan probabilitas hipotesis H berdasarkan kondisi X(posteori probabilitas). Dimana P(X|H) merupakan probabilitas X berdasarkan kondisi pada hipotesis H;X merupakan data class yang belum diketahui; H merupakan data hipotesis suatu class tertentu; P(H) merupakan probabilitas hipotesis H(prior probabilitas) dan P(X) merupakan probabilitas X[8].

Proses klasifikasi mengambil sejumlah petunjuk untuk menentukan kelas mana yang cocok bagi sampel yang dianalisis itu. Karena itu persamaan 1 dirubah menjadi sebagai berikut :

$$P(C \mid F_1 \ldots F_n) = \frac{P(C).P(F_1 \ldots F_n \mid C)}{P(F_1 \ldots F_n)} \quad (2)$$

Dimana variable C merepresentasikan kelas, sementara F1..Fn merepresentasikan karakteristik petunjuk yang dibutuhkan untuk melakukan klasifikasi.

Sedangkan Multilayer perceptron merupakan algortima yang mengadopsi cara kerja jaringan saraf pada makhluk hidup. MLP merupakan topologi yang paling umum dalam JST, dimana masing-masing perceptron terhubung membentuk beberapa lapisan/layer. Sebuah MLP memiliki lapisan masukan(input layer), minimal 1 lapisan hidden layer dan lapisan output[9]. MLP menggunakan pembelajaran propagasi balik(backpropagation) yang harus dilakukan dalam metode ini yaitu inisialisasi (initialization), aktivasi (activation), pelatihan bobot (weight training), dan iterasi (iteration). Pada langkah inisialisasi, nilai awal bobot dan ambang batas (threshold) ditentukan secara acak namun dalam batasan tertentu. Pada tahapan aktivasi, diberikan masukan dan nilai keluaran yang diharapkan (desired output). Proses penyesuaian bobot terjadi pada tahap pelatihan bobot, nilai luaran sebenarnya (actual output) dibandingkan dengan desired output dan dilakukan penyesuaian bobot. Langkah kedua dan ketiga diulangi sampai dengan tercapai kondisi yang ditentukan.

3. HASIL DAN PEMBAHASAN

Data yang digunakan dalam penelitian ini adalah tautan dari halaman utama sebuah website (homepage). Tautan website dikumpulkan terlebih dahulu dari direktori "All business directory (ABD)" dengan alamat http://www.allbusinessdirectory.biz. Jumlah website yang akan digunakan dalam proses pelatihan dibagi menjadi beberapa kelompok dengan jumlah masing-masing 100, 150 dan 200 website. Dalam proses pelatihan, halaman utama website akan diambil secara acak di setiap kategori. Semua website yang digunakan berbahasa Inggris, sesuai dengan kriteria yang telah dijelaskan sebelumnya. Proses pengujian untuk MLP dan naïve bayes dilakukan secara online dengan mengunduh langsung website untuk diuji.

Setelah halaman utama website dikumpulkan, data yang berupa halaman HTML ini akan diproses terlebih dahulu dengan mengambil teks utama dari sebuah halaman. Teks utama ini diambil dengan proses parsing. Pada penelitian ini proses parsing dilakukan dengan mengambil nilai teks yang berada didalam tag HTML seperti body dan title. Setelah proses parsing selesai dilakukan dilanjutkan dengan proses tokenisasi. Proses tokenisasi akan menguraikan kembali semua teks yang diperoleh dari hasil parsing dan dikumpulkan menjadi sebuah tabel yang berisi seluruh kata/term. Langkah selanjutnya adalah menghapus kata henti(stop words) dari tabel kata hasil parsing.

Proses pengolahan kata berikutnya yakni Stemming. Proses Stemming akan mencari kata dasar dari kata-kata yang telah diproses sebelumnya. Proses stemming akan mengurangi kemuncilan kata-kata yang memiliki kata dasar yang sama. Hasil dari semua proses ini dikumpulkan dalam tabel pengetahuan, yang berupa kata-kata unik dan penting. Sebelum dilakukan proses pelatihan dengan menggunakan metode MLP dan naïve bayes, dari setiap dokumen dalam tabel pengetahuan, semua kata/term dianalisis dengan menentukan nilai pembobotan setiap kata dengan metode tf-idf. Metode tf-idf akan menghitung kemunculan dari semua kata dalam tabel pengetahuan dari masing-masing dokumen

$$W_{ij}=TF_{ij}xIDF_{ij}$$

$$W_{ij}=TF_{ij}x\log(D/dfj)$$

Dimana Wij adalah bobot term (tj) terhadap dokumen (di). Sedangkan tfij adalah jumlah kemunculan term (tj) dalam dokumen (di). D adalah adalah jumlah semua dokumen yang ada dalam database dan dfj adalah jumlah dokumen yang mengandung term (tj). Hasil pembobotan ini kemudian diurutkan . Untuk mengurangi jumlah kosakata yang digunakan sebagai input jaringan saraf tiruan / naïve bayes, dari semua kata pada tabel pengetahuan diambil 30 kosakata dengan nilai tf-idf terbesar. Semua kosakata yang dikumpulkan itu unik.

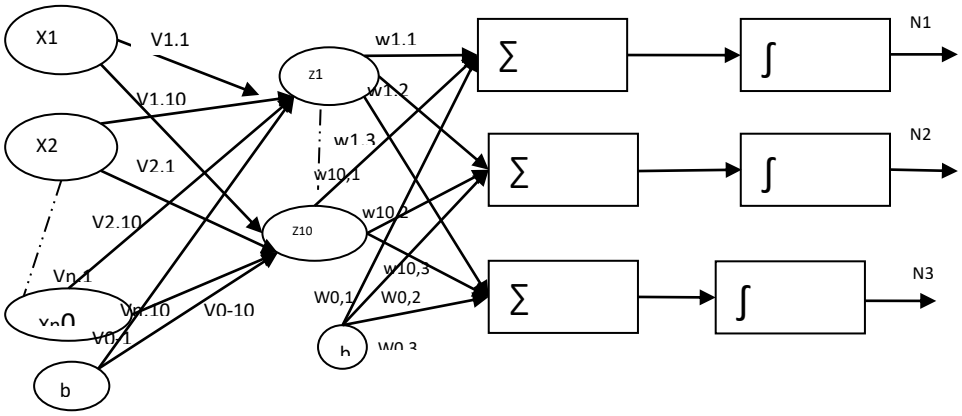
1. Multilayer Perceptron

Jurnal Sistem Komputer dan Informatika (JSON)
Volume 2, Nomor 2, Januari 2021
e-ISSN 2685-998X
DOI 10.30865/json.v2i2.2703

Secara garis besar proses yang terjadi pada klasifikasi teks dengan algoritma Multi layer Perceptron yakni:

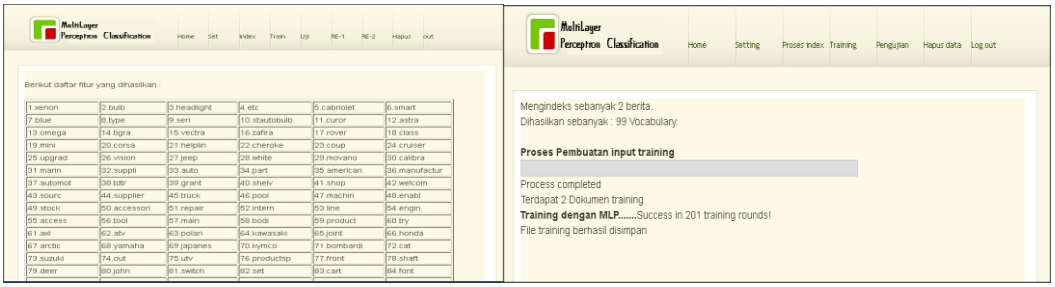
1. Hasil pembobotan kata disimpan kedalam table basis data tb_index. Pada table ini terdapat kolom kosakata, idWebsite dan bobotnya. Nilai bobot ini menjadi vektor input keldalam haring MLP. Masing-masing website hanya diambil 30 kosakata dengan nilai terbesar. Kosakata yang terbesar ini merupakan kata-kata yang penting dari sebuah halaman website.
2. Sebelum dilakukan proses pembelajaran/pelatihan, nilai parameter jaringan multilayer perceptron akan diset terlebih dahulu berupa jumlah epoch, MSE, momentum dan learning rate.
3. Kemudian dengan inputan vektor input dan nilai parameter jaringan, dilakukan proses pembelajaran dengan memanggil fungsi calculate.hasil proses pembelajaran ini disimpan dalam sebuah tabel basisdata pengetahuan yang akan digunakan juga pada proses pengujian.
4. Proses pengujian memiliki langkah yang sama dengan proses pelatihan hanya saja pada proses pengujian tidak ada lagi proses pelatihan dan pembuatan table pengetahuan. Disini juga tidak ada proses perbaikan bobot dengan jalan backpropagation. Proses pengujian diawali dengan tahapan praproses yang telah dijelaskan sebelumnya, kemudian kosakata dengan bobotnya ini akan dibandingkan dengan kosakata yang ada pada table pengetahuan. Kosakata yang tidak terdapat pada table pengetahuan akan diabaikan. Dari kosakata yang tersisa akan dihitung nilai bobot tf-idfnya dan akan menjadi vektor input bagi jaringan MLP.

Arsitektur jaringan Neural network yang dibuat adalah jaringan 3 layer feed forward. Neural Network 3 layer feed forward network, terdiri dari lapisan masukan, lapisan tersembunyi, dan lapisan keluaran. Nilai input neural network adalah vektor, di mana setiap dokumen memiliki jumlah vektor yang berbeda. Fungsi aktivasi akan digunakan di bagian masukan ke lapisan tersembunyi dan dari lapisan tersembunyi ke lapisan keluaran adalah fungsi aktivasi tangen hiperbolik. Nilai yang akan dihasilkan dari fungsi aktivasi tangen hiperbolik adalah -1 dan 1. Jadi untuk mengelompokkan 6 kategori dibutuhkan 3 keluaran untuk membedakannya. Ketiga keluaran tersebut akan menghasilkan kombinasi yang mewakili setiap kategori. Dalam penelitian ini kategori Otomotif dilambangkan dengan kombinasi output [1,1,1], komputer dengan kombinasi output [1,1, -1], pendidikan & pelatihan [1, -1, -1], kesehatan & beauty care [-1, 1,1], olah raga dengan kombinasi [-1,1, -1], travel & rekreasi dengan kombinasi [-1, -1,1].



Gambar 2. Arsitektur Neural Network yang digunakan

Penentuan nilai parameter algoritma MLP akan dapat menentukan keberhasilan pembelajaran. Dalam 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 dan bobot awal random dengan rentang nilai antara -0, 25 hingga 0,25.



Gambar 3. Proses pembentukan index dan training

Jurnal Sistem Komputer dan Informatika (JSON)
Volume 2, Nomor 2, Januari 2021
e-ISSN 2685-998X
DOI 10.30865/json.v2i2.2703

Pelatihan jaringan akan berhenti jika terdapat kesalahan yang lebih kecil dari target penelitian, yang disebut MSE (Mean Squared Error). Jika error tidak ketemu maka jaringan akan berhenti di iterasi maksimal (epoh) yang dimasukkan. Setelah melewati proses pelatihan, bobot yang dihasilkan pada proses pelatihan akan dibobotkan pada proses pengujian.

2. Naïve Bayes

Pada proses klasifikasi dengan naïve bayes, ada 2 proses utama yang dilakukan yakni proses pelatihan dan proses testing. Proses pelatihan dilakukan pada dataset yang telah diketahui kategorinya berdasarkan hasil klasifikasi yang dilakukan oleh reviewer secara manual. Sedangkan proses pengujian merupakan proses untuk mengetahui keakuratan dari sebuah model yang telah dibangun pada proses pelatihan. Data yang digunakan berupa kumpulan website selain yang digunakan pada saat pelatihan. Secara garis besar proses pelatihan dengan menggunakan algoritma Naïve Bayes adalah menentukan probabilitas tiap kategori berdasarkan contoh dari dokumen yang digunakan,. Sedangkan pada proses pengujian atau klasifikasi ditentukan nilai kategori setiap halaman website kemudian menghitung nilai probabilitas dari kemunculan katanya.

Tabel pengetahuan yang sama juga digunakan dalam klasifikasi dengan naïve bayes. Nilai dari tabel hash adalah frekuensi kemunculan kata (nk) di semua dokumen yang termasuk dalam kategori itu. Jumlah kata total (termasuk pengulangan) untuk setiap kategori yang disebut n juga dihitung. Probabilitas posterior dengan koreksi Laplace dihitung menggunakan rumus P (wk | c) = (nk +1) / (n + | Vocabulary |).

Untuk mengklasifikasikan dokumen, katakanlah X, probabilitas dari kategori tertentu dicari dalam tabel hash dan dikalikan bersama. Kategori yang menghasilkan probabilitas tertinggi adalah klasifikasi untuk dokumen X. Hanya kata-kata yang ditemukan di X yang akan dicari di tabel hash. Juga jika sebuah kata dalam X tidak ada dalam kosakata asli (dibangun dari set pelatihan), kata tersebut diabaikan. Persamaan yang digunakan untuk mengklasifikasikan X adalah C = arg max (P (c) Π P (wk | c)).

Akurasi pengklasifikasi sangat buruk yaitu sekitar 45%, ketika hanya 50 dokumen yang disediakan sebagai data pelatihan. Akurasi meningkat setiap kali pengklasifikasi dilengkapi dengan data pembelajaran tambahan. Pengklasifikasi mencapai akurasi 89% ketika hampir 250 dokumen disediakan sebagai input di setiap kategori

3. Perbandingan Hasil

Tabel 2. Hasil pengujian dengan MLP dan Naive bayes

No	Websites theme	MLP			Naive Bayes		
		Akurasi	*Recall*	Presisi	Akurasi	*Recall*	Presisi
1	*Automotive*	71%	76%	76%	88.65%	89.9%	91.37%
2	*Health & Beauty Care*	65%	65%	100%	85.13%	85.13%	85.13%
3	*Education & Training*	88%	88%	94%	93.36%	89.46%	91.37%
4	*Computers*	87%	94%	88%	86.44%	87.65%	87.04%
5	*Sports*	82%	82%	70%	93.36%	88.66%	90.95%
6	Travel & Recreation	88%	88%	94%	90%	89%	89%

4. KESIMPULAN

Berdasarkan hasil percobaan yang telah dilakukan dapat disimpulkan bahwa kelebihan dari penggunaan algoritma / multi layer perceptron adalah kemampuannya untuk melakukan proses klasifikasi dengan cukup baik. Nilai akurasi tertinggi diperoleh sebesar 80% untuk jumlah data latih 250, jumlah hidden units 10, MSE 0.01, learning rate 0.5 dan momentum 0.6. Nilai akurasi yang dihasilkan dapat diatur sedemikian rupa dengan mengkonfigurasi berbagai parameter jaringan sehingga menghasilkan akurasi yang optimal. Namun kekurangan yang paling mencolok dari penggunaan algoritma ini adalah lamanya waktu pelatihan, selain itu penggunaan data latih berupa website mengakibatkan proses perubahan data website menjadi vektor input yang cukup lama. Oleh karena itu, penentuan parameter jaringan yang tepat sangat diperlukan untuk mempercepat proses pelatihan. Konfigurasi parameter jaringan ini mungkin berbeda dari satu kumpulan data ke kumpulan lainnya, membutuhkan eksperimen lebih lanjut untuk menemukan nilai terbaik. Pendekatan NB untuk klasifikasi halaman rumah untuk 6 kategori yang dipertimbangkan di atas menghasilkan akurasi 89,05%. Juga diamati bahwa keakuratan klasifikasi dari pengklasifikasi sebanding dengan jumlah dokumen pelatihan. Akurasi terbaik diperoleh dengan menggunakan Naive bayes. Selain itu dari segi kecepatan, algoritma ini juga lebih cepat. Namun pada beberapa kategori MLP memiliki tingkat akurasi yang lebih baik, seperti pada kategori komputer.

Jurnal Sistem Komputer dan Informatika (JSON)
Volume 2, Nomor 2, Januari 2021
e-ISSN 2685-998X
DOI 10.30865/json.v2i2.2703

REFERENCES

[1] X. Qi and B. D. Davison, "Web page classification: Features and algorithms," *ACM Comput. Surv.*, vol. 41, no. 2, 2009, doi: 10.1145/1459352.1459357.
[2] K. Selvakuberan, M. Indra Devi, and R. Rajaram, "Feature selection for web page classification," *Soc. Implic. Data Min. Inf. Priv. Interdiscip. Fram. Solut.*, pp. 213–228, 2009, doi: 10.4018/978-1-60566-196-4.ch012.
[3] A. S. Patil and B. V. Pawar, "Automated classification of web sites using Naive Bayesian algorithm," *Lect. Notes Eng. Comput. Sci.*, vol. 2195, pp. 519–523, 2012.
[4] A. Muliantara and I. Widiartha, "Penerapan Multi Layer Perceptron Dalam Anotasi Image Secara Otomatis," *J. Ilmu Komput.*, vol. 4, no. 1, pp. 9–15, 2011.
[5] B. Kurniawan, M. A. Fauzi, and A. W. Widodo, "Klasifikasi Berita Twitter Menggunakan Metode Improved Naïve Bayes," *J. Pengemb. Teknol. Inf. dan Ilmu Komput. Univ. Brawijaya*, vol. 1, no. 10, pp. 1193–1200, 2017.
[6] D. Savio, "NETWORK," no. June 2001, 2016.
[7] R. Rajalakshmi and C. Aravindan, "Naive Bayes approach for website classification," *Commun. Comput. Inf. Sci.*, vol. 147 CCIS, no. January, pp. 323–326, 2011, doi: 10.1007/978-3-642-20573-6_55.
[8] I. Anggraini and Y. N. Kunang, "Telematika Penerapan Naïve Bayes pada Pendeteksian Malware dengan Diskritisasi Variabel," vol. 13, no. 1, pp. 11–21, 2020.
[9] N. Purwaningsih, "Penerapan multilayer perceptron untuk klasifikasi jenis kulit sapi tersamak," *J. TEKNOIF*, vol. 4, no. 1, pp. 1–7, 2016.