LONTAR KOMPUTER VOL. 7. NO.1. APRIL 2016 **p-ISSN 2088-1541**
DOI: 10.24843/LKJITI.2016.v07.i01.p03 **e-ISSN 2541-5832**

Optimasi Naïve Bayes Dengan Pemilihan Fitur Dan Pembobotan *Gain Ratio*

I. Gusti. A. Socrates[1], Afrizal L. Akbar[2], M. Sonhaji Akbar[3]

Teknik Informatika, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia
[1]socrates15@mhs.if.its.ac.id
[2]afrizal.la@gmail.com
[3]mson.akbar@gmail.com

Abstrak

Naïve Bayes merupakan salah satu metode data mining yang umum digunakan dalam klasifikasi dokumen berbasis text. Kelebihan dari metode ini adalah algoritma yang sederhana dengan kompleksitas perhitungan yang rendah. Akan tetapi, pada metode Naïve Bayes terdapat kelemahan dimana sifat independensi dari fitur Naïve Bayes tidak dapat selalu diterapkan sehingga akan berpengaruh pada tingkat akurasi perhitungan. Maka dari itu, metode Naïve Bayes perlu dioptimasi dengan cara pemberian bobot mengunakan Gain Ratio. Namun, pemberian bobot pada Naïve Bayes menimbulkan permasalahan pada penghitungan probabilitas setiap dokumen, dimana fitur yang tidak merepresentasikan kelas yang diuji banyak muncul sehingga terjadi kesalahan klasifikasi. Oleh karena itu, pembobotan Naïve Bayes masih belum optimal. Paper ini mengusulkan optimasi metode Naïve Bayes mengunakan pembobotan Gain Ratio yang ditambahkan dengan metode pemilihan fitur pada kasus klasifikasi teks. Hasil penelitian ini menunjukkan bahwa optimasi metode Naïve Bayes menggunakan pemilihan fitur dan pembobotan menghasilkan akurasi sebesar 94%.

***Kata Kunci**: Data Mining, Naïve Bayes, Weighted Naïve Bayes, Gain Ratio, Pemilihan Fitur.*

Abstract

Naïve Bayes is one of data mining methods that are commonly used in text-based document classification. The advantage of this method is a simple algorithm with low computation complexity. However, there is weaknesses on Naïve Bayes methods where independence of Naïve Bayes features can't be always implemented that would affect the accuracy of the calculation. Therefore, Naïve Bayes methods need to be optimized by assigning weights using Gain Ratio on its features. However, assigning weights on Naïve Bayes's features cause problems in calculating the probability of each document which is caused by there are many features in the document that not represent the tested class. Therefore, the weighting Naïve Bayes is still not optimal. This paper proposes optimization of Naïve Bayes method using weighted by Gain Ratio and feature selection method in the case of text classification. Results of this study pointed-out that Naïve Bayes optimization using feature selection and weighting produces accuracy of 94%.

***Keywords**: Data Mining, Naïve Bayes, Weighted Naïve Bayes, Gain Ratio, Feature Selection.*

1. Pendahuluan

Klasifikasi merupakan proses pengidentifikasian obyek ke dalam sebuah kelas, kelompok, atau kategori berdasarkan prosedur, karakteristik dan definisi yang telah ditentukan sebelumnya [1]. Salah satu bentuk klasifikasi yaitu klasifikasi dokumen atau teks. Klasifikasi dokumen atau teks adalah bidang penelitian dalam pengolahan informasi. Tujuan dari klasifikasi dokumen adalah mengembangkan sebuah metode dalam menentukan atau mengkategorikan suatu dokumen ke dalam satu atau lebih kelompok secara otomatis berdasarkan isi dokumen [2]. Pada era ini pengelompokkan teks atau dokumen digunakan untuk proses pencarian sebuah dokumen.

Maka dari itu, kebutuhan untuk pengelompokkan dokumen secara cepat dan mudah sangat penting. Sedangkan saat ini, pengelompokkan dokumen masih menggunakan cara manual.

Pengelompokkan dokumen dilakukan dengan cara pemberian label terhadap kategori dokumen. Dibutuhkan waktu yang cukup lama dalam mengklasifikasikan dokumen. Maka dari itu, dibutuhkan metode yang dapat digunakan dalam proses klasifikasi atau pengelompokkan dokumen secara cepat dan akurat.

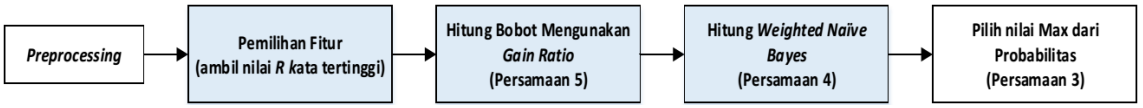
Salah satu metode klasifikasi yang biasa digunakan adalah *Naïve Bayes*. Klasifikasi *Naïve Bayes* pertama kali dikemukakan oleh Revered Thomas Bayes. Penggunaan metode *Naïve Bayes* sudah dikenalkan sejak tahun 1702-1761. *Naive Bayes* (atau dikenal sebagai *Simple Bayes*) menurut Lewis, Hand dan Yu merupakan pendekatan yang sangat sederhana dan sangat efektif untuk *classification learning* [3][4]. Sedangkan menurut Kononenko dan Langley menyimpulkan bahwa *Naïve Bayes* merupakan kemungkinan label kelas data atau bisa diasumsikan sebagai atribut kelas yang diberi label [5][6].

Menurut Hamzah Naïve Bayes memiliki beberapa kelebihan, yaitu algoritma yang sederhana, lebih cepat dalam penghitungan dan berakurasi tinggi [7]. Akan tetapi, pada metode *Naïve Bayes* juga memiliki kelemahan dimana sebuah probabilitas tidak bisa mengukur seberapa besar tingkat keakuratan sebuah prediksi. Maka dari itu, metode *Naïve Bayes* perlu dioptimasi dengan cara pemberian bobot mengunakan *Gain Ratio*. Pemberian bobot pada *Naïve Bayes* menimbulkan permasalahan pada penghitungan probabilitas setiap dokumen. Dimana fitur yang tidak merepresentasikan kelas yang diuji banyak muncul sehingga terjadi kesalahan klasifikasi. Oleh karena itu, pembobotan *Naïve Bayes* masih belum optimal.

Maka dari itu, Paper ini mengusulkan optimasi metode *Naïve Bayes* mengunakan pembobotan *Gain* Ratio yang ditambahkan dengan metode pemilihan fitur pada kasus pemilihan teks.

2. Metode Penelitian

Metode *Naïve Bayes* merupakan salah satu algoritma yang efektif dan efisien dalam proses klasifikasi [3][4]. Pada Gambar 1 menampilkan metode usulan *Weighted Naïve Bayes* dengan menggunakan *Gain Ratio*.



Gambar 1. Alur Metode Penelitian

2.1. Dataset

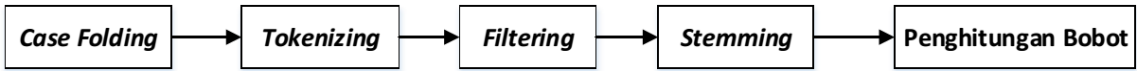
Dataset yang digunakan dalam penelitian ini diambil dari media online yaitu kompas, detik, dan tempo. Kemudian dilakukan proses penentuan kata dasar, penentuan kata umum yang sering muncul atau stopwords, dan penentuan kategori. Proses pengolahan dataset dapat dilihat pada Gambar 2.

Gambar 2. *Dataset*

2.2. Preprocessing

Preprocessing adalah proses awal pada klasifikasi dokumen yang bertujuan untuk menyiapkan data agar menjadi terstruktur. Hasil dari *preprocessing* akan berupa nilai numerik sehingga dapat dijadikan sebagai sumber data yang dapat diolah lebih lanjut. *Preprocessing* ini terbagi menjadi beberapa proses yang terdiri dari *case folding*, *tokenizing, filtering, stemming* dan penghitungan bobot kata.

Pada Gambar 3 terdapat proses *preprocessing. Case folding* merupakan tahap awal dari *preprocessing text* yang *mengubah* karakter huruf teks menjadi huruf kecil semua [8]. Karakter yang diterima hanya 'a' hingga 'z'. Karakter selain huruf akan dihilangkan dan dianggap sebagai *delimiter*. *Tokenizing* adalah tahap pemotongan *string input* berdasarkan tiap kata yang menyusunnya [9]. *Filtering* adalah proses menentukan kata-kata (*terms*) apa saja yang akan digunakan untuk merepresentasikan dokumen. Selain untuk menggambarkan isi dokumen, *term* ini juga berguna untuk membedakan dokumen yang satu dengan dokumen lainnya pada koleksi dokumen. Proses ini dilakukan dengan mengambil kata-kata penting dari hasil *token* dan menghapus *stop words*. *Stop words* adalah kata-kata yang tidak deskriptif sehingga dapat dibuang atau dihilangkan dan tidak berpengaruh ke dalam proses [8]. Dalam bahasa Indonesia, contoh *stop words* seperti "yang", "dan", "dari", "di", "seperti" dan lainnya. Tahap *stemming* adalah tahap mencari *root* (akar) kata dari kata hasil *filtering*. Pada tahap ini dilakukan proses pengambilan berbagai bentukan kata ke dalam suatu representasi yang sama. *Stem* (akar kata) merupakan bagian dari kata yang tersisa setelah dihilangkan imbuhannya (awalan dan akhiran). Contoh kata beri adalah *stem* dari memberi, diberikan, memberikan dan pemberian.



Gambar 3. *Preprocessing*

2.3. Penghitungan bobot

a. *Bayes*

Naive bayes adalah metode yang digunakan dalam statistika untuk menghitung peluang dari suatu hipotesis, *Naïve Bayes* menghitung peluang suatu kelas berdasarkan pada atribut yang dimiliki dan menentukan kelas yang memiliki probabilitas paling tinggi. *Naive bayes* mengklasifikasikan kelas berdasarkan pada probabilitas sederhana dengan mangasumsikan bahwa setiap atribut dalam data tersebut bersifat saling terpisah. Metode *Naive Bayes* merupakan salah satu metode yang banyak digunakan berdasarkan beberapa sifatnya yang sederhana, metode *Naive Bayes* mengklasifikasikan data berdasarkan probabilitas *P* atribut *x* dari setiap kelas *y* data. Pada model probablitas setiap kelas *k* dan jumlah atribut *a* yang dapat dituliskan seperti Persamaan (1) [2] berikut.

$$P(y_k|x_1, x_2, \ldots . x_a) \tag{1}$$

Penghitungan *Naïve Bayes* yaitu probabilitas dari kemunculan dokumen x_a pada kategori kelas y_k $P(x_a|y_k)$, dikali dengan probabilitas kategori kelas $P(y_k)$. Dari hasil kali tersebut kemudian dilakukan pembagian terhadap probabilitas kemunculan dokumen $P(x_a)$. Sehingga didapatkan rumus penghitungan *Naïve Bayes* dituliskan pada Persamaan (2) [2].

$$P(y_k|x_a) = \frac{P(y_k)P(x_a|y_k)}{P(x_a)} \tag{2}$$

Kemudian dilakukan proses pemilihan kelas yang optimal maka dipilih nilai peluang terbesar dari setiap probabilitas kelas yang ada. Sehingga didapatkan rumus untuk memilih nilai terbesar pada Persamaan (3) [10].

$$y(x_i) = \arg\max P(y) \prod_{i=1}^{a} P(X_i|y) \tag{3}$$

b. *Weighted Naive Bayes*

Menurut Hilden, Ferreira, dan Hall pembobotan atribut kelas dapat meningkatkan pengaruh prediksi [11][12][13]. Dengan memperhitungkan bobot atribut terhadap kelas, maka yang menjadi dasar ketepatan klasisifikasi bukan hanya probabilitas melainkan juga dari bobot setiap atribut terhadap kelas. Pembobotan *Naïve Bayes* dihitung dengan cara menambahkan bobot w_i pada setiap atribut. Sehingga didapatkan rumus untuk pembobotan *Naïve Bayes* dituliskan pada Persamaan (4).

$$P(y,x) = P(y)\prod\prod_{i=1}^{a} P(X_i|y)^{W_i} \tag{4}$$

Pembobotan dapat dirumuskan menggunakan *Gain Ratio* [10]. Dimana dari setiap atribut *Gain Ratio* dikali jumlah data *n* kemudian dibagi dengan rata-rata *Gain Ratio* semua atribut.

$$w_i = \frac{GainRatio(i)}{\frac{1}{a}\sum_{i=1}^{a} GainRatio(i)} \tag{5}$$

Atribut dari *Gain Ratio* sendiri merupakan hasil bagi dari *Mutual* Information dan *Entropy*. *Mutual Information* (MI) merupakan nilai ukur yang menyatakan keterikatan atau ketergantungan antara dua variabel atau lebih. Unit pengukur yang umum digunakan untuk menghitung MI adalah bit, sehingga menggunakan logaritma (*log)* basis 2. Secara formal, MI digunakan antara 2 variabel A dan B yang didefinisikan oleh Kulback dan Leibler [14], [15]. Selain MI, *Entropy* digunakan sebagai pembagi dari MI yang digunakan untuk menentukan atribut mana yang terbaik atau optimal. Penghitungan *Mutual Information* dituliskan pada Persamaan 6 [14][15].

$$MI(x_i, y) = \sum_{y}\sum_{x_1} P(x_i, y)\log\frac{P(x_1, y)}{P(x_1)P(y)} \tag{6}$$

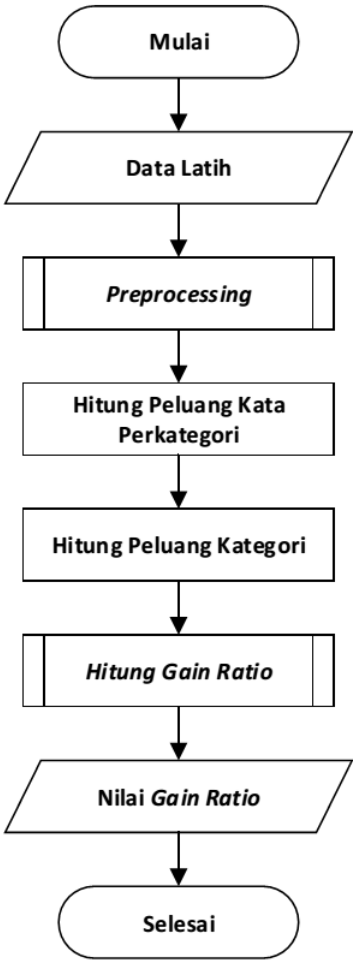
Sebelum mendapatkan nilai *Gain Ratio* dilakukan pencarian nilai *Entropy E*. *Entropy* digunakan untuk menentukan seberapa informatif sebuah masukan atribut untuk menghasilkan keluaran atribut. Penghitungan *Entropy* dengan menjumlahkan probabilitas dituliskan pada Persamaan (7).

$$E(x_i) = \sum_{x_1} P(x_1)\log\frac{1}{P(x_1)} \tag{7}$$

Maka dari itu penghitungan *Gain Ratio* adalah hasil dari penghitungan *Mutual Information* dibagi dengan hasil penghitungan *Entropy* Penghitungan *Gain Ratio* dituliskan pada Persamaan (8).

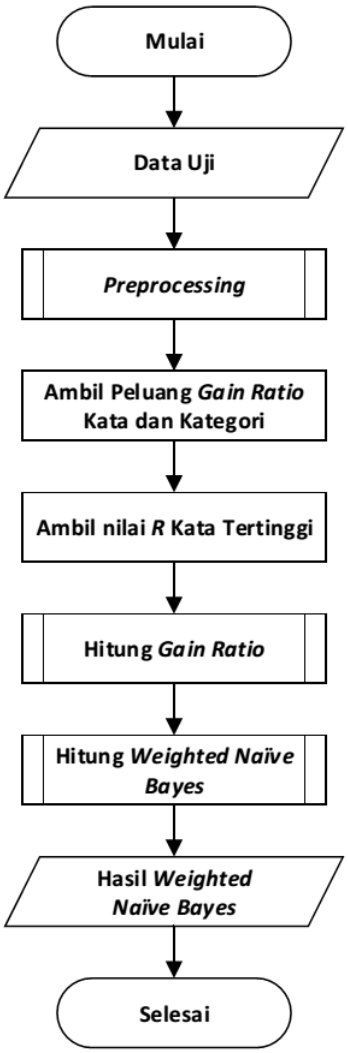
$$GainRatio(i) = \frac{MI(x_i, y)}{E(x_i)} = \frac{\sum_{y}\sum_{x_1} P(x_i, y)\log\frac{P(x_1, y)}{P(x_1)P(y)}}{\sum_{x_1} P(x_1)\log\frac{1}{P(x_1)}} \tag{8}$$

Proses penghitungan *Weighted Naïve Bayes* menggunakan *Gain Ratio* dibagi menjadi dua tahap. Tahap pertama adalah proses *training* (pelatihan). Pada proses *training* diambil data latih kemudian dilakukan preprocessing. Setelah itu hitung peluang kata *(term)* perkategori dan hitung peluang kategori *(class).* Kemudian dicari nilai *Gain Ratio* menggunakan Persamaan 8. Proses *training* dapat dilihat pada Gambar 4.



Gambar 4. Proses Training

Tahap kedua adalah proses *testing* (pelatihan). Pada proses *testing* diambil data uji kemudian dilakukan preprocessing. Setelah itu ambil nilai *Gain Ratio* tiap kata dan kategori. Setelah itu, dilakukan proses perankingan kata sebanyak *R* (jumlah kata yang ditentukan). Dari kata sebanyak *R* yang diambil dilakukakn proses penghitungan *Gain Ratio*. Kemudian dicari nilai *Weighted Naïve Bayes* menggunakan Persamaan 4. Proses *testing* dapat dilihat pada Gambar 5.



Gambar 5. Proses Testing

c. *Metode Evaluasi*

Pada tahap evaluasi bertujuan untuk mengetahui tingkat akurasi dari hasil penggunaan metode *Weighted Naïve Bayes*. Dari evaluasi akan tersedia informasi mengenai seberapa besar akurasi yang telah dicapai. Pada proses pengujian dikenal sebagai Matriks *Confusion* yang merepresentasikan kebenaran dari sebuah klasifikasi. Tabel Matriks *Confusion* dapat dilihat pada Tabel 1.

Tabel 1. Matriks Confusion

		Hasil Prediksi	
		+	**-**
Kenyataan	**+**	*True Positive*	*False Positive*
	-	*False Negative*	*True Negative*

- *True Positive* (TP) menunjukkan bahwa dokumen yang termasuk dalam hasil pengelompokkan oleh sistem memang merupakan anggota kelas.
- *False Positive* (FP) menunjukkan bahwa dokumen yang termasuk dalam hasil pengelompokkan oleh sistem ternyata seharusnya bukan merupakan anggota kelas.

- *False Negative* (FN) menunjukkan bahwa dokumen yang tidak termasuk dalam hasil pengelompokkan oleh sistem ternyata seharusnya merupakan anggota kelas.
- *True Negative* (TN) menunjukkan bahwa dokumen yang tidak termasuk dalam hasil pengelompokkan oleh sistem ternyata seharusnya bukan merupakan anggota kelas.

Untuk menghitung tingkat akurasi digunakan Persamaan 9 [16].

$$Akurasi = \frac{TP + TN}{TP + TN + FP + FN} \quad (9)$$

3. Eksperimen dan Hasil

Pengujian hasil menggunakan metode *Wighted Naïve Bayes* dilakukan dengan membandingkan hasil percobaan *Naïve Bayes* tanpa menggunakan pembobotan. Perbandingan dilakukan terhadap dokumen berita sejumlah 65 dokumen pada uji coba 1 dan 145 dokumen pada uji coba 2. Hasil yang dibandingkan adalah akurasi data yang dihasilkan dengan menghitung selisih antara *Weighted Naïve Bayes* dan *Naïve Bayes* biasa. Penghitungan akurasi tersebut dapat dilihat pada Persamaan 9.

Dilakukan uji coba 1 terhadap metode usulan dengan menggunakan data latih sebanyak 35 dokumen dan data uji sebanyak 30 dokumen. Pada uji coba 2, data uji yang digunakan sebanyak 110 dokumen dan data latih yang digunakan sama seperti uji coba 1. Dimana, pada data latih terdapat 7 kategori, yaitu Sepak Bola, Otomotif, Kesehatan, Teknologi, Ekonomi, Politik, dan Hukum. Pada masing-masing kategori berisi 5 dokumen.

Dari hasil uji coba 1 didapatkan hasil akurasi *Naïve Bayes* sebesar 92% sedangkan pada Weighted *Naïve Bayes* sebesar 94%. Selain itu, dari hasil uji coba 2 didapatkan hasil akurasi *Naïve Bayes* sebesar 92% dan *Weighted Naïve Bayes* sebesar 84%. Hasil akurasi dapat dilihat pada Tabel 2.

Tabel 2. Hasil Akurasi

Metode	Akurasi %	
	Uji Coba 1	Uji Coba 2
Naïve Bayes	92	**92**
Weighted Naïve Bayes	**94**	84

Berdasarkan uji coba 2, dilakukan proses pemilihan fitur sebanyak *R* (50, 30, dan 10 *term* terbaik). Dari hasil pemilihan fitur menggunakan 50 dan 30 *term* terbaik didapatkan akurasi sebesar 91% untuk metode usulan dan 95% untuk metode *Naïve Bayes* biasa. Sedangkan ketika menggunakan 10 *term* terbaik didapatkan akurasi sebesar 94% untuk metode usulan dan 91% untuk metode *Naïve Bayes* biasa. Hasil uji coba terhadap pemilihan fitur dapat dilihat pada Tabel 3.

Tabel 3. Pemilihan Fitur

Term Terbaik	Metode Usulan %	Naïve Bayes %
50	91	**95**
30	91	**95**
10	**94**	91

4. Pembahasan

Dari hasil uji coba 1 didapatkan nilai akurasi *Naïve Bayes* sebesar 92% sedangkan nilai akurasi untuk metode yang diusulkan atau *Weighted Naïve Bayes* sebesar 94%. Hasil metode yang diusulkan lebih tinggi disebabkan oleh pemberian bobot pada probabilitas dari setiap kata pada dokumen terhadap kategori. Pemberian bobot pada probabilitas mengakibatkan jarak antar peluang satu kata terhadap kategori semakin jauh. Hasil dari penelitian yang diusulkan sesuai

dengan penelitian Hilden, Ferreira dan Hall yang berpendapat bahwa pembobotan atribut kelas dapat meningkatkan pengaruh prediksi [11][12][13].

Akan tetapi pada uji coba 2, akurasi pada metode yang diusulkan cenderung rendah dibandingkan dengan *Naïve Bayes* biasa. Hal ini dikarenakan *term* yang sering muncul pada seluruh kategori dokumen menghasilkan nilai *Gain Ratio* yang tinggi dan mengakibatkan terjadinya kesalahan klasifikasi. Setelah diketahui hasil akurasi pada uji coba 2 rendah. Maka, dilakukan proses pemilihan fitur terbaik untuk mengatasi kesalahan klasifikasi yang disebabkan oleh sering munculnya *term* pada seluruh dokumen. Dari hasil uji coba pemilihan fitur menggunakan 50 dan 30 *term* terbaik didapatkan akurasi sebesar 91% untuk metode usulan dan 95% untuk metode *Naïve Bayes* biasa. Hal ini dikarenakan *term* yang sering muncul pada kelas lain terdapat pula pada kelas yang diuji. Sedangkan ketika menggunakan 10 *term* terbaik didapatkan akurasi sebesar 94% untuk metode usulan dan 91% untuk metode *Naïve Bayes* biasa. Hal ini dikarenakan *term* yang digunakan pada kelas yang diuji merepresentasikan kelas tersebut. Sehingga pada uji coba ini diketahui bahwa pemilihan fitur terbaik dapat mengurangi jumlah *term* yang sering muncul pada kelas lain.

5. Kesimpulan

Metode *Weighted Naïve Bayes* dapat mengoptimalkan nilai akurasi metode *Naïve Bayes* biasa. Hal ini dapat dilihat dari hasil akurasi *Weighted Naïve Bayes* sebesar 94% dibandingkan dengan *Naïve Bayes* biasa sebesar 92%. *Weighted Naïve Bayes* dapat menghasilkan tingkat akurasi yang lebih tinggi dikarenakan setiap probabilitas dari atribut diberi bobot yang menghasilkan nilai yang lebih tinggi. Ketika dilakukan pemilihan fitur mengunkan 10 *term* terbaik didapatkan akurasi sebesar 94% untuk metode usulan dan 91% untuk metode Naïve *Bayes* biasa. Hal ini dapat disimpulkan bahwa pemilihan fitur dapat mengatasi kesalahan klasifikasi.

Daftar Pustaka

[1] U. S. F. dan W. Service, “Definitions of the Terms and Phrases of Amer-,” *English*, 2013. [Online]. Available: http://www.fws.gov/stand/defterms.html. [Accessed: 12-Dec-2015].
[2] L. Tenenboim, B. Shapira, and P. Shoval, “Ontology-based classification of news in an electronic newspaper,” *Inf. Syst.*, 2008.
[3] D. D. Lewis, *Naive(Bayes)at forty: The independence assumption in information retrieval*. 1998.
[4] D. J. Hand and K. M. Yu, “Idiot’s Bayes - Not so stupid after all?,” *Int. Stat. Rev.*, 2001.
[5] I. Konokenko, “Comparison of inductive and naive Bayesian learning approaches to automatic knowledge acquisition,” *Current trends Knowledge Acquisition*, pp. 190–197, 1990.
[6] P. Langley and S. Sage, “Induction of Selective Bayesian Classifiers,” *Proceedings Tenth International Conference on Uncertainty in Artificial Inteligence*, 1994.
[7] A. Hamzah, “Klasifikasi Teks Dengan Naïve Bayes Classifier (NBC) Untuk Pengelompokan Teks Berita Dan Abstract Akademis,” *Prosiding Seminar Nasional Aplikasi Sains dan Teknologi Periode III*, 2012.
[8] S. Garcia, “Search Engine Optimisation Using Past Queries,” School of Computer Science and Information Technology, 2007.
[9] P. Baldi, P. Frasconi, and P. Smyth, “Modeling the Internet and the Web: Probabilistic Methods and Algorithms,” *Information Processing and Management*, 2003.
[10] H. Zhang and S. Sheng, “Learning weighted naive bayes with accurate ranking,” in *Proceedings - Fourth IEEE International Conference on Data Mining, ICDM 2004*, 2004.
[11] J. Hilden and B. Bjerregaard, *Computer-aided diagnosis and the atypical case*. North Holland Publishing Co., 1976.
[12] J. T. A. S. Ferreira, D. G. T. Denison, and D. J. Hand, “Weighted naive Bayes modelling for data mining,” *citeseerx*, pp. 1–20, 2001.
[13] M. Hall, “A Decision Tree-Based Attribute Weighting Filter for Naive Bayes,” *ACM*, vol. 20, no. 2, pp. 120–126, 2007.
[14] S. Kullback and R. A. Leibler, “On Information and Sufficiency,” *The Annals of Mathematical Statistic*, vol. 22, no. 1, pp. 79–86, 1951.
[15] A. Renyi, “On Information and Sufficiency,” in *Proceedings of the 4th Berkeley*

symposium on Mathematics, 1961, pp. 547–561.

[16] N. Hermaduanti and S. Kusumadewi, “Sistem Pendukung Keputusan Berbasis SMS Untuk Menentukan Status Gizi Dengan Metode K-Nearest Neighbor,” in *Seminar Nasional Aplikasi Teknologi Informasi (SNATI)*, 2008, pp. 49–56.