Metode Optimasi *Pembobotan Gain Ratio* Terhadap Metode Klasifikasi *Weighted Naive Bayes* Studi Kasus Ulasan Produk *White Perfect Clinical Day Cream*

Khoirul Umam[1], Deni Arifianto, M.Kom[2]
Jurusan Teknik Informatika
Fakultas Teknik, Universitas Muhammadiyah Jember
Email : khoirulumam2032@gmail.com

Abstrak

Media memiliki peran penting dalam menyebarkan segala informasi. Seiring berjalannya waktu, media berkembang menjadi banyak jenis. Salah satu media yang sangat cepat berkembang adalah media online. Female Daily adalah sebuah blog pribadi dengan konten fashion dan kecantikan yang dikelola oleh Hanifa pada tahun 2005. Female Daily Network merupakan sebuah situs media informasi bagi wanita yang menyajikan konten seputar dunia wanita di Indonesia.

Salah satu metode klasifikasi yang sering digunakan adalah *Naïve Bayes Classifier* yang pertama kali dikemukakan oleh Revered Thomas Bayes. Tujuan dari penelitian ini dalah untuk mengetahui tingkat *accuracy, precision* dan *recall* klasifikasi *Naïve Bayes Classifier* dengan melakukan optimasi pembobotan *Gain Ratio*. Berdasarkan hasil penelitian dan pembahasan tentang metode *Naïve Bayes Classifier* dan *Weighted Naïve Bayes* yang digunakan dalam klasifikasi review produk kecantikan.

Dari hasil implemetasi metode *Weighted NaÎve Bayes* lebih akurat dari *NaÎve Bayes*. Berdasarkan hasil implementasi pada pengujian K-fold 2 skenario I dimana accuracy 48% precission 43% dan recall 95% maka dapat disimpulkan bahwa weighted naïve bayes dapat optimal jika data training atau latih sebanding jumlah data positif dan negative.

Kata Kunci : media online, female daily, klasifikasi naïve bayes, weighted naïve bayes.

1. PENDAHULUAN

1.1.Latar Belakang

Media memiliki peran penting dalam menyebarkan segala informasi. Seiring berjalannya waktu, media berkembang menjadi banyak jenis. Salah satu media yang sangat cepat berkembang adalah media online. Di dalam media online, terdapat banyak jenis media yang digunakan dalam menyebarkan informasi. Apalagi, dengan media online, semua orang dapat menggunakannya.

Female Daily adalah sebuah blog pribadi dengan konten fashion dan kecantikan yang dikelola oleh Hanifa pada tahun 2005. Female Daily Network merupakan sebuah situs media informasi bagi wanita yang menyajikan konten seputar dunia wanita di Indonesia. Dalam perjalananya kini, Female Daily Network telah mengembangkan 4 situs, yaitu Fashionese Daily, Mommies Daily, Clozette Daily dan Travelers Daily.

Banyak dari konten promosi ini yang akhirnya membantu para pengguna yang ingin memakai maupun mengenal produk baru yang diluncurkan ke pasar. Dari berbagai ulasan-ulasan tersebut salah satu kategori yang sangat digemari saat ini terutama oleh kaum wanita adalah mengenai konten kecantikan. Fenomena ini ikut memunculkan pula hadirnya Female Daily yaitu para pembuat konten yang mengkhususkan dirinya untuk berbagi informasi terkait dunia kecantikan.

Sentiment Analysis atau *Opinion Mining* baru-baru ini menjadi topik menarik yang mencoba untuk menggabungkan statistik, kecerdasan buatan dan teknologi. Data Mining dalam kerangka terpadu (Pang, dan Lee, 2008) *opinion mining* adalah informasi tekstual yang berada di dalam web dan berisi tentang fakta dan *opini*. *Opini* merupakan pernyataan subjektif yang mencerminkan persepsi seseorang terhadap sesuatu peristiwa, misalnya tentang opini-opini yang berkembang seperti krisis di Libya dan Suriah, perdebatan tentang krisis ekonomi di Yunani, dan downdrating atas kredibilitas Amerika Serikat adalah beberapa topik kontroversial yang dimuat dalam berita sehari-hari. Menganalisa rating movie untuk mengetahui tingkat pendapatan dari pemutaran suatu movie (Pang, Lee, dan Vaithyanathan, 2002). Salah satu metode klasifikasi yang sering digunakan adalah *Naïve Bayes Classifier* yang pertama kali dikemukakan oleh Revered Thomas Bayes. Penggunaan *Naïve Bayes Classifier* sudah dikenalkan sejak tahun 1702-1761. Menurut Lewis, Hand dan Yu, *Naïve Bayes Classifier* merupakan pendekatan yang sangat sederhana dan sangat efektif untuk pelatihan klasifikasi (Lewis, 1998) (Hand and Yu, 2001). Sedangkan Kononenko dan Langley menyimpulkan bahwa *Naïve Bayes Classifier* merupakan kemungkinan label kelas data atau bias diasumsikan sebagai atribut kelas yang diberi label (Kononenko, 1990) (Langley, 1994).

Menurut Hamzah Naïve Bayes memiliki beberapa kelebihan, yaitu algoritma yang sederhana, lebih cepat dalam penghitungan dan berakurasi tinggi (Hamzah, 2012). 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 optima.

Untuk data skala besar sangat dibutuhkan kecepatan dalam proses pencarian data. Sehingga dibutuhkan klasifikasi data terlebih dahulu. *Naive Bayes* merupakan algoritma pembelajaran

untuk klasifikasi dengan efisiensi komputasi dan akurasi yang baik, khususnya untuk dimensi dan jumlah data yang besar.untuk itu dalam penelitian ini akan membuktikan kemampuan *naïve bayes classifier* untuk mengklasifikasikan review produk kecantikan yang diperoleh dari *femaledaily*.

1.2.Rumusan Masalah

Berdasarkan latar belakang yang telah diuraikan, maka dapat diidentifikasikan beberapa rumusan masalah sebagai berikut:

1. Berapa tingkat akurasi *Naïve Bayes* tanpa menggunakan pembobotan *Gain Ratio* ?
2. Berapa tingkat akurasi *Naive Bayes* dengan melakukan optimasi dengan menggunakan pemboboan *Gain Ratio* ?

1.3.Batasan Peneltian

Agar penlitian tidak menyimpang dari topik peneltian maka dibuat batasan penelitian seperti berikut:

1. Dataset yang digunakan adalah review produk *White Perfect Clinical Day Cream* yang diperoleh dari *https://reviews.femaledaily.com* yang berjumlah 166 review produk dengan masing-masing label data positif 53 dan negatif 113 review produk yang di unduh pada periode 19 mei 2019.
2. Hasil klasifikasi yang diperoleh oleh sistem adalah Positif dan Negatif.
3. Hasil penelitian adalah perbandingan *naïve bayes* tanpa pembobotan *Gain Ratio* dan dengan pembobotan *Gain Ratio*.
4. Akurasi dilakukan dengan membandingkan klasifikasi manual dengan klasifikasi sistem.

1.4.Tujuan Peneltian

Tujuan dari penelitian ini adalah:

1. Menghitung tingkat akurasi klasifikasi sentiment menggunakan *Naïve Bayes Classifier*.
2. Menghitung tingkat akurasi klasifikasi *Naive Bayes Classifier* dengan melakukan optimasi dengan menggunakan pembobotan *Gain Ratio*.

1.5.Manfaat Penelitian

Manfaat yang diharapkan dari penelitian ini adalah:

1. Membantu perusahaan dalam mengetahui tingkat kepuasan terhadapat produk kecantikan yang dipasarkan.
2. Penulis dapat memahami sentiment analis dengan menggunakan metode Naïve Bayes.

2. TINJAUAN PUSTAKA

2.1.Naïve Bayes

Naïve Bayes Clasifier atau disebut juga dengan *Bayesian Classification* merupakan metode pengklasifikasian statistik yang didasarkan pada teorema bayes yang dapat digunakan untuk memprediksi probabilitas keanggotaan suatu kelas. *Bayesian Classification* terbukti memiliki akurasi dan kecepatan yang tinggi saat diaplikasikan ke dalam *database* yang besar.

Bentuk umum teorema bayes adalah sebagai berikut:

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

Dimana :

X = Data dengan kelas yang belum diketahui

H = Hipotesa data X merupakan suatu kelas spesifik

P(H|X) = Probabilitas hipotesis H berdasarkan kondisi X (posterior probability)

P(H) = Probabilitas hipotesis H (prior probability)

Peluang bersyarat atribut kategorikal dinyatakan dalam bentuk sebagai berikut:

$$P(A_i|C_j) = \frac{|A_{ij}|}{N_{cj}}$$

2

Dimana $|A_{ij}|$ adalah jumlah contoh pelatihan dari kelas A_i yang menerima nilai C_j. Jika hasilnya adalah nol, maka menggunakan pendekatan berikut:

$$P(A_i|C_j) = \frac{n_c + n_{equiv} p}{n + n_{equiv}}$$

3

Dimana n adalah total dari jumlah hasil dari kelas Cj. nC adalah jumlah contoh pelatihan dari kelas Ai yang menerima nilai C_j. n_{equiv} adalah nilai konstan dari ukuran sampel yang ekuivalen. P adalah peluang estimasi prior, $P = 1/k$ dimana k adalah jumlah kelas dalam variabel target.

Peluang bersyarat atribut kontinu dinyatakan dalam bentuk berikut:

$$P(A_i|C_j) = \frac{1}{\sqrt{2\pi\sigma ij}} exp^{[-\frac{(A_i-\mu_{ij})^2}{2(\sigma_{ij})^2}]}$$

2

2.2.Weighted Naïve Bayes

Menurut Hilden, Ferreira, dan Hall pembobotan atribut kelas dapat meningkatkan pengaruh prediksi (Hilden and Bjerregaard, 1976), (Ferreira, Denison, and Han, 2001) dan (Hall, 2007). 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 wi pada setiap atribut. Sehingga didapatkan rumus untuk pembobotan Naïve Bayes dituliskan pada Persamaan (3).

$$P(y,x) = P(y) \prod_{i=1}^{\alpha} P(x_i|y)^{w_i}$$

(3)

Pembobotan dapat dirumuskan menggunakan Gain Ratio (Zhang and Sheng, 2004). Dimana dari setiap atribut Gain Ratio dikali jumlah data n kemudian dibagi dengan rata-rata Gain Ratio semua atribut.

$$w_i = \frac{GainRatio(i)}{\frac{1}{\alpha}\sum_{i=1}^{\alpha} GainRatio(i)}$$

(4)

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 (Kullback and Leibler, 1951), (Renyi, 1961). Selain MI, Entropy digunakan sebagai pembagi dari MI yang digunakan untuk menentukan atribut mana yang terbaik atau optimal. Penghitungan Mutual Information dituliskan pada Persamaan 5 (Kullback and Leibler, 1951), (Renyi, 1961).

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

(5)

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 (6).

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

(6)

Maka dari itu penghitungan Gain Ratio adalah hasil dari penghitungan Mutual Information dibagi dengan hasil penghitungan Entropy Penghitungan Gain Ratio dituliskan pada Persamaan (7).

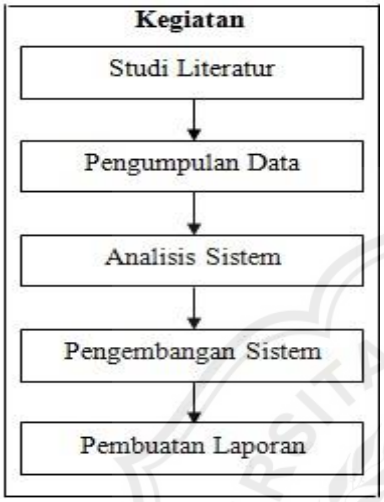
$$GainRatio(i) = \frac{MI(x_i,y)}{E(x_i)} = \frac{\sum_y \sum_{x_1} P(x_i,y) log \frac{P(x_1,y)}{P(x_1)P(y)}}{\sum_{x_1} P(x_1) log \frac{1}{P(x_1)}}$$

(7)

3. METODOLOGI PENELITIAN

3.1.Metode Penelitian

Untuk membantu dalam penyusunan penelitian ini, maka perlu adanya susunan kerangka kerja (frame work) yang jelas tahapan-tahapannya. Kerangka kerja ini merupakan langkah-langkah yang akan dilakukan dalam penyelesaian masalah yang akan dibahas. Adapun kerangka kerja penelitian yang di gunakan seperti terlihat pada gambar 3.1 :



Gambar 3.1 Metode Penelitian

Berdasarkan kerangka kerja penelitian yang telah digambarkan di atas, maka dapat diuraikan pembahasan masing-masing tahap dalam penelitian adalah sebagai berikut :

1. Studi Literatur
   Pada tahap ini dilakukan pencarian landasan-landasan teori yang diperoleh dari berbagai buku dan juga internet untuk melengkapi perbendaharaan konsep dan teori, sehingga memiliki landasan dan keilmuan yang baik dan sesuai.
2. Pengumpulan Data
   Pada tahap ini dilakukan proses pengumpulan data dengan mengunduh komentar atau testimoni dari website *femailedaily.com*.
3. Analisis Sistem
   Pada tahap ini dilakukan identifikasi masalah pada sistem dalam sentiment analis. Dengan demikian, diharapkan peneliti dapat menemukan kendala-kendala dan permasalahan yang terjadi dalam penentuan sentiment analis.
4. Pengembangan Sistem
   Pada Tahap ini dilakukan Pengembangan sistem dengan menggunakan model *waterfall.*
5. Pembuatan Laporan
   Pada tahapan ini dilakukan pembuatan laporan yang disusun berdasarkan hasil penelitian dengan menggunakan teknik pengumpulan data primer dan sekunder sehingga menjadi laporan penelitian yang dapat memberikan gambaran secara utuh tentang sistem yang sedang dibangun.

Penelitian ini akan mengusulkan sebuah metode baru untuk melakukan klasifikasi terhadap teks. Dengan menggunakan pembobotan *Gain Ratio* untuk menyeleksi fitur, dan algoritma *Naïve Bayes Classifier* (NBC) sebagai algoritma klasifikasi. Sedangkan metode pemilihan sampel menggunakan stratified sampling.

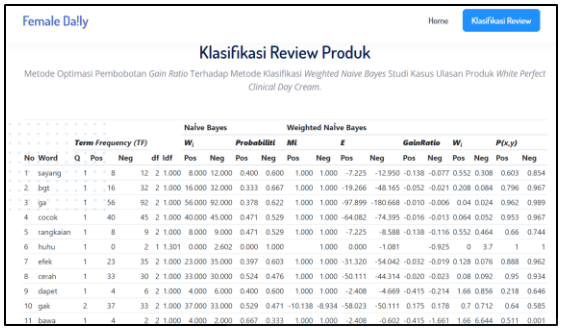
Model desain ini akan melakukan pemrosesan data training dan testing untuk menguji metode algoritma yang digunakan. Tahapan yang akan dilalui dibagi menjadi tiga bagian, yaitu *preprocessing*, seleksi fitur (*feature selection*), dan validation yang di dalamnya berisi *sub proses training* dan testing juga *performance measure* (lihat Gambar 3.2).

4. IMPLEMENTASI SISTEM

4.1.Implementasi Sistem

Pada tahap ini dilakukan pemodelan data, metode yang dipakai pada penelitian ini adalah probabilitas (prediksi) dengan menggunakan algoritma *Naive Bayes* dan *Weighted NaÎve Bayes*. Data yang telah dikumpulkan, diseleksi dan ditransformasi akan dikelola menggunakan probabilitas. Metode ini dapat digunakan dalam memprediksi peluang di masa depan berdasarkan pengalaman di masa sebelumnya sebagai perbandingan.

Data yang akan diujikan dibagi menjadi dua bagian yaitu training dan testing. Data review memiliki 100 record, dengan skema pengujian yang beda-beda yang sudah dijabarkan pada baba sebelumnya.



		Term Frequency (TF)				Naive Bayes				Weighted Naive Bayes									
						W_i		Probabilitii		Mi		E		GainRatio		W_i		P(x,y)	
No	Word	Q	Pos	Neg	df	Idf	Pos	Neg	Pos	Neg	Pos	Neg	Pos	Neg	Pos	Neg	Pos	Neg	Pos	Neg
1	sayang	1	8	12	2	1.000	8.000	12.000	0.400	0.600	1.000	1.000	-7.225	-12.950	-0.138	-0.077	0.552	0.308	0.603	0.854
2	bgt	1	16	32	2	1.000	16.000	32.000	0.333	0.667	1.000	1.000	-19.266	-48.165	-0.052	-0.021	0.208	0.084	0.796	0.967
3	ga	1	56	92	2	1.000	56.000	92.000	0.378	0.622	1.000	1.000	-97.899	-180.668	-0.010	-0.006	0.04	0.024	0.962	0.989
4	cocok	1	40	45	2	1.000	40.000	45.000	0.471	0.529	1.000	1.000	-64.082	-74.395	-0.016	-0.013	0.064	0.052	0.953	0.967
5	rangkaian	1	8	9	2	1.000	8.000	9.000	0.471	0.529	1.000	1.000	-7.225	-8.588	-0.138	-0.116	0.552	0.464	0.66	0.744
6	huhu	1	0	2	1	1.301	0.000	2.602	0.000	1.000		1.000	0.000	-1.081		-0.925	0	3.7	1	1
7	efek	1	23	35	2	1.000	23.000	35.000	0.397	0.603	1.000	1.000	-31.320	-54.042	-0.032	-0.019	0.128	0.076	0.888	0.962
8	cerah	1	33	30	2	1.000	33.000	30.000	0.524	0.476	1.000	1.000	-50.111	-44.314	-0.020	-0.023	0.08	0.092	0.95	0.934
9	dapet	1	4	6	2	1.000	4.000	6.000	0.400	0.600	1.000	1.000	-2.408	-4.669	-0.415	-0.214	1.66	0.856	0.218	0.646
10	gak	2	37	33	2	1.000	37.000	33.000	0.529	0.471	-10.138	-8.934	-58.023	-50.111	0.175	0.178	0.7	0.712	0.64	0.585
11	biasa	1	4	2	2	1.000	4.000	2.000	0.667	0.333	1.000	1.000	-2.408	-0.602	-0.415	-1.661	1.66	6.644	0.511	0.001

Gambar 4.2 Halaman Perhitungan Naïve Bayes dan Weight Naïve Bayes

Dari hasil implementasi *NaÎve Bayes* dan *Weighted NaÎve Bayes* data testing dihasilkan Naïve Bayes Positif = 15.547 dan Negatif = 11.453 maka hasil Naïve Bayes diklasifikasi menjadi Positif dan sedangkan *Weighted NaÎve Bayes* Positif = 20.145 dan Negatif = 20.649 maka hasil *Weighted NaÎve Bayes* diklasifikasi menjadi Negatif.

5. KESIMPULAN DAN SARAN

5.1.Kesimpulan

Dari hasil implementasi dan pengujian yang dilakukan pada bab sebelumnya maka didapatkan kesimpulan sebagai berikut:

1. Dari hasil implemetasi metode *Weighted NaÎve Bayes* lebih akurat dari *NaÎve Bayes* berdasarkan hasil implementasi pada bab sebelumnya.
2. Dari hasil implemetasi maka dapat disimpulkan bahwa metode Weighted NaÎve Bayes lebih akurat dari NaÎve Bayes jika data training (latih) sebanding dengan jumlah data positif dan negatif. Berdasarkan K-fold 2 skenario I dimana accuracy 48%, precision 43% dan recall 95%.

5.2.Saran

Saran untuk penelitian berikutnya agar sistem dapat lebih baik sebagai berikut:

1. System dapat dikembangkan dengan memanfaatkan curl dengan membaca review secara langsung dari *http://femaledaily.com/*.
2. Produk yang digunakan bisa lebih dari satu produk.

DAFTAR PUSTAKA

Basari, A.S.H., Hussin, B., Ananta, I.G.P., Zeniarja, J. 2013. Opinion Mining of Movie Review using Hybrid Method of Support Vector Machine and Particle Swarm Optimization. Procedia Engineering 53 (2013) 453 – 462. Malaysia.

Bramer, M., 2007, *Principles of Data Mining*. Springer, London.

Chaovalit, P., dan Zhou, L., 2005, Movie Review Mining: a Comparison between Supervised and Unsupervised Classification Approaches, IEEE, pp. 1-9.

Fadillah Z. Tala, A Study of Stemming Effect on Information Retrieval in Bahasa Indonesia, Netherland, Universiteit van Amsterdam, http://ucrel.lancs.ac.uk/acl/P/P00/P00-1075.pdf, diakses terakhir tanggal 25 Juli 2009.

Ferreira, J, T, A, S., Denison, D, G, T., dan Hand, D, J., 2001, "Weighted naive Bayes modelling for data mining," citeseerx, pp. 1–20, 2001.

Fish, U.S., Dan Service, W., 2013, *Definition Of Terms And Phrases*, *Http://Www.Fws.Gov/Stand/Devterms.Html*, Diakses 1 Maret 2019.

Hamzah, A., 2012, "Klasifikasi Teks Dengan Naïve Bayes Classifier (NBC) Untuk Pengelompokan Teks Berita Dan Abstract Akademis," Prosiding Seminar Nasional Aplikasi Sains dan Teknologi Periode III.

Hand, D, J., & YU, K., 2001, *Idiot's Bayes: Not So Stupid after All?. International Statistical Review*, 69 (3), 385-398.

Hilden, J., Bjerregaard, B., 1976, Computer-aided diagnosis and the atypical case. North Holland Publishing Co.

Kotler, P., & Keller, K, L., 2012, Marketing management. Upper Saddle River, N.J: Pearson Prentice Hall.

Langley dan Sage, S., 1994, *Induction of Selective Bayesian Classifier. Proceeding of The Tenth Conference on Uncertainty in Artificial Intelligence*. Morgan Kaufmann, US.

Lewis, D., 1998. *Naïve Bayes at forty: The independence assumption in information retrieval. Proceedings of the Tenth European Conference on Machine Learning*. April, Berlin, Germany. 4-15.

Nazief, and M. Adriani, 1996. Confix-stripping: Approach to stemming algorithm for Bahasa Indonesia. Internal publication, Faculty of Computer Science, University of Indonesia, Depok, Jakarta.

Ohana, B., Tierney, B., 2011, Supervised Learning Methods for Sentiment Classification with RapidMiner, RapidMiner Community Meeting And Conference, RCOMM, pp.

Pang, B., dan Lee, L., 2008, Opinion mining and sentiment analysis, Foundations and Trends in Information Retrieval, vol. 2, no. 1-2, pp. 1-135.

Pang, B., Lee, L., dan Vaithyanathan, S., 2002, "*Thumbs up? Sentiment Classification using Machine Learning Tehniques*," Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 79-86.

Saputra, A., 2011, Trik dan Solusi Jitu Pemrograman PHP, PT. Elex Media Komputindo, Jakarta.

Socrates, A, G., Akbar A, L., dan Akbar, M, S., 2016, Optimasi Naïve Bayes Dengan Pemilihan Fitur dan Pembobotan Gain Ratio, Surabaya, Institut Teknologi Sepuluh November.

Stylios, G., et al., 2010, *Public Opinion Mining for Governmental Decisions*, Electronic Journal of eGovernment, vol. 8, no. 2, pp. 202-213.

Zhang, H., and Sheng, S., 2004, "Learning weighted naive bayes with accurate ranking," in Proceedings - Fourth IEEE International Conference on Data Mining, ICDM 2004.