Computer Based Information System Journal

ISSN (Print): 2337-8794 | E- ISSN : 2621-5292
web jurnal : http://ejournal.upbatam.ac.id/index.php/cbis

SELEKSI FITUR INFORMATION GAIN DAN ALGORITMA NAÏVE BAYES UNTUK REVIEW OPINI KONSUMEN

Nurfaizah[1)], Nandang Hermanto[2)], Yanuar Ibnu Romadon[3)]

[1,3)]*Program Studi Sistem Informasi, Fakultas Ilmu Komputer, Universitas Amikom Purwokerto*
[2)]*Program Studi Informatika, Fakultas Ilmu Komputer, Universitas Amikom Purwokerto*

INFORMASI ARTIKEL

Diterima Redaksi: Juli 2020
Diterbitkan *Online*: September 2020

KATA KUNCI

Information Gain, Naïve Bayes, Review Opini

KORESPONDENSI

E-mail: nurfaizah@amikompurwokerto.ac.id

A B S T R A C T

The growth of internet users in Indonesia is increasing, this is in line with online shopping habits or often referred to as e-commerce which continues to increase. Various things are done by e-commerce companies to maintain customer loyalty, one of which is through product evaluation using consumer opinion reviews. The number of reviews that are too many will be biased, so it is necessary to do a classification method that will help e-commerce companies to find out the extent of their customer loyalty. Consumer review becomes something important because all assessments of the products they buy are all in the review column. In this research, a consumer review is carried out using the Naive Bayes classification method and to improve the accuracy of attributes using the Information Gain feature selection and using the Select by Weight operator which will display the best attributes of the pre processing process. The review data set is taken from consumers' comments on Google Play. The results of this study are classifying consumer reviews into positive reviews and negative reviews with Cross Validation using 10 fold, the accuracy of the Naive Bayes method is 78.4% using the Information Gain feature selection method, the accuracy increases to 81.2%

I. Latar Belakang

Salah satu tempat belanja yang gemari saat ini oleh masyarakat adalah melalui situs online marketplace [1]. Situs jual beli online memeberikan kemudahan bagi para konsumen karena dapat diakses selama 24 jam. Hal tersebut juga didukung dengan perkembangan penggunaan teknologi internet di Indonesia, berdasarlam hasil survey yang dilakukan Asosiasi Penyelenggara Jasa Internet Indonesia pengguna internet terus meningkat[2].

Pada tahun 2020 jumlah online shopper di Indonesia dengan jumlah pengunjung web bulanan 71,5 juta. Tercatat sekitar 74 juta unduhan di Indonesia sampai akhir tahun 2018 dan menjadi aplikasi belanja online nomor satu di Google Play dan App Store. Google Play merupakan layanan konten digital yang berisi toko produk-produk online. Google Play dapat diakses melaui aplikasi android atau yang disebut dengan Play Store, melalui web dan Google TV. Layanan Google Play dilengkapi dengan adanya fitur berisi review dari para pengguna yang dapat digunakan untuk melihat review dari pengguna aplikasi.

Penelitian yang dilakukan oleh Farki [3] menyatakan bahawa hasil penelitian Online Customer Review (OCR) baik review maupaun rating menjadi fitur yang sangat penting saat ini dalam e-commerce karena minat pelanggan untuk membeli dipengaruhi OCR. Pengguna internet saat ini tergantung pada rekomendasi opini sebelum melakukan transaksi suatu produk, karena opini dari pengguna sebelumnya menjadi informasi baru dari produk tersebut.

Salah satu alat analisis kalimat opini konsumen adalah analisis text mining yang dapat mengklasifikasikan semua opini konsemen ke dalam suatu kesimpulan positif, negatif maupun opini netral terhadap suatu produk yang mereka jual. Hal ini dapat mempermudah perusahaan membuat sistem pendukung keputusan atas produk tersebut.

Algoritma yang digunakan dalam penelitian ini menggunakan algoritma Naive Bayes. Algoritma yang popular dan banyak digunakan untuk mengklasifikasikan suatu teks, dimana Naïve Bayes memiliki performa yang baik pada banyak domain [4]. Dalam penelitian lain yang dilakukan oleh [5] mengungkapkan bahwa penelitian tersebut mampu mengembangkan klasifikasi sentimen dari data pemilu presiden 2014 dari halaman facebook dengan menggunakan algoritma Naïve Bayes.

Selain menggunakan algoritma Naive Bayes, dalam mengklasifikasikan sentimen analisis pada text dalam dataset yang digunakan banyak terdapat atribut yang menjadi masalah, sehingga perlu menggunakan metode seleksi fitur untuk mengurangi atribut yang kurang relevan pada dataset, penelitian yang dilakukan oleh [6] hasil dari komparasi seleksi fitur, Information Gain memiliki hasil yang paling baik.

Penelitian ini membuat klasfikasi review opini pengguna yang diambil dari opini pada google play menggunakan algoritma Naïve Bayes dengan klasifikasi fitur yang digunakan yaitu Information Gain.

II. Kajian Literatur

Hidayatullah dalam penelitiannya telah membangun model untuk melakukan klasifikasi tweet terhadap tokoh public berdasarkan sentimen dan kategori dengan algoritma naïve bayes. Klasifikasi tweet pada penelitian ini diperoleh berdasarkan kombinasi antara kelas sentimen dan kelas kategori. Klasifikasi sentimen terdiri dari positif dan negatif sedangkan klasifikasi kategori terdiri dari kapabilitas, integritas, dan akseptabilitas [7].

Penelitian analisis sentiment dibidang politik juga dilakukan oleh saif penelitian dimulai dari dataset pelatihan, mengumpulkan tweets langsung dan melakukan sentimen dengan berbagai klasifikasi dengan kategori yang digunakan positif, negatif dan netral, kemudian memprediksi partai mana yang memiliki kemungikanan tinggi untuk memenangkan pemilu [8].

Penelitian yang dilakukan oleh [7] [5] telah mampu mengembangkan sistem klasifikasi sentimen berdasarkan data Pemilu presiden Indonesia 2014 dari Facebook Page menggunakan Naïve bayes. Algoritma Naïve Bayes merupakan lagoritma pendukung untuk melakukan klasifikasi. Penelitian komparasai algoritma juga dilakukan oleh [6] menghasilkan pengujian algoritma seleksi fitur terbaikuntuk setiap algoritma klasifikasi adalah Information Gain mendapat hasil terbaik untuk digunakan pada algoritma Naïve Bayes.

Perbedaan dengan penelitian ini akan membahas review opini konsumen yang menghasilkan nilai akurasi menggunakan algoritma Naive Bayes dan untuk mengurangi permasalahan pada atribut dalam penelitian ditambahakn seleksi fitur Information Gain.

III. Metodologi

Tahapan penelitian diperlukan sebagai kerangka dan panduan proses penelitian, sehingga rangkaian proses penelitian dapat dilakukan secara terarah, teratur dan sistematis.

Berdasarkan gambar 1 tahapan penelitian yang dilakukan sebagai berikut:

1. Proses Pengumpulan Dataset

Proses ini dilakukan dengan melakukan pengambil data *google play* yang nantinya akan dioleh menjadi dataset yang nantinya akan dijadikan awal dalam penelitian ini.

2. Text Processing

Tahap *Text Processing* merupakan tahap pengolahan *dataset* yang bertujuan untuk mempersiapkan dokumen teks yang tidak terstruktur menjadi data terstruktur yang siap untuk digunakan pada tahap selanjutnya. *Text Processing* dilakukan dengan menggunakan metode-metode berikut:

a. *Case folding* adalah tahap merubah semua huruf capital pada semua dokumen komentar menjadi huruf kecil. Tujuannya untuk menghilangkan *redudansi* data yang hanya berbeda pada hurufnya saja.
b. *Tokenizing* proses untuk merubah proses penguraian deskripsi yang semula berupa kalimat-kalimat menjadi kata-kata dan menghilangkan delimiter-delimiter seperti tanda titik (.), koma(,), spasi dan karakter angka yang ada pada kata tersebut.
c. *Filtering* yaitu tahap mengambil kata-kata penting dari hasil *token* dalam penelitian ini menggunakan *stoplist/stopword* agar kata-kata yang kurang penting dan sering muncul dalam suatu dokumen dibuang sehingga hanya menyisakan kata-kata yang penting.

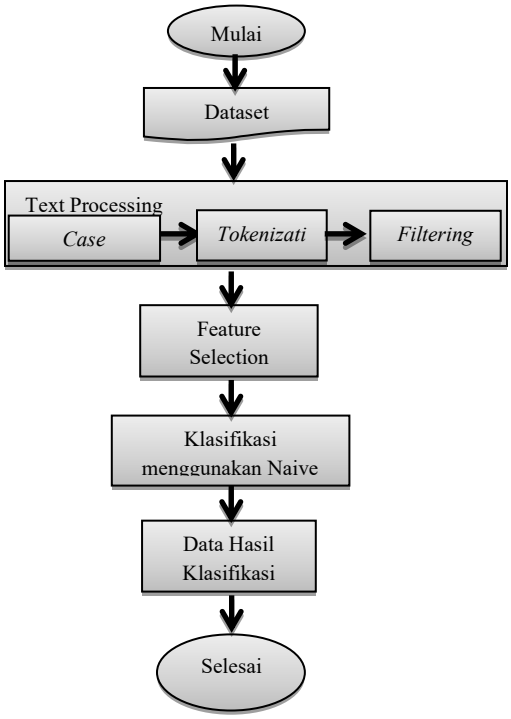
3. *Feature Selection Information Gain*

Dalam tahap ini dilakukan seleksi fitur dari data training yang digunakan sebelum dilkaukan proses klasifikasi.

4. Klasifikasi

Proses ini dilakukan untuk menguji akurasi dengan menggunakan algoritma naïve bayes.

Secara detail tahap penelitian yang dilakukan seperti pada gambar 1 berikut:



Gambar 1 Tahapan Penelitian

IV. Pembahasan

Dataset diperoleh dari data ulasan dari aplikasi google play melalui metode scraping. Setelah proses scrapping selesai dilakukan, langkah selanjutnya mengolah dataset dengan preprocessing, langkah yang perlu dilakukan selanjutnya adalah pembersihan data yang memiliki tujuan mengurangi dimensi-dimensi kata yang tidak memiliki pengaruh pada hasil pengolahan data. Karena hasil dari scraping memiliki bentuk kalimat atau teks yang tidak berstruktur yang memiliki banyak noise melalui tahap text prosesing.

Data ulasan sebanyak 250 yang telah diperoleh akan dibagi data menjadi dua yaitu sebagai data training dan data testing. Pada penelitian ini data yang menjadi data training sebesar 125 data ulasasan positif dan 125 ulasan negatif dan 100 data ulasan sebagai data testing.

Tahap pertama tahap pre prosesing terhadap text dataset yaitu tahap transform cases dalam tahap ini dilakukan penyamaan bentuk huruf dengan menggunakan transform cases, karena data yang diperoleh dalam komentar google play tidak sama mengandung huruf besar dan kecil pada proses ini huruf akan diubah menjadi huruf kecil, operator yang digunakan transform cases. Selanjutnya text dipisah menggunakan tokenisasi. Setelah text berubah menjadi kata dilakukan proses stopword Bahasa Indonesia untuk menghilangkan kata yang tidak penting pada dataset.

Data review produk yang digunakan sebanyak 125 pada google play, sebanyak 122 data diprediksi sesuai yaitu negatif, 51 data diprediksi negatif tetapi ternyata positif, 3 data diprediksi positif ternyata negatif dan 74 data sesuai dengan prediksi positif.

Perhitungan nilai dari akurasi dilakukan dengan menggunakan rumus akurasi. Sehingga, untuk penghitungan nilai akurasi pada metode Naive Bayes sebagai berikut:

Accuracy: 81,20%			
	True Negative	True Positive	Class Precision
Prediction Negative	101	23	81,45%
Prediction Positive	24	102	80,95%
Class recall	80,80%	81,60%	

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

$$Akurasi = \frac{(122 + 74)}{(122 + 3 + 51 + 74)}$$

$$Akurasi = \frac{196}{250} = 0{,}784 = 78{,}4\%$$

Detail Confution Matrix algoritma Naive Bayes dengan tingkat akurasi sebesar 78,40% seperti pada tabel 1 berikut:

Tabel 1
Confution Matrix Algoritma Naive Bayes

Accuracy: 78, 40%			
	True Negative	**True Positive**	**Class Precision**
Prediction Negative	122	51	70,52%
Prediction Positive	3	74	95,10%
Class recall	97,60%	59,20%	

Peningkatkan akurasi dalam penelitian ini selain menggunakan metode Naive Bayes, juga dilakukan seleksi fitur dengan menggunakan metode Information Gain, gambaran proses review opini keseluruhan dengan menggunakan metode Naive Bayes dan Information Gain seperti pada tabel 2 berikut:

Data review produk sebanyak 125 yang diambil dari google play, sebanyak 101 data diprediksi sesuai yaitu negatif, 23 data diprediksi negatif tetapi ternyata positif, 24 data diprediksi positif ternyata negatif dan 102 data sesuai dengan prediksi positif.

Proses perhitungan nilai dari akurasi dilakukan dengan menggunakan rumus akurasi. Sehingga, untuk penghitungan nilai akurasi pada metode Naive Bayes sebagai berikut:

$$Akurasi = \frac{(TN + TP)}{(TN + TF + TP + FP)}$$

$$Akurasi = \frac{(101 + 102)}{(101 + 24 + 23 + 102)}$$

$$Akurasi = \frac{203}{250} = 0{,}812 = 81{,}2\%$$

Detail Confution Matrix algoritma Naive Bayes dengan tingkat akurasi sebesar 81,20% seperti pada tabel 2 berikut:

Tabel 2
Confution Matrix Algoritma Naive Bayes Dan Information Gain

Implementasi metode menggunakan software rapidminer, dalam penelitian ini seleksi fitur Information Gain dengan Select by Weights dengan menampilkan kata paling banyak muncul parameter Weight Relation dipilih top k dengan k berisi 10. Selanjutnya melakukan visualisasi dengan tujuan yaitu untuk melihat informasi sentimen positif dan sentimen negatif serta, untuk mengekstraksi informasi berupa topik yang paling sering diulas atau dibahas oleh konsumen. Dari sekian banyak teks ulasan yang ada nantinya dapat diambil informasi yang dianggap paling penting.

Tabel distribusi dari algoritma Naive Bayes dengan 2 parameter yaitu mean dan standard deviation, tabel distribusi menggunakan parameter standard deviation seperti pada Tabel 3.

Tabel 3
Distribution Table Parameter Mean

Atribut	Negatif	Positif
bermasalah	0.011	0.001
jaringan	0.017	0.002
jaringan_bermasalah	0.010	0.000
maret_suka	0.000	0.008
suka	0.001	0.012
membantu	0.000	0.011
online	0.001	0.014
belanja	0.007	0.014
sukses	0.001	0.011
update	0.012	0.000

Sedangkan tabel distribusi algoritma Naive Bayes dengan parameter standard deviation seperti pada Tabel 4

Tabel 4
Distribution Table Parameter Standard Deviation

Atribut	Negatif	Positif
bermasalah	0.023	0.009
jaringan	0.027	0.010
jaringan_bermasalah	0.023	0.004
maret_suka	0.001	0.023
suka	0.006	0.027
membantu	0.001	0.030
online	0.010	0.027
belanja	0.013	0.017
sukses	0.005	0.026
update	0.030	0.005

Penelitian ini juga menghasilkan data revew terbanyak dari review opini produk yang dituliskan oleh konsumen, seperti pada gambar 2 berikut

Gambar 2 Review Terbanyak

Terdapat 10 kata review yang paling banyak muncul, kata bermasalah sebanyak 222, jaringan sebanyak 203, jaringan_bermasalah sebanyak 227, maret_suka 237 kali muncul, suka 222 kali muncul, kata membantu muncul sebanyak 233, kata online muncul sebanyak 218, kata belanja sebanyak 170 kali muncul, sukses muncul sebanyak 228 dan kata update muncul sebanyak 229 kali.

V. Kesimpulan

Algoritma Naïve Bayes merupakan algoritma pendukung untuk melakukan klasifikasi opini review konsumen. Dalam penelitian ini dilakukan dengan 2 tahap yaitu dengan melihat akurasi dari metode Naive Bayes. Selanjutnya untuk menghilangkan bias kata digunakan 2 metode yaitu Naive Bayes dengan seleksi fitur Information Gain.

Berdasarkan hasil uji diperoleh nilai akurasi sebesar 78,4% dengan menggunakan metode Naive Bayes, untuk meningkatkan nilai akurasi dengan menggunakan seleksi fitur Information Gain diperoleh tingkat akurasi sebesar 81,2%.

Ucapan Terima Kasih

Ucapan terima kasih kami sampaikan kepada Universitas Amikom Purwokerto yang telah mendanai kegiatan penelitian ini.

Daftar Pustaka

[1] Muljono, D. P. Artanti, A. Syukur, A. Prihandono, and D. R. I. M. Setiadi, “Analisa Sentimen Untuk Penilaian Pelayanan Situs Belanja Online Menggunakan Algoritma Naïve Bayes,” pp. 8–9, 2018.

[2] APJII, “Buletin APJII Edisi 22 - Maret 2018,” 2018.

[3] A. Farki, I. Baihaqi, and M. Wibawa, “Pengaruh Online Customer Review dan Rating Terhadap Kepercayaan dan Minat Pembelian pada Online Marketplace di Indonesia,” vol. 5, no. 2, 2016.

[4] L. D. Utami, “Integrasi Metode Information Gain Untuk Seleksi Fitur dan Adaboost Untuk Mengurangi Bias Pada Analisis Sentimen Review Restoran Menggunakan Algoritma Naïve Bayes,” vol. 1, no. 2, pp. 120–126, 2015.

[5] A. Rachmat and Y. Lukito, “Klasifikasi Sentimen Komentar Politik dari Facebook Page Menggunakan Naive Bayes,” J. Inform. dan Sist. Inf. Univ. Ciputra, vol. 2, no. 2, pp. 26–34, 2016.

[6] V. Chandani, F. I. Komputer, and U. D. Nuswantoro, “Komparasi Algoritma Klasifikasi Machine Learning Dan Feature Selection pada Analisis Sentimen Review Film,” vol. 1, no. 1, pp. 56–60, 2015.

[7] A. F. Hidayatullah and A. Sn, “Analisis Sentimen dan Klasifikasi Kategori Terhadap Tokoh Publik Pada Twitter,” Semin. Nas. Inform. 2014, vol. 2014, no. August 2013, pp. 0–8, 2014.

[8] H. Saif, “Semantic Sentiment Analysis in Social Streams,” vol. 29, pp. 229–233, 2017.