RAINSTEK
Jurnal Terapan Sains & Teknologi
Fakultas Sains dan Teknologi – Universitas PGRI Kanjuruhan Malang
ejournal.unikama.ac.id/index.php/jtst

e - ISSN: 2721-6209
Vol. 6, No. 3, 2024

Implementasi Naive Bayes Dengan Menggunakan Metode Laplace Smoothing

Yusril Adek Rizky, Abdul Aziz, Wahyudi Harianto
Universitas PGRI Kanjuruhan Malang
Email : yusriladek@gmail.com, abdul.aziz@unikama.ac.id, wahyudiharianto@unikama.ac.id

ABSTRAK

Abstrak. Algoritma Naïve Bayes memiliki kelebihan karena kesederhanaannya dan tingkat akurasi yang relatif tinggi dibandingkan dengan metode lainnya. Namun, Naïve Bayes juga memiliki kekurangan, seperti asumsi independensi antara atribut yang dapat mengurangi akurasi dan masalah probabilitas nol ("Zero Frequency"). Salah satu cara untuk mengatasi kekurangan metode Naïve Bayes adalah dengan menggunakan metode Laplace Smoothing. Metode ini membantu menghilangkan dampak probabilitas nol dan hasil akurasi dapat ditingkatkan. Dalam penelitian ini, nilai zero frequency dapat dihilangkan sepenuhnya dengan menggunakan Laplace Smoothing dan proses klasifikasi dibagi menjadi tiga skenario dengan perbandingan pengujian yang berbeda, yaitu 95% dan 5%, 80% dan 20%, serta 75% dan 25% data testing dan data uji .Hasil pengujian menunjukkan bahwa metode Naive Bayes mendapatkan nilai akurasi tertinggi sebesar 0,49, sedangkan Naive Bayes dengan metode Laplace Smoothing menghasilkan nilai akurasi tertinggi sebesar 0,69. Dengan demikian, dapat disimpulkan bahwa penggunaan Laplace Smoothing pada Naive Bayes dapat meningkatkan nilai akurasi dalam klasifikasi data.

Kata Kunci: *: Laplace Smoothing, Naïve Byaes, Klasifikasi, Teks minning, Zero Frequency, Probabilitas nol, analisis sentimen*

How to cite: Rizky, Y.A., Aziz, A., & Wahyudi, H. (2024). Implementasi Naive Bayes Dengan Menggunakan Metode Laplace Smoothing. *Jurnal Terapan Sains dan Teknologi*, *6* (2), 164-172. https://doi.org/ 10.21067/jtst.v6i3.9132

Pendahuluan

Bagian Klasifikasi merupakan proses menganalisis data dalam bentuk besar, seperti komentar di YouTube. Dalam proses klasifikasi, dilakukan ekstraksi informasi relevan dari teks dan analisis sebagian atau keseluruhan teks yang tidak terstruktur. Klasifikasi teks juga melibatkan penggunaan aturan tertentu [1]. Klasifikasi teks adalah model pengolahan data yang tidak terstruktur dan sulit ditangani, sehingga diperlukan proses text mining. Proses text mining diharapkan dapat mengeluarkan informasi yang jelas dari teks tersebut untuk keperluan analisis [2]. Oleh karena itu, text mining sangat mempermudah analisis sentimen.

Dalam penelitian analisis sentimen, diperlukan algoritma agar dapat mencapai akurasi maksimal. Salah satu algoritma yang umum digunakan untuk analisis sentimen adalah algoritma Naïve Bayes. Naïve Bayes adalah teknik pembelajaran algoritma data mining yang menggunakan metode probabilitas dan statistik. Dalam klasifikasi menggunakan Naïve Bayes, terdapat dua proses penting: pembelajaran (pelatihan) dan pengujian. Algoritma Naïve Bayes memiliki performa yang baik dibandingkan dengan model klasifikasi lainnya. Dalam jurnal oleh Nanang [3], yang berjudul "Analisis Sentimen terhadap Penerapan Sistem Ganjil/Genap pada Twitter

dengan Metode Naïve Bayes," disebutkan bahwa Naïve Bayes memiliki tingkat akurasi yang lebih tinggi.

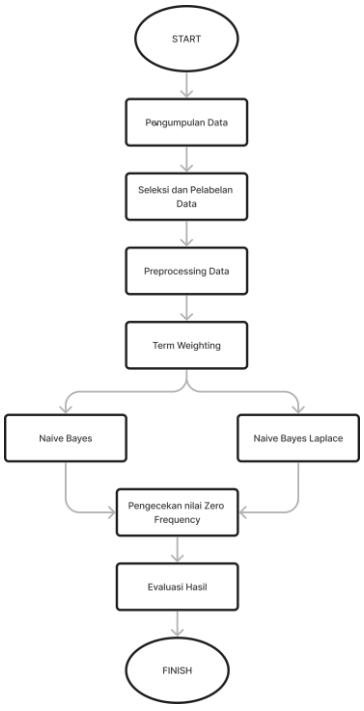
Kelebihan algoritma Naïve Bayes meliputi kesederhanaannya dan tingkat akurasi yang lebih tinggi dibandingkan metode lain. Algoritma Naïve Bayes yang sederhana dan proses pelatihan serta klasifikasinya yang cepat menjadikannya metode menarik dalam klasifikasi [3]. Namun, Naïve Bayes juga memiliki kekurangan, seperti asumsi independensi antar atribut yang dapat mengurangi akurasi dan ketidakberlakuan probabilitas jika ada nilai nol ("Zero Frequency") [4]. Kekurangan ini dapat diatasi dengan menggunakan metode Smoothing, sehingga probabilitas nol dapat diminimalisir dan hasil akurasi lebih tinggi [5]

Sebagai contoh penelitian sebelumnya, [6] telah membahas "Klasifikasi Pola Asuh Orang Tua terhadap Anak Usia Dini." Penelitian ini menggunakan 30 data yang sudah dilabeli, dan hasilnya dibandingkan. Metode Naïve Bayes dengan Laplace Smoothing dalam penelitian ini menghasilkan akurasi yang lebih baik daripada Naïve Bayes tanpa Laplace Smoothing.

Berdasarkan tinjauan pustaka, metode Naïve Bayes dengan teknik laplace smoothing memiliki tingkat akurasi yang tinggi dalam klasifikasi teks. Penelitian ini bertujuan untuk membuktikan bahwa dengan menggunakan teknik laplace smoothing, tingkat akurasi metode Naïve Bayes dapat lebih maksimal. Data dalam penelitian ini akan diambil dari komentar pada saluran YouTube "Detective Aldo," yang membahas tentang Metaverse.

Metode Penelitian

Penelitian ini menggunakan metode penelitian kuantitatif, yang didasarkan pada filsafat positivisme. Metode ini digunakan untuk menguji hipotesis dengan melakukan pengukuran yang teliti terhadap variabel yang diteliti. Prosesnya melibatkan pengumpulan data dalam bentuk angka dan pengolahan statistik[10]. Rancangan penelitian yang digunakan dalam penelitian ini adalah sebagai berikut:



Gambar 1. 1 Rancangan Penelitian

Pengumpulan Data

Proses pengumpulan data dalam penelitian ini bertujuan untuk memperoleh informasi dan fakta yang relevan yang akan diuji. Setelah data terkumpul, akan dilakukan proses pengolahan dan analisis menggunakan metode Naïve Bayes. Data yang digunakan dalam penelitian ini berasal dari komentar video YouTube berjudul "METAVERSE BAKAL BIKIN BUMI MAKIN NGERI?" yang diunggah oleh kanal Detective Aldo. Pengumpulan data dilakukan melalui teknik web scraping, yakni pengambilan data komentar YouTube yang dilakukan dengan menggunakan pemrograman dalam bahasa Python.

Seleksi dan Pelabelan Data

Tahap ini melibatkan seleksi data yang akan digunakan. Data yang tersedia di seleksi untuk menentukan mana yang akan disimpan untuk penggunaan dalam tahap selanjutnya, sambil menghapus data duplikat dan yang tidak relevan. Data yang berhasil dikumpulkan kemudian diurai menjadi kalimat-kalimat. Selanjutnya, dilakukan pelabelan data dengan kategori positif dan negatif.

Preprocessing Data

Preprocessing data adalah tahapan yang bertujuan untuk membersihkan kata-kata yang tidak diperlukan atau tidak memiliki makna. Tahap ini dilakukan berdasarkan kondisi data yang telah diperoleh sebelumnya.

Term Weighting

Dengan menggunakan metode TF/IDF, dilakukan pengukuran pembobotan kata, yang akan mempengaruhi klasifikasi data uji selanjutnya. Berikut adalah langkah-langkah dalam proses pembobotan kata menggunakan TF-IDF:

1. Hitung frekuensi kemunculan kata (TF) di semua dokumen.
2. Hitung nilai Inverse Document Frequency (IDF) menggunakan rumus yang sesuai.
3. Terapkan proses TF-IDF dengan memasukkan nilai IDF ke dalam kolom dengan nilai 1 pada masing-masing dokumen.
4. Selanjutnya, hitung bobot untuk setiap kata dalam dokumen.
5. Hasil yang dihasilkan akan digunakan sebagai dasar untuk menghitung panjang vektor sebelum menerapkan perhitungan jarak pada Confusion Matrix.

Naïve Bayes

Metode Naïve Bayes merupakan pendekatan yang digunakan untuk mengklasifikasikan data komentar guna mendapatkan analisis sentimen. Dalam proses klasifikasi sentimen, data akan melalui tahapan preprocessing hingga pembobotan kata menggunakan TF-IDF. Setelah data berhasil melalui tahap pelatihan (training), langkah selanjutnya adalah menguji hasil klasifikasi menggunakan data uji (test) untuk mengevaluasi akurasi klasifikasi yang telah dilakukan. Berikut adalah langkah-langkah yang perlu dilakukan :

1. Menghitung nilai prior
2. Menghitung nilai Likelihood
3. Menghitung V_{MAP}
4. Menentukan hasil Label

Laplace Smoothing

Laplace Smoothing dilakukan untuk mengatasi nilai frekuensi nol. Teknik Laplace Smoothing dilaksanakan dengan menambahkan nilai 1 pada setiap perhitungan data yang terdapat dalam set

data pelatihan (training set). Penambahan ini tidak akan secara signifikan mempengaruhi estimasi probabilitas, sehingga membantu menghindari nilai probabilitas yang nol.

Evaluasi Hasil

Evaluasi hasil dilakukan untuk mengukur performa model. Evaluasi hasil melibatkan analisis akurasi metode melalui confusion matrix dan tabel akurasi serta presisi untuk masing-masing model. Setelah data uji diuji dengan data pelatihan, hasilnya akan menghasilkan daftar kelas yang diprediksi dari data uji, yang disebut prediksi kelas. Prediksi kelas kemudian dibandingkan dengan kelas sebenarnya dari data uji yang sebelumnya disembunyikan. Dengan demikian, kita dapat melihat dan menghitung nilai akurasi (accuracy), presisi (precision), dan recall dari model tersebut.

Hasil dan Pembahasan

Data

Pengumpulan data dalam penelitian ini menggunakan teknik scrapping. Dengan data yang diambil berupa teks komentar dari video YouTube dari chanel Detective Aldo yang berjudul "METAVERSE BAKAL BIKIN BUMI MAIN NGERI?". Dalam scrapping ini berhasil mendapatkan 1534 data.

Preprocessing

Tahap preprocessing bertujuan untuk mengolah data mentah yang sudah didapatkan dari proses scrapping. Kemudian data akan melewati tahap Case Folding, Tokenizing, Filtering dan Stemming.

Tabel 1. 1 Hasil Preprocessing

Sebelum	Sesudah
Sebagus apapun virtual gak bisa ngalahin nyata	duniabagus apa dunia virtual gak bisa duniakalah dunia nyata

Term Weighting

Hasil Pada tahap ini dilakukan pembobotan dari suatu data. Kata yang didapatkan adalah kata dari hasil preprocessing. Pada tabel menghitung jumlah kata yang muncul dari setiap dokumen, sehingga dapat memberi nilai sesuai kondisi kemunculan suatu kata. Sebelum mencari TF-IDF, perlu mencari nilai TF,DF dan IDF.

Tabel 1. 2 TF-IDF

No	Term	TF			df	Idf
		D1	D2	D3		
1	Bill		1		1	2,8810
2	Gates		1		1	2,8810
3	Sindir		1		1	2,8810

4	Nft		1		1	2,8810
5	Duit		1		1	2,8810
6	Ditambang		1		1	2,8810
7	Orang	1	1	1	3	2,5800
8	bodoh		1		1	2,8810
Total						22,747

Pada sebagai contoh nilai TF adalah angka kemunculan term dan bernilai 1. Term "bill" hanya muncul dalam D2, maka jumlah term yang muncul akan dibagi jumlah kata dalam D2.

$TF = 0{,}5 + 0{,}5^{*}\frac{1}{7}$ = 0,5714

Kemudian menghitung nilai DF dengan cntoh term"bill" total kemunculannya sebanyak satu kali dalam D2. Maka nilai DF dalam term "bill" adalah 1.setelah nilai TF dan DF telah diketahui, kemudian menghitung nilai IDF

$$idf(t,d) = \log\left(\frac{1521}{1+1}\right)$$

Hasil dari idf dari term "bill" yaitu =2,8810

Sebagai cntoh nilai IDF didapatkan dari total kemunuclan dari suatu term. Contohnya kata "bill" muncul sebanyak satu kali, maka jumlah total data akan dibagi dengan keunculan term.

Setelah mengetahui nilai TF, DF dan IDF langkah selanjutnya dalah menghitung bobot dari nilai TF-IDF.

$TFIDF_{t,d} = 0{,}5714\ x\ 2{,}8810$ = 1,6462

Diatas merupakan contoh perhitungan TF-IDF. Dimana diambil contoh dari nilai TF dan IDF dari term "bill". Dan hasilnya bobot dari term "bill" menghasilkan nilai 1,6462.

Hasil bobot data tersebut akan digunakan untuk menghitung nilai probabilitas dalam tahap selanjutnya.

Naïve Bayes

Naive bayes memperhitungkan frekuensi setiap term yang muncul dalam suatu dokumen. Contoh pada perhitungan berikut yaitu data yang ada pada tabel 4.12. D2 memiliki kelas positif, D1 dan D3 adalah kelas negatif. Dari kata TF-IDF sebelumnya , didapatkan nilai keseluruhan W untuk kelas positif adalah W(+) = 13,3614 yang didapatkan dari total W D2, jumlah keseluruhan W pada kelas negatif adalah W(-) = 2,7934 yang didapatkan dari menambahkan total W D1 dan D3, dan jumlah keseluruhan idf pada seluruh kelas adalah B = 22,747

Tabel 1. 3 Bobot Kelas

Term	W	
	Poitif(c1)	**Negatif(c2)**
Bill	1,6462	0
Gates	1,6462	0
Sindir	1,6462	0
Nft	1,6462	0
Duit	1,6462	0
Ditambang	1,6462	0
Orang	1,8380	2,7934
bodoh	1,6462	0

Dari tabel bobot , selanjutnya mengitung nilai probabilitas sebagai berikut :

$P(c1) = \frac{1}{3} = 0{,}333$

$P(c2) = \frac{2}{3} = 0{,}666$

Setelah mendapatkan probabilitas prior, Kemudian menghitung likelihood. Berikut adalah contoh perhitungannya dari kata bill:

$$bill(c1) = \frac{1{,}6462}{13{,}3614 + 22{,}747} = 0{,}45590$$

$$bill(c2) = \frac{0}{2{,}7934 + 22{,}747} = 0$$

Tabel 1. 4 Hasil Bobot Naïve Bayes

P(w|c)	Positif	Negatif
Bill	0,45590	0
Gates	0,45590	0
Sindir	0,45590	0
Nft	0,45590	0
Duit	0,45590	0
Ditambang	0,45590	0
Orang	0,05090	0,10937
Bodoh	0,45590	0

Kemudian menghitung Vmap psitif dan negatif pada kalimat data testing untuk mendapatkan validasi apakah kalimat itu bernilai positif atau negatif. Sebagai contoh pada kalimat " bill gates sindir" :

Vmap(c1) = P(bill)*P(gates)*P(sindir)*P(C)

= 0,45590*0,45590*0,45590*0,333

=0,03155

Vmap(c2) = P(bill)*P(gates)*P(sindir)*P(C)

=0 *0*0*0,666

=0

Dalam kalimat “bill gates sindir” bernilai kalimat positif karena nilai positif lebih besar daripada nilai negatif.

Setelah itu dapat dinilai validasinya dengan cara mencocokan label sbenarnya dan label setelah proses klasifikasi. Jika sama,label sebenarnya Positif dan label setelah klasifikasi juga Positif maka bisa dinilai valid. Dan jika labelnya berbeda, label sebenarnya Positif dan label setelah klasifikasi Negatif maka dinilai tidak valid.

Laplace Smoothing

Disini Naive Bayes mengambil nilai bobot suatu kata dari TF,IDF dan TF-IDF. Rumus perhitungannya hampir sama seperti Naive Bayes. Tapi yang membedakan adalah menambah nilai 1 dalam pembobotan kelas positif dan negatif agar menghindari probabilitas nol. Berikut adalah cntoh perhitungannya dari kata bill:

$$bill(c1) = \frac{1{,}6462 + 1}{13{,}3614 + \ 22{,}747} = 0{,}0732$$

$$bill(c2) = \frac{0 + 1}{2{,}7934 + \ 22{,}747} = 0{,}0391$$

Nilai bobot probalitas yang dipakai diambil dari tabel 1.3.

Tabel 1. 5 Hasil Bobot NB+Laplace

P(w|c)	Positif	Negatif
Bill	0,0732	0,0391
Gates	0,0732	0,0391
Sindir	0,0732	0,0391
Nft	0,0732	0,0391
Duit	0,0732	0,0391
Ditambang	0,0732	0,0391
Orang	0,0785	0,1485
Bodoh	0,0732	0,0391

kemudian langkah selanjutnya sama seperti di metode Naïve Bayes.

Evaluasi Hasil

Dalam evaluasi hasil akan diketahui bepara hasil akurasi, presisi dan recall yang didapatkan pada masing-masing data uji dan data latih. Ketika dataset hanya memiliki dua kelas maka salah satu akan dianggap positif dan yang lain sebagai negatif. Dala confusion matrix ini dalam hasil positif akan menjadi True Positif dan False Negatif, sedangkan dalam hasil negatif akan menjadi True Negatif dan False Negatif.

Tabel 1. 6 Confusion Matrix Naïve Bayes

Data	Confusion Matrix				Akurasi	Presisi	Recall
	TP	FP	FN	TN			
75%	35	0	42	0	0,45	0,45	1,00
80%	144	0	161	0	0,47	0,47	1,00
95%	187	0	194	0	0,49	0.49	1,00

Tabel 1. 7 Confusion Matrix NB+Laplace

Data	Confusion Matrix				Akurasi	Presisi	Recall
	TP	**FP**	**FN**	**TN**			
75%	20	15	12	30	0,64	0,57	0,62
80%	87	57	44	117	0,66	0,60	0,64
95%	119	68	50	144	0,69	0,63	0,70

Penutup

Kesimpulan

Berdasarkan penelitian yang dilakukan, berikut merupakan kesimpulannya,

1. Penggunaan Laplace Smoothing dapat menghilangkan nilai Zero Frequency
2. Laplace Smoothing dapat meningkatkan nilai akurasi. Nilai akurasi tertinggi pada metode Naive bayes yaitu 0,49 dengan pengujian menggunakan data training 95% dan data testing 5%. Sedangkan dalam metde Naive Bayes Laplace mendapatkan nilai akurasi tertinggi sebesar 0,69 dengan data training 95% dan data testing 5%.

Dengan demikian dapat disimpulkan bahwa Laplace Smoothing dapat menghilangkan nilai Zero Frequency dan dapat meningkatkan nilai akurasi dari Naïve Bayes.

Saran

Untuk penelitian selanjutnya disarankan untuk menggunakan metode Smoothing yang lain seperti, Witten-Bell Smoothing, Additive Smoothing, Lidstone Smoothing dan sebagainya. Agar mengetahui teknik Smoothing yang lebih otimal serta kelebihannya.

Daftar Pustaka

I. Listiowarni. 2019. Implementasi Naïve Bayessian dengan Laplacian Smoothing untuk Peminatan dan Lintas Minat Siswa SMAN 5 Pamekasan. *Jurnal Sisfokom (Sistem Informasi dan Komputer)*, vol. 8, no. 2, pp. 124–129.

Ashyaksa, H. 2019. SENTIMENT ANALYSIS MENGGUNAKAN NAÏVE BAYES.

N. Ruhyana. 2019. ANALISIS SENTIMEN TERHADAP PENERAPAN SISTEM PLAT NOMOR GANJIL/GENAP PADA TWITTER DENGAN METODE KLASIFIKASI NAIVE BAYES.

Hinde, C.J., Stone, R., Xhemali, D., CC BY-NC-ND 4.0 REPOSITORY RECORD Naïve Bayes vs. Decision Trees vs. Neural Networks in the Classification of Training Web Pages. *IJCSI International Journal of Computer Science Issues*, Vol. 4, No. 1.

Listiowarni, I., Setyaningsih, E.R. 2018. *Analisis Kinerja Smoothing pada Naive Bayes untuk Pengkategorian Soal Ujian*.

FAJAR DARWIS DZIKRIL HAKIMI. 2018. SISTEM ANALISIS SENTIMEN PUBLIK TENTANG OPINI PEMILIHAN KEPALA DAERAH JAWA TIMUR 2018 PADA DOKUMEN TWITTER MENGGUNAKAN NAIVE BAYES CLASSIFIER.

R. Puspita and A. Widodo. 2021. Perbandingan Metode KNN, Decision Tree, dan Naïve Bayes Terhadap Analisis Sentimen Pengguna Layanan BPJS. *Jurnal Informatika Universitas Pamulang*, vol. 5, no. 4, p. 646.

Hakimi Fajar. 2018. SISTEM ANALISIS SENTIMEN PUBLIK TENTANG OPINI PEMILIHAN KEPALA DAERAH JAWA TIMUR 2018 PADA DOKUMEN TWITTER MENGGUNAKAN NAIVE BAYES CLASSIFIER.

Listiowarni. 2019. Analisis Kinerja Smoothing pada Naive Bayes untuk Pengkategorian Soal Ujian Indah