

Akurasi Sentimen Pada Komentar Youtube Untuk Permendikbud Tentang Pencegahan Dan Penanganan Kekerasan Seksual Berbasis Naïve Bayes Dan Support Vector Machine

Accuracy Of Sentiments On Youtube Comments For The Permendikbud Regarding The Prevention And Treatment Of Sexual Violence Based On Naïve Bayes And Support Vector Machine

Muhammad Derryl Qinanda ¹⁾, Agung Nilogiri^{2)*}, Taufiq Timur W³⁾

¹Fakultas Teknik, Universitas Muhammadiyah Jember
Email : derrylqinanda@gmail.com

²Dosen Fakultas Teknik, Univeritas Muhammadiyah Jember *Koresponden Author
Email : agungnilogiri@unmuhjember.ac.id

³Dosen Fakultas Teknik, Univeritas Muhammadiyah Jember
Email : taufiqtimur@unmuhjember.ac.id

Abstrak

RETRACTED

Perkembangan teknologi informasi membawa perubahan modern untuk masyarakat. Beberapa penggunaan media sosial yang digunakan untuk menjangkau dan menyampaikan informasi kepada masyarakat atau kerabat. Salah satu dari beberapa media sosial yang saat ini sering digunakan oleh masyarakat yaitu situs Youtube. Sentimen Analisis adalah teknik dimana mengekstrak sebuah data yang berbentuk teks yang digunakan untuk memperoleh sebuah informasi tentang sentimen bernilai positif dan negatif. Ruang lingkup penelitian ini dilakukan hanya pada komentar masyarakat terhadap permasalahan pencegahan dan penanganan kekerasan seksual di lingkungan perguruan tinggi pada channel MataNajwa. Selanjutnya data tersebut akan dilakukan pengoianan data menggunakan metode Naïve Bayes dan Support Vector Machine menggunakan ekstrasi fitur TF-IDF. Hasil akurasi yang didapatkan pada penelitian ini ialah 64% dengan menggunakan Naïve Bayes dan dilakukan pengujian dengan menggunakan unseen data test yang dipilih secara random menghasilkan akurasi sebesar 51,7%. Sedangkan nilai akurasi yang didapatkan saat menggunakan Support Vector Machine ialah sebesar 92% dengan dilakukan pengujian menggunakan unseen data test yang dipilih random menghasilkan nilai akurasi sebesar 62%.

Keywords: analisis sentimen, akurasi, naive bayes, support vector machine, youtube

Abstract

The development of information technology brings modern changes to society. Several uses of social media are used to obtain and convey information to the public or relatives. One of the several social media currently frequently used by the public is the Youtube site. Sentiment Analysis is a technique which extracts data in the form of text that is used to obtain information about positive and negative sentiments. The scope of this research was carried out only on public comments on the problem of preventing and handling sexual violence in the university environment on the MataNajwa channel. Furthermore, the data will be processed using the Support Vector Machine and Naïve Bayes methods with TF-IDF feature extraction. The accuracy results obtained in this study were 64% using Naïve Bayes and testing was carried out using an unseen data test that was selected randomly resulting in an accuracy of 51.7%. While the accuracy value obtained using the Support Vector Machine is 92% by testing using an unseen test data selected randomly resulting in an accuracy value of 62%.

Keywords: analisis sentimen, akurasi, naive bayes, support vector machine, youtube

1. PENDAHULUAN

Sentimen Analisis merupakan teknik dimana mengekstrak sebuah data yang berbentuk teks yang digunakan untuk memperoleh sebuah informasi yang bernilai positif dan negatif. Sentimen Analisis tersebut bisa didapatkan dari seseorang yang menggunakan internet untuk menyampaikan suatu komentar yang bersifat menilai atau opini pribadi (Sari, 2019). Melalui penelitian ini, akan dilakukan sebuah penelitian yang bertujuan untuk melakukan analisis sentimen terhadap komentar pengguna youtube di ruang lingkup penelitian ini dilakukan hanya pada komentar masyarakat terhadap permasalahan tindakan preventif dan penanganan kekerasan seksual pada lingkup perguruan tinggi. Pada tahun 2020, sebanyak 962 laporan yang diterima tentang kasus kekerasan seksual pada perguruan tinggi, selanjutnya pemerintah bersepakat untuk membuat peraturan dimana nantinya peraturan tersebut bisa berfungsi untuk melakukan pencegahan dan bisa melakukan penanganan pada kekerasan seksual di perguruan tinggi.

Selanjutnya data tersebut akan dilakukan pengolahan data menggunakan dua metode yaitu metode Naïve Bayes dan Support Vector Machine dengan ekstraksi fitur TF-IDF. Perlunya melakukan perbandingan antara hasil kerja metode Naïve Bayes dan Support Vector Machine dikarenakan terdapat beberapa artikel dengan judul diantaranya yaitu “Perbandingan Metode Naïve Bayes dan Support Vector Machine pada Analisis Sentimen Twitter” oleh Fiktri, dimana mengatakan bahwa hasil kerja metode Naïve Bayes didapatkan beberapa nilai yaitu nilai akurasi, presisi, dan recall dengan nilai yang lebih tinggi dibandingkan menggunakan metode Support Vector Machine. Selain itu, pada artikel dengan judul “Komparasi Algoritma Naive Bayes Dan Support Vector Machine Untuk Analisa Sentimen Review Film” oleh Elly Indrayuni dimana jurnal tersebut mengatakan bahwa nilai akurasi yang didapatkan oleh metode Naïve Bayes sebesar 84.50%, dan Support Vector Machine sebesar 90%.

Dengan perbedaan pendapat dari jurnal yang didapat maka akan dilakukan penelitian hasil kinerja antara metode Support Vector Machine

dan Naïve Bayes guna membuktikan adanya kemungkinan perbedaan hasil nilai rata – rata akurasi. Pada percobaan menggunakan kedua metode tersebut diharapkan mampu mendapatkan nilai yang lebih baik dibandingkan dengan penelitian yang sebelumnya, selain itu melalui penelitian ini dapat diketahui perbandingan kinerja hasil metode mana yang lebih baik antara algoritma Support Vector Machine dan Naïve Bayes berdasarkan nilai akurasi yang diperoleh.

2. TINJAUAN PUSTAKA

A. Pre-processing Text

Dilakukannya pre-processing pada data yang digunakan melalui tahapan seperti berikut:

1. Cleansing

Pada tahapan ini, semua karakter yang bukan alfabet akan dihapus sehingga mengurangi karakter yang tidak dikehendaki dan tidak memiliki arti dalam analisis sentimen.

2. Case Folding

Pada tahap ini, setiap huruf yang terdapat dalam kalimat akan diubah menjadi lowercase atau menjadi huruf kecil semua.

3. Tokenizing

Tahapan ini, komentar yang telah melalui tahapan cleansing dan case folding akan dipisahkan dari kalimatnya menjadi per kata atau token.

4. Stopword Removal

Tahapan ini ialah proses memilih kata yang penting dari hasil token yaitu kata-kata apa saja yang digunakan dalam dokumen atau disebut dengan filtering.

5. Stemming

Proses pada tahapan ini yaitu mengembalikan kata menjadi kata dasar dengan cara membuang awalan, akhiran atau sisipan atau bisa disebut dengan imbuhan. Proses ini dilakukan agar hanya kata dasar yang tersimpan dalam index database dalam sebuah dokumen.

B. TF-IDF

TF-IDF atau singkatan dari Term Frequency Inverse Document Frequency adalah suatu proses untuk melakukan transformasi data dari data tekstual ke dalam data numerik untuk dilakukan pembobotan pada tiap kata atau fitur.

Metode ini terkenal efisien, mudan dan akurat (Ria, 2018)

$$tf_{t,d} = \frac{n_{t,d}}{N} \quad (1)$$

Keterangan:

$n_{t,d}$ = nilai istilah yang muncul

N = total dokumen dalam aset

Tf = frekuensi kemunculan kata pada sebuah dokumen

$$idf_d = \log\left(\frac{N}{df}\right) \quad (2)$$

Keterangan:

N = total semua dokumen

df = banyak dokumen yang mengandung term tersebut

$$tfidf_{t,d} = tf_{t,d} \times idf_d \quad (3)$$

Keterangan:

TF-IDF = Term Frequency Inverse Document Frequency

TF = nilai TF

IDF = nilai IDF

C. K – Fold Cross Validation

K – Fold Cross-validasi adalah suatu teknik validasi untuk menilai bagaimana hasil statistik analisis akan menggeneralisasi kumpulan data independen. Teknik ini digunakan untuk melakukan prediksi model dan memperkirakan seberapa akurat sebuah model prediktif ketika dijalankan dalam praktiknya. Salah satu Teknik dari validasi silang adalah k-fold cross validation, dimana Teknik ini memecah data menjadi bagian set data dengan ukuran yang sama. Penggunaan k-fold cross validation untuk menghilangkan bias pada data. Pelatihan dan pengujian dilakukan sebanyak k kali.

D. Naïve Bayes

Naive Bayes adalah sebuah algoritma yang terdapat pada data mining dimana penggunaan metode Naïve Bayes untuk melakukan pemrosesan data memiliki waktu yang cepat, selain pemrosesan yang cepat, metode ini mudah untuk diimplementasikan dengan struktur yang cukup sederhana dan memiliki tingkat efektifitas yg tinggi (Fahriza, 2021). Berikut adalah persamaan teorema bayes sebagai berikut:

$$P(H | X) = \frac{P(X|H)P(H)}{P(X)} \quad (4)$$

Keterangan:

$P(H | X)$ = Data dengan class yang belum diketahui

H = Hipotesis data X yang merupakan suatu class spesifik

$P(H|X)$ = Probabilitas hipotesis H berdasarkan kondisi X (posteriori probabilitas)

$P(H)$ = Probabilitas hipotesis H (prior probabilitas)

$P(X|H)$ = Probabilitas X berdasarkan kondisi pada hipotesis H

$P(X)$ = Probabilitas X

E. Support Vector Machine

Support Vector Machine (SVM) merupakan metode yang bisa digunakan untuk menyelesaikan permasalahan secara linier maupun permasalahan non-linier. Dalam menyelesaikan permasalahan non-linier digunakan konsep kernel pada ruang kerja berdimensi tinggi, dengan mencari hyperplane yang dapat memaksimalkan margin antar kelas data. Hyperplane berfungsi dalam memisahkan 2 kelompok class +1 dan class -1 dimana setiap class memiliki pattern masing-masing (Rian, 2020).

Berikut adalah langkah-langkah metode Support Vector Machine:

- Menentukan kata pada dokumen dimana kata tersebut sering muncul atau komentar yang digunakan pada youtube.
- Menentukan inisialisasi awal dimana nilai $\alpha=0.5$, $C=1$, $\lambda=0.5$, $\gamma=0.5$ dan $\epsilon=0.001$.
- Melakukan perhitungan matriks dengan rumus (5):

$$D_{ij} = y_i y_j (K(\vec{x}_i \cdot \vec{x}_j) + \lambda^2) \quad (5)$$

Keterangan:

D_{ij} = anggota matriks data ke-ij

y_i = label atau kelas data ke-i

y_j = label atau kelas data ke-j

λ = turunan batasi teoritis

$K(\vec{x}_i \cdot \vec{x}_j)$ = fungsi kernel

- Bagi data ke $n=1,2,3,\dots,n$ gunakan persamaan (6)(7)(8) berikut ini :

$$E_i = \sum_{j=1}^i a_j D_{ij} \quad (6)$$

$$\delta a_i = \min \{ \max [\gamma (1 - E_i), -\alpha_i], C - \alpha_i \} \quad (7)$$

$$a_i = \alpha_i + \delta a_i \quad (8)$$

Keterangan:

- E_i = nilai error data ke- i
- Γ = tingkat pembelajaran
- $\max_{(i)} D_{ij}$ = nilai maksimum diagonal matriks hessian

e. Melakukan pencari nilai bias (b) menggunakan persamaan (10)

$$b = -\frac{1}{\sum w} [w \cdot x + w \cdot c] \quad (10)$$

f. Menguji pada dokumen yang akan diuji

g. Perhitungan keputusan

$$h(x) = \begin{cases} +1, & \text{if } w \cdot x + b \geq 0 \\ -1, & \text{if } w \cdot x + b < 0 \end{cases} \quad (10)$$

Jika hasil sign $h(x)$ dari perhitungan keputusan lebih dari sama dengan 0 maka hasil dari nilai tersebut adalah +1 dan termasuk ke dalam kelas positif dan sebaliknya, jika hasil sign $h(x)$ dari perhitungan keputusan kurang dari 0 maka hasil dari nilai -1 dan termasuk ke dalam kelas negatif. Penggunaan persamaan untuk mencari nilai keputusan seperti pada persamaan (11).

$$h(x) = w \cdot x + b \quad (11)$$

atau

$$h(x) = \sum_{i=1}^m a_i y_i K(x, x_i) + b$$

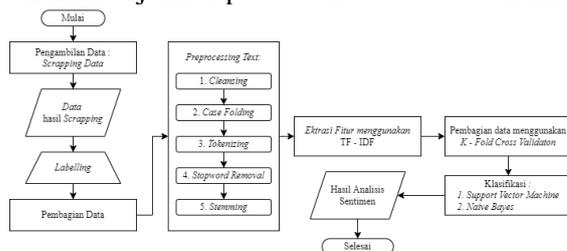
3. METODE PENELITIAN

A. Pengumpulan Data

Data yang digunakan pada penelitian ini yaitu data komentar Youtube. Dataset tersebut didapatkan dari channel Najwa Shihab dengan judul “Pro Kontra Permendikbud soal Kekerasan Seksual – (Part 4)”. Proses pengumpulan data tersebut dilakukan dengan menggunakan teknik scrapping pada bulan November tahun 2021

B. Penerapan Tahapan Penelitian

Tahapan penelitian dari proses awal hingga akhir ditunjukkan pada Gambar 1 di bawah ini



Gambar 1. Tahapan penelitian

Sumber: Hasil Penelitian

4. HASIL DAN PEMBAHASAN

Berikut adalah hasil nilai akurasi yang dihasilkan pada masing-masing metode.

a. Naive Bayes

Berikut adalah nilai akurasi yang didapatkan menggunakan metode Naive Bayes yang bisa dilihat pada **Tabel 1**.

Tabel 1. Nilai Akurasi Metode Naive Bayes

K-fold cross	Langkah Uji	Naive Bayes
2-fold	Langkah Uji 1	0,488
	Langkah Uji 2	0,512
	Langkah Uji 3	0,52
5-fold	Langkah Uji 1	0,52
	Langkah Uji 2	0,52
	Langkah Uji 3	0,56
	Langkah Uji 4	0,54
	Langkah Uji 5	0,46
10-fold	Langkah Uji 1	0,4
	Langkah Uji 2	0,56
	Langkah Uji 3	0,64
	Langkah Uji 4	0,44
	Langkah Uji 5	0,48
	Langkah Uji 6	0,6
	Langkah Uji 7	0,56
	Langkah Uji 8	0,6
	Langkah Uji 9	0,52
	Langkah Uji 10	0,44

Sumber: Hasil Perhitungan

b. Support Vector Machine

Berikut adalah nilai akurasi yang didapatkan menggunakan metode Support Vector Machine yang bisa dilihat pada **Tabel 2**.

Tabel 2. Nilai Akurasi Metode Support Vector Machine

K-fold cross	Langkah Uji	Support Vector Machine
2-fold	Langkah Uji 1	0,672
	Langkah Uji 2	0,68
5-fold	Langkah Uji 1	0,64
	Langkah Uji 2	0,82
	Langkah Uji 3	0,72
	Langkah Uji 4	0,74
10-fold	Langkah Uji 5	0,62
	Langkah Uji 1	0,64
	Langkah Uji 2	0,6
	Langkah Uji 3	0,92
	Langkah Uji 4	0,72

Langkah Uji 5	0,68
Langkah Uji 6	0,84
Langkah Uji 7	0,72
Langkah Uji 8	0,72
Langkah Uji 9	0,6
Langkah Uji 10	0,6

Sumber: Hasil Perhitungan

Langkah Uji 3	0,92	0,62
Langkah Uji 4	0,72	0,586
Langkah Uji 5	0,68	0,62
Langkah Uji 6	0,84	0,586
Langkah Uji 7	0,72	0,551
Langkah Uji 8	0,72	0,586
Langkah Uji 9	0,6	0,551
Langkah Uji 10	0,6	0,586

Sumber: Hasil Perhitungan

Berikutnya adalah pengujian data dimana pengujian data ini, masing-masing langkah uji pada setiap fold akan dilakukan menggunakan unseen data test dimana komposisi data tersebut adalah 15 positif dan 16 negatif. Berikut adalah tabel dari hasil pengujian data menggunakan unseen data baik metode Naïve Bayes dan Support Vector Machine.

Berdasarkan langkah pengujian data yang telah dilakukan, diperoleh rekapitulasi data hasil akurasi yang disajikan pada tabel berikut:

Tabel 3. Hasil Perhitungan pada metode Naïve Bayes dan Support Vector Machine

Tabel 3. Pengujian Unseen Data Test Naïve Bayes

K-fold cross	Langkah Uji	Naïve Bayes	
		Data Training	Data Unseen
2-fold	Langkah Uji 1	0,488	0,448
	Langkah Uji 2	0,512	0,551
5-fold	Langkah Uji 1	0,52	0,482
	Langkah Uji 2	0,52	0,551
	Langkah Uji 3	0,56	0,551
	Langkah Uji 4	0,54	0,586
	Langkah Uji 5	0,46	0,551
10-fold	Langkah Uji 1	0,4	0,517
	Langkah Uji 2	0,56	0,517
	Langkah Uji 3	0,64	0,517
	Langkah Uji 4	0,44	0,62
	Langkah Uji 5	0,48	0,551
	Langkah Uji 6	0,6	0,551
	Langkah Uji 7	0,56	0,551
	Langkah Uji 8	0,6	0,551
	Langkah Uji 9	0,52	0,482
	Langkah Uji 10	0,44	0,62

Sumber: Hasil Perhitungan

Tabel 4. Pengujian Unseen Data Test Support Vector Machine

K-fold cross	Langkah Uji	Support Vector Machine	
		Data Training	Data Unseen
2-fold	Langkah Uji 1	0,672	0,551
	Langkah Uji 2	0,68	0,551
5-fold	Langkah Uji 1	0,64	0,551
	Langkah Uji 2	0,82	0,551
	Langkah Uji 3	0,72	0,62
	Langkah Uji 4	0,74	0,586
	Langkah Uji 5	0,62	0,551
10-fold	Langkah Uji 1	0,64	0,62
	Langkah Uji 2	0,6	0,551

K-fold cross	Langkah Uji	Naïve Bayes		Support Vector Machine	
		Data Training	Data Unseen	Data Training	Data Unseen
2-fold	Langkah Uji 1	0,488	0,448	0,672	0,551
	Langkah Uji 2	0,512	0,551	0,68	0,551
5-fold	Langkah Uji 1	0,52	0,482	0,64	0,551
	Langkah Uji 2	0,52	0,551	0,82	0,551
	Langkah Uji 3	0,56	0,551	0,72	0,62
	Langkah Uji 4	0,54	0,586	0,74	0,586
	Langkah Uji 5	0,46	0,551	0,62	0,551
10-fold	Langkah Uji 1	0,4	0,517	0,64	0,62
	Langkah Uji 2	0,56	0,517	0,6	0,551
	Langkah Uji 3	0,64	0,517	0,92	0,62
	Langkah Uji 4	0,44	0,62	0,72	0,586
	Langkah Uji 5	0,48	0,551	0,68	0,62
	Langkah Uji 6	0,6	0,551	0,84	0,586
	Langkah Uji 7	0,56	0,551	0,72	0,551
	Langkah Uji 8	0,6	0,551	0,72	0,586
	Langkah Uji 9	0,52	0,482	0,6	0,551
	Langkah Uji 10	0,44	0,62	0,6	0,586

Sumber: Hasil Perhitungan

Tabel nilai akurasi diatas menunjukkan perolehan nilai akurasi tertinggi metode naïve bayes yaitu 64% dengan k=10 pada langkah uji ketiga, sedangkan untuk nilai akurasi tertinggi yang dihasilkan oleh metode support vector machine yaitu 92% dengan k=10 pada langkah uji ketiga. Dari hasil akurasi tersebut menunjukkan bahwa nilai yang dihasilkan oleh Support Vector Machine yaitu berupa nilai akurasi menghasilkan nilai yang lebih besar dari pada nilai akurasi yang dihasilkan oleh metode naïve bayes.

5. KESIMPULAN DAN SARAN

1. Pada penelitian ini yakni pada proses K-fold cross validation nilai akurasi tertinggi yang dihasilkan oleh metode naïve bayes berdasarkan langkah uji ketiga pada $K = 10$ yaitu 64%, sedangkan nilai akurasi tertinggi yang dihasilkan oleh metode Support Vector Machine (SVM) berdasarkan langkah uji ketiga pada $K = 10$ yaitu 92%. Untuk masing-masing metode dilakukan pengujian dengan menggunakan 29 unseen data test menghasilkan nilai akurasi sebesar 51,7% dan 62% pada langkah uji ketiga dengan $K=10$.
2. Metode Support Vector Machine (SVM) pada penelitian ini mampu menghasilkan nilai akurasi yang lebih baik daripada menggunakan metode Naïve Bayes, dengan rata-rata nilai akurasi yang diperoleh ialah 70,18% dibandingkan dengan 52%.

6. DAFTAR PUSTAKA

- Buntoro, Ghulam Asrofi. 2017. Analisis Sentimen Calon Gubernur DKI Jakarta 2017 Di Twitter. Universitas Muhammadiyah Ponorogo.
- Firman, T., Miftah, M., Amal, K. 2018. Perbandingan Klasifikasi Antara Knn Dan Naive Bayes Pada Penentuan Status Gunung Berapi Dengan K-Fold Cross Validation. Universitas Khairun Ternate
- Hakim, Sulton Nur. 2021. Analisis Sentimen Persepsi Pengguna Myindihome Menggunakan Metode Support Vector Machine (Svm) Dan Naïve Bayes Classifier (Nbc). Universitas Islam Indonesia
- Cahyono, A G. 2016. Pengaruh Media Sosial Terhadap Perubahan Sosial Masyarakat Di Indonesia.
- Sari. 2019. Analisis Sentimen Pelanggan Toko Online Jd.Id Menggunakan Metode Naïve Bayes Classifier Berbasis Konversi Ikon Emosi. Bandung
- Sunoto, Y. 2017. *Analisis Testimonial Wisatawan Menggunakan Text Mining Dengan Metode Naive Bayes Dan Decision Tree, Studi Kasus Pada Hotel – Hotel Di Jakarta*. Institut Bisnis Dan Informatika Kwik Kian Gie
- Tuhuteru. 2020. Analisis Sentimen Masyarakat Terhadap Pembatasan Sosial Berskala Besar Menggunakan Algoritma Support Vector Machine. Universitas Kristen Indonesia Maluku
- Luqyana, 2018. *Analisis Sentimen Cyberbullying Pada Komentar Instagram Dengan Metode Klasifikasi Support Vector Machine*. Universitas Brawijaya
- Ria, M. 2018. Penerapan Metode Term Frequency Inverse Document Frequency (Tf-Idf) Dan Cosine Similarity Pada Sistem Rekomendasi Informasi Untuk Mengetahui Syarat Hadit Berbasis Web (Studi Kasus: Syarah Umdatul Ahkani), Universitas Islam Negeri Syarif Hidayatullah Jakarta
- Rian, T. 2020. Analisis Sentimen Terhadap Layanan Indihome Berdasarkan Twitter Dengan Metode Klasifikasi Support Vector Machine (Svm). Universitas Nasional, Jakarta
- Fahriza, F. 2021. Klasifikasi Sentimen Terhadap Gubernur DKI Jakarta Di Media Sosial Twitter Menggunakan Metode Naive Bayes Classifier. Universitas Negeri Sultan Syarif Qasim Riau Pekanbaru
- Fikri, M. 2020. Perbandingan Metode Naïve Bayes Dan Support Vector Machine Pada Analisis Sentimen Twitter. Universitas Muhammadiyah Malang