Syntax:
Journal of Software Engineering, Computer Science and Information Technology
p-ISSN: 2776-7027, e-ISSN: 2723-0538
Volume: 5, Nomer: 1, Juni 2024

KLASIFIKASI OPINI MASYARAKAT TERHADAP VIDEO DOKUMENTER DIRTY VOTE DENGAN ALGORITMA KNN (K-NEAREST NEIGHBOR) DAN NAÏVE BAYES

Silvi Mutia[1], Andreas[2], Jendraja Husein Kotan[3], Hafidz Irsyad[4]

[1,2,3]*Universitas Multi Data Palembang*
Jl. Rajawali No.14, 9 Ilir, Kec. Ilir Tim. II, Kota Palembang, Sumatera Selatan 30113
[1]*silvimutia7@mhs.mdp.ac.id*, [2]*andreas14477@mhs.mdp.ac.id,*
[3]*jendrajahk03@mhs.mdp.ac.id*, [4]*hafizirsyad@mdp.ac.id*

ABSTRAK

Penelitian ini menggabungkan 2 algoritma yaitu algoritma K-Nearest Neighbor (KNN) dan Naïve Bayes dalam mengklasifikasikan opini masyarakat terhadap video dokumenter "Dirty Vote". KNN mengklasifikasikan data berdasarkan kemiripan dengan data yang ada, sementara Naïve Bayes menggunakan pendekatan probabilistik dengan asumsi independensi antar fitur. Tujuan penelitian ini adalah mengevaluasi efektivitas dan akurasi kedua algoritma dalam analisis sentimen. Hasil menunjukkan Naïve Bayes lebih mendapatkan tingkat akurasi sebesar 0.76. Kesimpulannya, Klasifikasi Opini Masyarakat terhadap Video Dirty Vote dengan menggunakan Algoritma KNN dan Naïve Bayes sangat penting dan bermanfaat untuk meningkatkan edukasi masyarakat mengenai pentingnya integritas pemilu dan mendorong partisipasi aktif dalam proses demokrasi, sehingga memperkuat sistem demokrasi dan memastikan suara masyarakat didengar dalam pengambilan keputusan politik.

Kata Kunci— **KNN, Naïve Bayes, Sentimen, Dirty Vote.**

ABSTRACT

This research combines 2 algorithms, namely the K-Nearest Neighbor (KNN) algorithm and Naïve Bayes in classifying public opinion on the documentary video "Dirty Vote". KNN classifies data based on similarity to existing data, while Naïve Bayes uses a probabilistic approach with the assumption of independence between features. The aim of this research is to evaluate the effectiveness and accuracy of the two algorithms in sentiment analysis. The results show that Naïve Bayes has a higher accuracy rate of 0.76. In conclusion, Classification of Public Opinion on Dirty Vote Videos using the KNN and Naïve Bayes Algorithms is very important and useful for increasing public education regarding the importance of election integrity and encouraging active participation in the democratic process, thereby strengthening the democratic system and ensuring that people's voices are heard in political decision making.

Keywords— **KNN, Naïve Bayes, Sentiment, Dirty Vote.**

Syntax:
Journal of Software Engineering, Computer Science and Information Technology
p-ISSN: 2776-7027, e-ISSN: 2723-0538
Volume: 5, Nomer: 1, Juni 2024

I. PENDAHULUAN

Dalam era digital saat ini, media sosial dan platform berbagi video telah menjadi saluran utama bagi masyarakat untuk mengakses dan mendiskusikan berbagai isu sosial dan politik. Salah satu isu yang sering menjadi sorotan adalah masalah integritas pemilu. Kepercayaan masyarakat terhadap proses pemilihan umum sangat krusial untuk memastikan legitimasi dan kelangsungan sistem pemerintahan.

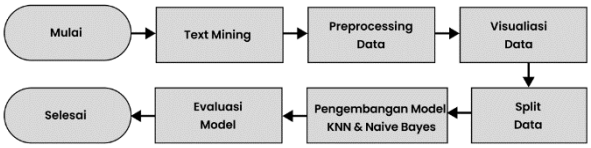
Namun, dalam beberapa tahun terakhir, kepercayaan ini sering kali terguncang akibat berbagai isu dan skandal, salah satunya terkait dengan kecurangan dalam pemilihan umum seperti yang diangkat dalam video dokumenter "Dirty vote". Dirty vote merupakan sebuah video dokumenter yang memuat hasil riset terhadap kecurangan pemilu dari tahun sebelumnya yang dimana para penonton dipaparkan oleh fenomena politik yang terjadi menjelang pemilu. Mulai dari ketidaknetralan pemerintah, anggaran dan penyeluran bansos, pelanggaran etik, dan lain-lain [1]. Dengan penyajian konten dan pemublikasian video melalui saluran youtube sebelum pemilu menimbulkan berbagai macam opini dari masyarakat , baik itu pendapat mendukung argumen dari video terkait atau pendapat menentang/ketidaksetujuan argurmen dari video terkait di kolom komentar. Dari tanggapan tersebut perlu digolongkan kedalam bentuk tanggapan positif atau negatif melalui analisis sentimen.

Analisis sentimen merupakan proses untuk mengidentifikasi dan mengklasifikasikan opini atau emosi dari teks yang ditulis oleh pengguna [2]. Dalam konteks video dokumenter "*Dirty Vote*", analisis sentimen dapat membantu untuk memahami bagaimana masyarakat menanggapi konten yang dipublikasikan. Hal ini sangat penting karena dapat memberikan wawasan tentang persepsi publik terhadap isu kecurangan pemilu dan integritas proses demokrasi. Dengan memahami sentimen masyarakat, pihak-pihak terkait dapat mengambil langkah yang lebih tepat dalam merespons kekhawatiran publik dan memperbaiki sistem pemilu yang ada.

Algoritma *K-Nearest Neighbor (KNN)* dan *Naïve Bayes* dalam klasifikasi opini memiliki pendekatan yang berbeda dan memiliki keunggulan serta keterbatasannya masing-masing. Algoritma KNN mencari kelompok k objek dalam data training yang paling dekat atau mirip dengan objek pada data atau data testing [3]. Metode ini sangat berguna dalam mengelompokkan opini masyarakat ke dalam kategori positif, negatif, atau netral berdasarkan fitur-fitur yang ada dalam teks komentar. Sedangkan, Algoritma *Naïve Bayes* adalah salah satu algoritma klasifikasi yang didasarkan pada asumsi "kemurnian" (naive) antara fitur-fitur yang digunakan dalam proses klasifikasi. Algoritma ini memanfaatkan perhitungan probabilitas kelas dan probabilitas fitur untuk memprediksi kategori atau abel yang tepat untuk data yang belum diklasifikasikan [4]. Penelitian ini bertujuan untuk menggabungkan kinerja antara algoritma *KNN* dan *Naïve Bayes* dalam mengklasifikasikan opini masyarakat terhadap video dokumenter "*Dirty Vote*". Dengan menggabungkan kedua metode ini, kita dapat mengevaluasi efektivitas dan akurasi masing-masing algoritma dalam konteks analisis sentimen. Penelitian ini juga bertujuan untuk mengatasi isu-isu yang mempengaruhi kepercayaan publik terhadap proses pemilihan umum. Selain itu, wawasan yang diperoleh dari analisis ini dapat digunakan untuk meningkatkan edukasi masyarakat mengenai pentingnya integritas pemilu dan mendorong partisipasi yang lebih aktif dalam proses demokrasi. Dengan demikian, penelitian ini berpotensi untuk memperkuat sistem demokrasi dan memastikan bahwa suara masyarakat didengar dan diperhatikan dalam pengambilan keputusan politik.

II. METODOLOGI PENELITIAN

Pada penelitian ini pengujian di lakukan dengan platform Google Colab sebagai IDE berbasis cloud dengan bahasa program Python. Pendekatan metode machine learning yang digunakan untuk melakukan klasifikasi sentimen, yaitu *K-Nearest Neighbor* dan *Naïve Bayes*. Pada Gambar 1. terdapat flowchart yang menunjukan beberapa tahapan akan di lakukan untuk menghasilkan hasil yang maksimal mulai dari text mining, preprocessing, visualisasi, pemisahan data, pembobotan, pembangunan model *KNN* dan *Naïve Bayes*, dan evaluasi.



Gambar. 1 Flowchart Tahap Penelitian

A. *Text Mining*

Text mining adalah proses untuk memperoleh informasi berkualitas tinggi dari teks [5]. Text mining juga masih bagian dari data mining dimana akan memproses data – data atau text – text serta dokumen – dokumen yang bisa jadi dalam jumlah sangat besar [6]. Langkah ini diperlukan untuk menyusun struktur data teks yang acak/tidak beraturan, dengan beberapa tahapan untuk menghasilkan teks yang lebih mudah dianalisis. Tahap labeling mengkategorikan sentimen menjadi 3 kategori, yaitu positif, negatif, dan netral.

Syntax:
Journal of Software Engineering, Computer Science and Information Technology
p-ISSN: 2776-7027, e-ISSN: 2723-0538
Volume: 5, Nomer: 1, Juni 2024

TABEL I
Labeling Sentimen Komentar Youtube

Deskripsi	Sentimen
sejuta terima kasih atas ketekunan dan kejujuran kalian. Perjuangan masih panjang. Kita bergerak bersama	Positif
dibayar ganjar piro leng	Negatif
apakabar	Netral

B. *Preprocessing Data*

Preprocessing data adalah tahap yang sangat penting dalam text mining, yang bertujuan untuk mengubah data teks mentah menjadi format yang lebih terstruktur dan siap untuk dianalisis.
Tahap ini mempersiapkan data seperti pembersihan data, pengubahan data, dan pengintegrasian data [7]. Proses ini melibatkan beberapa langkah penting yang membantu menghilangkan noise dan ketidakteraturan dalam data teks sehingga informasi yang relevan dapat diekstraksi dengan lebih mudah dan akurat. Berikut Langkah-langkah utama dalam preprocessing data :

1. *Text Cleaning*
   Tahap ini adalah proses membersihkan teks dari elemen-elemen yang tidak diinginkan atau tidak relevan. Proses ini biasanya melibatkan penghapusan simbol-simbol seperti tanda baca, angka, spasi berlebih, karakter khusus dan karakter non-alfabet [8].

2. *Stopword Removal*
   Stopword merupakan kata-kata yang tidak deskriptif yang dapat dibuang [9] seperti kata-kata umum yang sering muncul dalam teks tetapi tidak memberikan banyak nilai dalam analisis, seperti "dan", "atau", "adalah", dan "dengan". Menghapus stopwords membantu mengurangi dimensi data dan meningkatkan fokus pada kata-kata yang lebih signifikan dalam konteks analisis sentimen.

3. *Tokenized*
   Tokenized adalah proses pemilahan data berupa kalimat atau frasa menjadi beberapa kata [10]. Kata - kata tersebut adalah token. Setiap token biasanya berupa kata atau frasa yang akan dianalisis lebih lanjut.

4. *Stemming*
   Stemming adalah proses mengurangi kata-kata ke bentuk dasarnya atau akarnya. Tujuan dari proses stemming adalah menghilangkan imbuhan-imbuhan baik itu berupa prefiks, sufiks, maupun konfiks yang ada pada setiap kata [11]. Ini dilakukan untuk menghilangkan variasi kata yang berbeda yang memiliki makna dasar yang sama.

TABEL III
Proses Preprocessing Data

Sebelum Preprocessing	Setelah Preprocessing
FILM DIRTY VOTE DISEBARKAN PADA MASA TENANG JELANG HARI "H" PENCOBLOSAN.	film dirty vote sebar masa tenang jelang hari h coblos
Semua kru tlg siapkan mental anda utk menyaksikan pelantikan pak Prabowo	semua kru siap mental utk saksi lantik pak prabowo
Film yg bikin rusuh,	film buat rusuh

C. *Visualisasi Data*

Visualisasi data adalah sebuah visual yang menunjukan sentimen berdasarkan kelasnya. Pada proyek ini visualisasi berupa *WordCloud* pada Gambar 2. dan Gambar 3. yang menampilkan kata-kata sesuai dengan jumlah kemunculannya dalam sentimen positif ataupun negatif.

Gambar. 2 *WordCloud* Positif & Negatif

Wordcloud adalah gambaran visual berdasarkan frekuensi kemunculan kata-kata pada suatu kumpulan teks, dimana ukuran huruf menentukan frekuensi kemunculan sebuah kata yang artinya semakin besar ukuran huruf maka semakin besar kemunculan kata tersebut dan sebaliknya, semakin kecil huruf maka semakin kecil frekuensi kemunculan kata tersebut [12] yang berguna untuk menganalisis teks untuk membantu mengidentifikasi dan menyoroti kata-kata yang paling sering muncul dalam kumpulan data. Pada Visualisasi positif kata terbanyak di dapatkan oleh kata ”semua” sedangkan visuali negatif kata terbanyak di dapatkan oleh kata ”film”

Syntax:
Journal of Software Engineering, Computer Science and Information Technology
p-ISSN: 2776-7027, e-ISSN: 2723-0538
Volume: 5, Nomer: 1, Juni 2024

D. Split Data

Pada tahap pemisahan data, data akan dibagi menjadi 2, yaitu data training sebesar 85% atau sebanyak 1.364 data dan data testing sebanyak 15% atau sebanyak 241 data. Data training digunakan untuk melatih model dengan menggunakan Algoritma KNN dan Naïve Bayes, kemudian data testing digunakan untuk mengevaluasi kinerja model Algoritma KNN dan Naïve Bayes yang telah dibuat.

E. Penggabungan Model K-Nearest Neighbor & Naïve Bayes

K-Nearest Neighbor (KNN) adalah salah satu algoritma dalam supervised learning yang digunakan untuk tugas klasifikasi dan regresi. KNN bekerja berdasarkan prinsip kemiripan (similarity) antara data baru dengan data yang telah diklasifikasikan sebelumnya. Algoritma ini sederhana dan mudah dilakukan, tidak sedikit peneliti yang meragukan nilai dari k yang dihasilkan. Nilai k yang tinggi akan mengurangi efek noise, tetapi akan membuat hasil prediksi semakin kabur, sedangkan jika nilai k terlalu kecil atau 1, akan mengakibatkan hasil prediksi terasa kaku [13].

Sedangkan, Naïve Bayes adalah algoritma klasifikasi berbasis probabilitas yang didasarkan pada Teorema Bayes dengan asumsi independensi antara fitur-fitur. Meskipun asumsi independensi ini sering kali tidak realistis, Naïve Bayes tetap efektif dan efisien dalam banyak tugas klasifikasi teks. Algoritma ini memliki Teknik prediksi probabilitas berdasarkan pada penerapan teorema bayes dimana antara suatu fitur dengan fitur lain dalam suatu data itu tidak saling keterkaitan, Teknik [14].

Di Tahap ini Algoritma KNN dan Naive Bayes akan membentuk sebuah model dengan cara melatih data latih yang telah di hasilkan dari tahap Split Data yang akan menghasilkan sebuah model.

F. Evaluasi Model

Evaluasi model adalah tahap penting dalam pengembangan algoritma K-Nearest Neighbor (KNN) dan Naïve Bayes. Evaluasi bertujuan untuk menilai kinerja model dalam melakukan tugas klasifikasi, memastikan model bekerja dengan baik pada data baru yang belum pernah dilihat sebelumnya.

Perbandingan kinerja antara KNN dan Naïve Bayes dilakukan menggunakan metrik-metrik evaluasi yang telah dijelaskan. Akurasi, precision, recall,dan F1-score performa masing-masing model. Cross-validation memastikan bahwa model dievaluasi secara robust, menghindari overfitting dan memastikan kemampuan generalisasi model. Berikut merupakan formula atau rumus persamaan untuk mendapatkan nilai dari presisi, recall, dan akurasi [15] :

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

$$\text{Recall} = \frac{\text{TP}}{TP + FN}$$

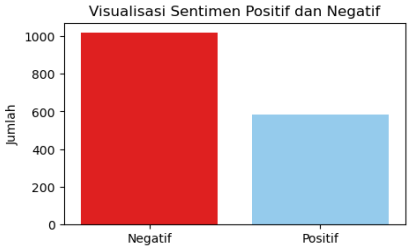
$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

Keterangan :
TP = True Positive
TN = True Negative
FP = False Positive
FN = False Negative

Dengan menggabungkan hasil dari kedua model ini, kita dapat mendapatkan hasil kklasifikasi yang baik berdasarkan karakteristik data dan tujuan analisis. Evaluasi yang tepat akan membantu dalam memilih model yang memberikan keseimbangan terbaik antara kompleksitas dan kinerja prediktif.

III. HASIL DAN ANALISIS

Pada proyek ini data yang digunakan berjumlah 1.605 setelah melalui proses preprocessing dengan jumlah sentimen positif sebanyak 584 dan sentimen negatif sebanyak 1021 yang ditunjukkan pada Gambar 3.



Gambar. 3 Jumlah Sentimen Positif dan Negatif

Penelitian dilakukan dengan menguji model KNN dan Naïve Bayes, Hasil perbandingan kedua algoritma tersebut dapat di lihat pada Tabel 3.

TABEL IIIII
HASIL DARI ALGORITMA KNN dan NAIVE BAYES

	K-Nearest Neighbor (KNN)		*Naïve Bayes*	
	Positif	Negatif	Positif	Negatif
Accuracy	0.66		0.76	
Precision	0.54	0.73	0.73	0.77
Recall	0.52	0.75	0.55	0.88

F1 Score	0.53	0.74	0.62	0.82

Berdasarkan dari Tabel 3. terlihat perbedaan nilai accuracy, precision, dan recall pada 2 Algoritma. Pada KNN didapatkan akurasi sebesar 0.66 sedangkan Naive Bayes didapatkan akurasi sebesar 0.76. Perbedaan akurasi antara kedua algoritma terlihat signifikan.

IV. KESIMPULAN

Berdasarkan penelitian yang telah dilakukan, didapatkan kesimpulan bahwa Algoritma Naive Bayes terbukti cukup baik dibandingkan dari Algoritma KNN dalam melakukan klasifikasi sentimen Film Dokumenter Dirty Vote melalui platform Youtube. Algoritma KNN didapatkan akurasi sebesar 0.66 sedangkan Naive Bayes didapatkan akurasi sebesar 0.76. Terlihat juga nilai Precision tertinggi di dapatkan oleh Algoritma Naïve Bayes sebesar 0.73 untuk positif dan 0.77 untuk negative, lalu nilai Recall tertinggi di dapatkan oleh Algoritma Naïve Bayes sebesar 0.55 untuk positif dan 0.88 untuk negative, Dan Nilai F1 Score terbesar juga di dapatkan oleh Algoritma Naïve Bayes sebesar 0.62 untuk positif dan 0.82 untuk negative.

REFERENSI

[1] S. Ardhi, “Ahli Hukum UGM Zainal Arifin Tanggapi Kontroversi Film ‘Dirty Vote,’” ugm.ac.id. Accessed: May 30, 2024. [Online]. Available: https://ugm.ac.id/id/berita/ahli-hukum-ugm-zainal-arifin-tanggapi-kontroversi-film-dirty-vote/

[2] H. ,Muchammad Nurrun, I. ,Paulus Santosa, and W. W. Winarno, “STUDI LITERATUR TENTANG PERBANDINGAN METODE UNTUK PROSES ANALISIS SENTIMEN DI TWITTER,” 2016.

[3] R. Alfiyyah, R. Andreswari, and E. Sutoyo, “ANALISIS DAN DETEKSI FRAUD PADA DATA PANGGILAN MENGGUNAKAN ALGORITMA K NEAREST NEIGHBOR (STUDI KASUS: PT XYZ),” 2020.

[4] P. Sofyan Zakaria, R. Julianto, and R. Surya Bernada, “IMPLEMENTASI NAIVE BAYES MENGGUNAKAN PYTHON DALAM KLASIFIKASI DATA,” 2023. [Online]. Available: https://jurnalmahasiswa.com/index.php/biikma

[5] A. Deolika and E. Taufiq Luthfi, “ANALISIS PEMBOBOTAN KATA PADA KLASIFIKASI TEXT MINING,” Jurnal Teknologi Informasi, vol. 3, no. 2, 2019

[6] A. H. Tri Jaka, “Preprocessing Text untuk Meminimalisir Kata yang Tidak Berarti dalam Proses Text Mining,” 2015.

[7] F. Putra, H. F. Tahiyat, R. M. Ihsan, R. Rahmaddeni, and L. Efrizoni, “Penerapan Algoritma K-Nearest Neighbor Menggunakan Wrapper Sebagai Preprocessing untuk Penentuan Keterangan Berat Badan Manusia,” MALCOM: Indonesian Journal of Machine Learning and Computer Science, vol. 4, no. 1, pp. 273–281, Jan. 2024, doi: 10.57152/malcom.v4i1.1085.

[8] S. Syafrizal, M. Afdal, and R. Novita, “Analisis Sentimen Ulasan Aplikasi PLN Mobile Menggunakan Algoritma Naïve Bayes Classifier dan K-Nearest Neighbor,” MALCOM: Indonesian Journal of Machine Learning and Computer Science, vol. 4, no. 1, pp. 10–19, Dec. 2023, doi: 10.57152/malcom.v4i1.983.

[9] L. Aji Andika and P. Amalia Nur Azizah, “Analisis Sentimen Masyarakat terhadap Hasil Quick Count Pemilihan Presiden Indonesia 2019 pada Media Sosial Twitter Menggunakan Metode Naive Bayes Classifier,” 2019.

[10] Y. P. Akbar, M. Sri Satyawati, and N. Putra Sastra, “Analisis Sentimen Kata Anjay pada Media Sosial Twitter Dalam Kajian Linguistik Komputasi,” Stilistika : Journal of Indonesian Language and Literature, vol. 1, no. 2, p. 62, Apr. 2022, doi: 10.24843/stil.2022.v01.i02.p06.

[11] A. Z. Amrullah, A. Sofyan Anas, M. Adrian, and J. Hidayat, “Analisis Sentimen Movie Review Menggunakan Naive Bayes Classifier Dengan Seleksi Fitur Chi Square,” Jurnal, vol. 2, no. 1, 2020, doi: 10.30812/bite.v2i1.804.

[12] J. Indri and Lindawati, “Analisis Sentimen Terhadap Sistem Informasi Akademik Mahasiswa Institut Teknologi Garut,” 2022. [Online]. Available: https://jurnal.itg.ac.id/

[13] P. Yulianto and S. Darwis, “Penerapan Metode K-Nearest Neighbors (kNN) pada Bearing,” 2021, doi: 10.29313/.v1i01.7090.

[14] M. E. Lasulika, "KOMPARASI NAÏVE BAYES, SUPPORT VECTOR MACHINE DAN K-NEAREST NEIGHBOR UNTUK MENGETAHUI AKURASI TERTINGGI PADA PREDIKSI KELANCARAN PEMBAYARAN TV KABEL," ILKOM Jurnal Ilmiah, vol. 11, no. 1, pp. 11–16, May 2019, doi: 10.33096/ilkom.v11i1.408.11-16.

[15] M. M. Baharuddin, H. Azis, and T. Hasanuddin, "ANALISIS PERFORMA METODE K-NEAREST NEIGHBOR UNTUK IDENTIFIKASI JENIS KACA," ILKOM Jurnal Ilmiah, vol. 11, no. 3, pp. 269–274, Dec. 2019, doi: 10.33096/ilkom.v11i3.489.269-274.