HOAQ: JURNAL TEKNOLOGI INFORMASI, Volume 15 Nomor 1, Mei 2024, 22-32
p-ISSN: 2337-5280, e-ISSN: 26207427
DOI: https://doi.org/10.52972/hoaq.vol15no1.p22-32

ANALISIS SENTIMEN MENGGUNAKAN METODE NAIVE BAYES DAN METODE SUPPORT VECTOR MACHINE PADA KASUS PELANTIKAN ARTIS SEBAGAI ANGGOTA ANGGOTA DPR RI TAHUN 2024

Sebastianus Adi Santoso Mola[1], Patrisius Remby Lete[2], Triyanto[3], Bernard Jose Adrian Junio Ajilo Pa[4], Tiwuk Widiastuti[5]

[1]Program Studi Ilmu Komputer, Universitas Nusa Cendana
Jl. Adisucipto Penfui, Kupang – Nusa Tenggara Timur, Indonesia
Email : [1]adimola@staf.undana.ac.id, [2]dokterpatrisius@gmail.com, [3]triyanto16082003@gmail.com, [4]bernardpa2104@gmail.com, [5]tritiwuk@gmail.com

ABSTRAK

Penelitian ini bertujuan untuk menganalisis sentimen publik terkait pelantikan artis sebagai anggota DPR tahun 2024 melalui komentar pada berbagai video *YouTube* terkait. Melalui analisis ini, hasil penelitian diharapkan memberikan wawasan bagi akademisi, pemerintah, dan *platform* media sosial tentang persepsi publik terhadap keterlibatan figur publik dalam politik, mendukung penelitian lebih lanjut, serta memberikan pertimbangan bagi kebijakan yang relevan dalam memahami pola sentimen masyarakat terhadap isu politik tertentu. Metode yang digunakan adalah metode *Naïve Bayes* dan *Support Vector Machine (SVM),* yang terbukti memiliki akurasi tinggi dan hasil yang bervariasi ketika dibandingkan kedua metode dalam mengklasifikasikan komentar menjadi tiga kelas sentimen: positif, netral, dan negatif pada berbagai kasus. Data set awal terdiri dari 6.438 komentar yang telah melalui proses pembersihan, pengolahan,penerjemahan dan pelabelan, yang kemudian digunakan *TF-IDF* untuk pembobotan kata serta metode *SMOTE* diterapkan untuk menyeimbangkan data sebelum klasifikasi dengan pembagian 80% data latih dan 20% data uji. Hasil klasifikasi menunjukkan bahwa *SVM* mencapai akurasi lebih unggul sebesar 91% dibandingkan Naïve Bayes sebesar 80%, terutama dalam mendeteksi sentimen negatif dan netral, sedangkan Naïve Bayes lebih efektif pada sentimen ekstrem (positif dan negatif) tetapi kurang optimal dalam mengidentifikasi sentimen netral.Sehingga perlu perbaikan pada kelas positif dan netral untuk klasifikasi yang jauh lebih optimal. Persepsi masyarakat berdasarkan hasil kedua metode menunjukan kelas negatif cenderung lebih rendah yang memberikan gambaran bahwa mayoritas masyarakat mungkin lebih menerima atau tidak terlalu terpengaruh oleh isu tersebut.

Kata kunci: Analisis Sentimen, Naïve Bayes, *Support Vector Machine*, DPR, Politik

ABSTRACT

This study aims to examine public sentiment regarding the appointment of celebrities as members of the Indonesian parliament in 2024, analyzing comments on various related YouTube videos. The findings are expected to provide valuable insights for academics, the government, and social media platforms on public views of public figures' involvement in politics, encourage further research, and inform relevant policy considerations on public sentiment toward political issues. The study uses Naïve Bayes and Support Vector Machine (SVM) methods, both known for their high accuracy and varied results when applied to classify comments into three sentiment categories: positive, neutral, and negative in different contexts. The initial dataset consists of 6,438 comments, which were cleaned, processed, translated, and labeled, followed by the application of TF-IDF for word weighting and the SMOTE technique to balance the data before classification with an 80% training and 20% testing split. The results show that SVM outperforms Naïve Bayes with a 91% accuracy rate compared to 80%, particularly excelling in detecting negative and neutral sentiments. In contrast, Naïve Bayes is more effective for extreme sentiments (positive and negative) but less effective in identifying neutral sentiments. Therefore, further improvements are needed for the positive and neutral sentiment classes to achieve optimal classification. Public perception, based on the results of both methods, indicates that negative sentiment is generally lower, suggesting that most of the public may be more accepting or less impacted by the issue.

Keywords: Sentiment Analysis, Naïve Bayes, Support Vector Machine, House of Representatives, Politics

1. PENDAHULUAN

Latar Belakang

Dengan kemajuan teknologi dan internet, media sosial telah menjadi platform utama untuk masyarakat luas menyuarakan pendapat dan saling berdiskusi mengenai berbagai topik, termasuk tentang politik, seperti peningkatan jumlah artis yang dilantik menjadi anggota DPR, mencapai 24 orang pada tahun 2024 [1]. Respon masyarakat terhadap keterlibatan artis di dunia politik bervariasi, baik dukungan maupun kritik. *YouTube*, sebagai media berbasis video dengan 73,4 miliar kunjungan (peringkat kedua setelah *Google*),menjadi media untuk mengekspresikan pandangan melalui komentar pada video terkait fenomena ini. Analisis sentimen terhadap komentar-komentar tersebut penting untuk memahami persepsi masyarakat secara umum [2].

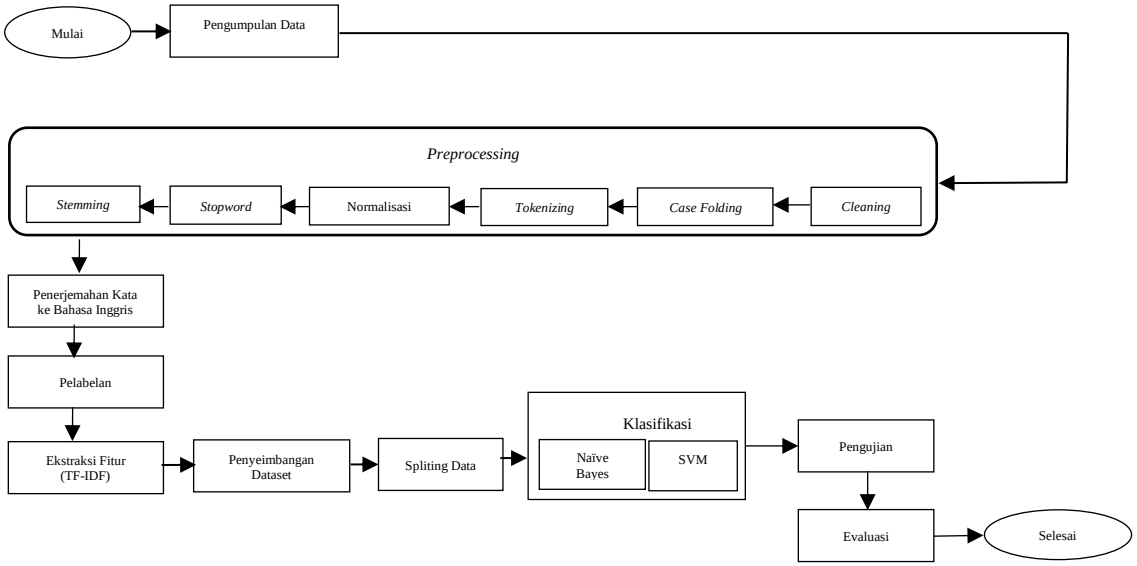
Dengan adanya pendekatan pembelajaran mesin yang memanfaatkan kata-kata emosional sebagai fitur untuk klasifikasi teks memungkinkan pemilihan emosi dengan cepat. Metode yang umum digunakan dalam analisis sentimen salah satunya metode *Naïve Bayes* dan *Support Vector Machine* (SVM) [3]. Analisis sentimen pada berbagai topik menggunakan metode *SVM* dan *Naïve Bayes* menunjukkan hasil yang bervariasi. Pada analisis tweet terhadap sentimen *ChatGPT*, *SVM* dengan Vader mengungguli *Naïve Bayes* dengan akurasi, presisi, dan *recall* 59% dibandingkan 47% [4]. Di sentimen aplikasi Pluang, *SVM* juga lebih unggul dengan akurasi 99,50% dibandingkan Naïve Bayes yang mencapai 99,25% [5]. Sedangkan untuk sentimen vaksin COVID-19, *Naïve Bayes* lebih baik dengan rata-rata performa 57,10%, sementara *SVM* mencatatkan 53,32% [6]. Pada sentimen perekrutan PPPK di Twitter, Naïve Bayes juga lebih unggul dengan akurasi 96,14%, dibandingkan *SVM* dengan 94,80% [7]. Sementara itu, untuk kenaikan harga bahan pokok, *Naïve Bayes* memiliki akurasi tinggi 94,38% [8]. Hasil dari berbagai penelitian ini menunjukkan bahwa kedua metode memiliki performa yang baik, meskipun dengan hasil yang bervariasi saat dibandingkan.

Pada sebuah penelitian ketidakseimbangan data set dapat saja terjadi. Ketidakseimbangan dataset dapat mempengaruhi performa model analisis sentimen, sehingga perlu dilakukan penyeimbangan sebelum klasifikasi. Misalnya, penelitian tentang aplikasi Ruangguru melakukan penyeimbangan dataset agar hasil lebih akurat [9]. Ketidakseimbangan data menyebabkan model lebih fokus pada kelas mayoritas dan menghasilkan performa klasifikasi yang tidak optimal. Untuk mengatasi hal ini, metode *SMOTE* (*Synthetic Minority Oversampling Technique*) digunakan untuk mengatasi ketidakseimbangan distribusi kelas melalui penambahan sampel sintetik pada kelas minoritas [10].

Dengan tujuan penelitiaan ini untuk menganalisis persepsi masyarakat melalui sentimen komentar pada beberapa video YouTube terkait pelantikan artis sebagai anggota DPR, dengan menggunakan metode *SVM* dan *Naïve Bayes*. Kedua metode ini dipilih karena, berdasarkan penelitian sebelumnya, keduanya terbukti memiliki performa yang cukup baik, meskipun dengan hasil yang bervariasi saat dibandingkan. Oleh karena itu, kedua metode ini menjadi menarik untuk diterapkan pada kasus ini. Urgensi perbandingan ini terletak pada pentingnya memilih metode yang paling optimal dalam mengklasifikasikan sentimen publik secara akurat, terutama untuk isu yang bersifat sensitif seperti keterlibatan *figur* publik dalam politik.Selain itu, karena ini merupakan penelitian pertama kami, kami memutuskan untuk fokus pada perbandingan dua metode ini sebagai langkah awal. Metode *Naïve Bayes* dipilih karena proses pelatihannya yang cepat, cocok untuk dataset kami yang besar, sementara *SVM* lebih sesuai dengan struktur teks komentar pada dataset kami yang kompleks. Evaluasi model dilakukan menggunakan *confusion matrix* pada tahap akhir. Akurasi dan keandalan hasil klasifikasi sangat diperlukan untuk mendukung tujuan penelitiaan ini. Diharapkan hasil penelitian ini dapat memberikan wawasan mengenai persepsi masyarakat, membantu pihak terkait memahami pandangan publik dengan lebih mendalam, serta mendukung penelitian lebih lanjut mengenai hasil kedua metode dalam kasus ini.

2. METODE PENELITIAN

Pada penelitian ini menggunakan alur proses pengerjaan analisis sentimen sesuai pada Gambar 1.



Gambar 1. Tahapan Penelitian

Sesuai pada Gambar 1 penelitian ini diawali dengan mengumpulkan data dari komentar pada beberapa video *YouTube* terkait, menghasilkan total 6.438 komentar yang kemudian disimpan untuk analisis lebih lanjut. Tahap selanjutnya adalah data *cleaning*, yang mencakup pembersihan karakter khusus, *URL*, elemen tidak relevan, penghapusan data duplikat, penghapusan data kosong (*NaN*), serta konversi ke huruf kecil (*case folding*). Kemudian dilakukan preprocessing dengan langkah-langkah normalisasi kata, tokenisasi, dan *stemming* menggunakan sastrawi untuk mempersiapkan data.

Selanjutnya, data diubah menjadi bahasa Inggris menggunakan *googletran*s agar dapat dilabeli menggunakan nltk dan *TextBlob*. Untuk memvisualisasikan kata-kata dalam data, word cloud dibuat berdasarkan keseluruhan data serta per label sentimen (positif, netral, negatif). Setelah itu, dipastikan tidak ada data kosong pada kolom '*tweet_eng'* dan 'klasifikasi', yang kemudian dipisahkan sebagai teks dan labelnya masing-masing, lalu dilakukan vektorisasi teks menggunakan TF-IDF.

Karena distribusi dataset pada setiap sentimen tidak seimbang, metode *SMOTE* diterapkan untuk menyeimbangkan distribusi data, khususnya dengan menambah data positif, sehingga distribusinya lebih merata. Dataset ini kemudian dilakukan pemisahan menjadi data latih dan uji dengan komposisi 80%:20%. Pada tahap terakhir, dilakukan klasifikasi dengan metode *Naïve Bayes* dan *SVM*, dilanjutkan dengan penyajian *confusion matrix* dan evaluasi model untuk kedua metode tersebut.

Pengumpulan data

Sumber data utama pada penelitiaan ini dikumpulkan dari komentar-komentar di *platform* media sosial YouTube pada beberapa video yang membahas pelantikan artis sebagai anggota DPR RI tahun 2024. Data utama ini mencakup opini, sentimen, dan reaksi masyarakat mengenai keterlibatan artis dalam dunia politik. Sentimen yang akan dianalisis mencakup kategori positif, negatif, dan netral. Penggunaan data untuk penelitian ini diambil dari komentar-komentar pada video yang ada di berbagai channel *YouTube*. Metode *scraping* digunakan untuk mengambil data tersebut, dengan bantuan *Google Colaboratory* sebagai alat untuk melakukan proses *web scraping*. Pengambilan data dilakukan dengan menyalin tautan video, lalu data akan diunduh dan disalin ke *Ms. Excel* dengan format *CSV*. Dari data yang diambil, hanya digunakan satu kolom teks komentar [11].

Text Preprocessing

Setelah pelabelan dilakukan pada seluruh data komentar yang dikumpulkan terkait pelantikan artis sebagai anggota DPR tahun 2024, langkah selanjutnya adalah melakukan *text preprocessing* untuk menyusun data menjadi lebih terstruktur. Tahapan *text preprocessing* mencakup beberapa langkah berikut:

1. ***Cleaning***: Proses ini dilakukan untuk membersihkan data komentar dari karakter, simbol, atau tanda baca yang tidak relevan pada data komentar.Sebagai contoh, pada kasus pelantikan artis sebagai anggota DPR tahun 2024, proses ini dilakukan terhadap 6.438 data komentar.
2. ***Case Folding***: Mengubah semua huruf dalam data komentar menjadi huruf kecil untuk menjaga konsistensi.
3. **Normalisasi Sinonim**: Mengganti kata dengan sinonim yang lebih umum atau standar untuk mengurangi variasi. Misalnya, "mobil" dan "kendaraan" dapat dinormalisasi menjadi "kendaraan".
4. ***Tokenizing***: Memecah kalimat dalam komentar menjadi potongan kata setelah proses *case folding*, sehingga setiap kata dapat dianalisis sebagai satuan data yang terpisah.
5. ***Stopword Removal***: Menghapus kata-kata biasa yang tidak memiliki kontribusi signifikan, seperti "yang," "di," dan "ke," agar analisis sentimen lebih terfokus.

6. ***Stemming***: Mengembalikan kata-kata berimbuhan menjadi bentuk dasar katanya sesuai dengan KBBI untuk memastikan keseragaman kata.

Tahapan *text preprocessing* ini membantu menghasilkan data yang lebih bersih dan siap dianalisis, sehingga mendukung analisis sentimen terhadap komentar terkait fenomena pelantikan artis sebagai anggota DPR tahun 2024 dengan lebih akurat [12].

Penerjemahan Kata ke Bahasa Inggris

Pada tahapan ini setiap data bersih diterjemahkan ke dalam bahasa inggris dengan menggunakan *library* dari phyton yaitu *googletrans* untuk dilakukan tahapan pelabelan dengan pustaka *ntlk.*

Pelabelan

Tahap pelabelan adalah proses di mana hasil dari tahap sebelumnya dianalisis untuk menghitung polaritas dari setiap ulasan yang diambil.Pada tahap ini, *TextBlob* digunakan untuk memproses data dalam dataset, sementara pustaka *nltk* digunakan untuk menganalisis polaritas, sehingga ulasan dapat dikategorikan menjadi tiga kelompok utama: positif, netral, dan negatif.

Pembobotan Kata (TF-IDF)

Pada tahapan pemmberian bobot kata *TF-IDF (Term Frequency-Inverse Document Frequency)*, data tekstual dikonversi ke data numerik berdasarkan frekuensi kata dalam dokumen. *TF* mengukur seberapa sering kata muncul dalam setiap dokumen, sementara *IDF* menghitung frekuensi dokumen yang mengandung kata tersebut. *IDF* menggunakan logaritma untuk mengurangi pengaruh kata yang sering muncul di banyak dokumen, sesuai dengan rumus dalam Persamaan 1 [10] .

$$TF - IDF_{t,d} = TF_{t,d} \times IDF_t \qquad (1)$$

Dimana, $IDF_t = log \frac{N}{DF_t}$

t = kata-kata yang diperhitungkan;
d = bobot kalimat (d);
$TF - IDF_{t,d}$ = Bobot (d) yang diterapkan pada kata (t);
$TF_{t,d} = Term\ Frequency$;
$IDF_t = Inverse\ Document\ Frequency$;
N = total kalimat;
$DF_t =$ jumlah kemunculan suatu kata;

Dalam penelitiaan ini menggunakan pustaka *scikit-learn* untuk menerapkan *TF-IDF* pada data set.

Penyeimbang Data

Pada langkah ini, masalah ketidakseimbangan dataset diatasi dengan mengaplikasikan metode oversampling *SMOTE* (*Synthetic Minority Over-sampling Technique*) yang terdapat dalam pustaka imblearn.

Splitting Data

Setelah diproses, data akan dibagi menjadi dua bagian, dengan 80% digunakan untuk melatih model dan 20% sisanya untuk pengujian. Pembagian ini bertujuan untuk memberikan kesempatan bagi model untuk belajar dari data yang berbeda dengan yang digunakan untuk pengujian.Komposisi tersebut dipilih karena *data set* yang cukup besar lebih dari 5000 data sehingga memerlukan cukup data untuk proses pelatihan dan jumlah data yang sama baiknya untuk melakukan proses pengujian.

Klasifikasi

Pada tahap klasifikasi data di modelkan dengan metode naïve baiyes dan metode *SVM* untuk dilakukan pemrosesan dengan data uji dan tes yang telah di bagi dari *data set*.

Metode *Support Vector Machine*

Teknik *Support Vector Machine (SVM*) digunakan dalam prediksi, baik untuk proses klasifikasi dan regresi. Secara dasar, SVM adalah linear *classifier* yang pertama kali dirancang untuk kasus klasifikasi *linier*. Meskipun SVM awalnya dirancang untuk masalah linier, teknik ini dapat dikembangkan untuk menangani masalah *non-linier* melalui pendekatan kernel di ruang kerja dengan banyak dimensi. Metode ini berfokus pada penentuan batas pemisah antara dua kelas dengan memaksimalkan jarak antar data terdekat. Untuk menemukan batas optimal antar kelas, dibangun *hyperplane* atau garis pembagi yang optimal di ruang masukan, yang ditentukan melalui pengukuran *margin hyperplane* dan pencarian nilai tertinggi [13]. Dimana tahap klasifikasi dilakukan melalui pencarian *hyperplane* atau *decision boundary* yang memisahkan kelas, di mana *SVM* berfokus pada pencarian ukuran *hyperplane* terbaik dengan memanfaatkan *support vector* serta *margin* yang tepat [9].

Metode *Naïve Bayes*

Naïve Bayes adalah teknik pelatihan berbasis probabilitas dalam pembelajaran mesin yang dipakai untuk mengklasifikasikan data uji secara individual [14]. Algoritma ini menggunakan Teorema Bayes untuk memperkirakan kemungkinan suatu peristiwa terjadi berdasarkan data yang telah ada sebelumnya [15]. Kelebihan utama dari *Naïve Bayes* adalah kesederhanaannya dalam implementasi dan kemampuannya memberikan akurasi tinggi dalam berbagai jenis data [16]. Algoritma ini mengidentifikasi apakah suatu data termasuk dalam kategori positif atau negatif. Berikut adalah rumus yang dipakai untuk melakukan perhitungan *Naïve Bayes* :

$$P(c_k|d_j) = P(c_k) \prod_{i=1}^{T} p\,(t_i|c_k) \qquad (2)$$

Dalam persamaan 2 $P(t_i|c_k)$ merujuk pada probabilitas yang bergantung pada kondisi kata t_i yang muncul dalam dokumen kelas c_k, sementara $P(c_k)$ menunjukkan probabilitas awal dokumen yang termasuk dalam kelas c_k.Kedua nilai tersebut diperoleh dari data pelatihan [10].

Pengujian

Proses evaluasi dilakukan dengan menerapkan metode *confusion matrix,* yang digunakan untuk menilai efektivitas atau tingkat akurasi model klasifikasi [10]. *Confusion Matrix* berfungsi sebagai alat untuk mengevaluasi efektivitas model klasifikasi dalam kecerdasan buatan untuk dua kelas maupun lebih. Ini disajikan dalam bentuk tabel yang mencakup empat kombinasi berbeda antara prediksi dan nilai sebenarnya [17].

Evaluasi

Setelah pelatihan pada train set dengan metode naïve baiyes dan *SVM* , model akan diuji menggunakan test set. Pengukuran dalam proses evaluasi model ini melibatkan metrik seperti akurasi, presisi, recall, dan F1-Score. Selain itu, dilakukan analisis kesalahan untuk memahami kasus-kasus di mana model gagal dalam melakukan prediksi yang akurat pada kedua metode tersebut.

3. HASIL DAN PEMBAHASAN

Pengumpulan Data

Data diambil melalui *scraping* pada komentar beberapa video *YouTube* terkait pelantikan artis sebagai anggota DPR RI tahun 2024, menghasilkan 6.438 komentar yang mencakup berbagai opini masyarakat. Dengan bantuan *Google Colaboratory* data di *scraping* menggunakan API untuk mengambil data komentar yang dimasukan kedalam satu kolom comment yang berisi komentar youtube. Setelah data berhasil diambil seluruhnya selanjutnya disimpan dengan format *CSV*. Karena keterbatasan jumlah komentar proses ini diulangi pada beberapa video digabung menjadi satu file *CSV*. Contoh data set pengumpulan data disajikan pada Tabel 1.

Tabel 1. Contoh Hasil Pengumpulan Data

No	Comment
1	dewan rakyat jaman kolonial belanda dan bpupki bentukan pemeritahan militer jepang lebih bermutu daripada ini dr kph sotardjo kartohadikusumo prof dr mr moh yamin sh kh agus salim raden mas h o s cokroaminoto dr k r t rajiman wedyodiningrat prof dr mr soepomo ir soekarno drs moh hatta dr mr a a maramis prof k h abdul kahar moezakkir dan masih banyak lagi sekarang isinya malah pelawak foto model youtuber tiktoker yg dulu bisa memerdekakan bangsa indonesia yg sekarang cuma manggutmanggut membiarkan rakyat dijajah preman berdasi preman berseragam oreman bersorban
2	indonesia semakin ruwet anjay orang gak berguna banyak yg jadi dpr anjay ok gas ok gas makan siang gratis
3	republik selebsalah siapa

Preprocessing

Setelah data dikumpulkan, tahap selanjutnya adalah *preprocessing* data. Proses ini meliputi enam tahapan: *stopword,removal, cleaning, case folding*, normalisasi kata, *tokenizing dan stemming* yang dapat dilihat contoh sebuah data dengan melalui 6 tahapan tersebut pada Tabel 2.

Tabel 2. Contoh Hasil Preprocessing Data

No	Preprocessing	Input	Output
1	Cleaning	seharus nya publik pigur tidak blh ikut jd kut jd dprkasih peluang anak bangsa iniyg lain diseroboti semua rezeki +0	Seharus nya publik pigur tidak blh ikut jd kut jd dprkasih peluang anak bangsa iniyg lain diseroboti semua rezeki
2	Case Folding	Seharus nya publik pigur tidak blh ikut jd kut jd dprkasih peluang anak bangsa iniyg lain diseroboti semua rezeki	seharus nya publik pigur tidak blh ikut jd kut jd dprkasih peluang anak bangsa iniyg lain diseroboti semua rezeki

No	Preprocessing	Input	Output
3	Tokenizing	seharus nya publik pigur tidak blh ikut jd kut jd dprkasih peluang anak bangsa iniyg lain diseroboti semua rezeki	[seharus, nya, public, pigur, tidak, blh, ikut, jd, kut, jd dprkasih, peluang, anak, bangsa iniyg, lain, diseroboti, semua, rezeki]
4	Normalisasi Kata	[seharus, nya, public, pigur, tidak, blh, ikut, jd, kut, jd dprkasih, peluang, anak, bangsa, iniyg, lain, diseroboti, semua, rezeki]	[seharusnya, publik, figur, tidak, boleh, ikut, jadi,dpr,kasih, peluang, anak, bangsa,ini,yang, lain, diseroboti, semua, rezeki]
5	Stopword Removal	[seharusnya, public, pigur, tidak, boleh, ikut, jadi,dpr,kasih, peluang, anak, bangsa,ini,yang, lain, diseroboti, semua, rezeki]	[seharusnya, public, pigur, ikut, jadi,dpr,kasih, peluang, anak, bangsa, lain, diseroboti, semua, rezeki]
6	Stemming	[seharusnya, public, pigur, ikut, jadi,dpr,kasih, peluang, anak, bangsa, lain, diseroboti, semua, rezeki]	[harus, publik, figur, ikut, jadi, dpr, kasih, peluang, anak, bangsa, lain, serobot, semua, rezeki]

Data Set Hasil Transleting dan Pelabelan

Pada proses ini data yang sudah melewati proses *preprocessing*, akan di lakukan proses *translate* menggunakan pustaka *google trans* sehingga akan menghasilkan data seperti yang ada pada Tabel 3 yang akan diklasifikasikan menjadi positif, negatif dan netral. Dengan hasil analisis data menggunakan *textlob* dengan Pustaka ntlk sebanyak 6291 dengan positif 2983, netral 2375 dan negatif 933.

Tabel 3.Contoh Hasil Pelabelan Data

NO	Comment	tweet_eng	klasifikasi
1	indonesia makin sulit sekali orang guna banyak jadi dpr sekali gas gas makan siang gratis.	*In Indonesia, it is becoming increasingly difficult for people to use a lot of gas and free lunches*	Positif
2	jadi nih carry orang.	*So, people carry it.*	Netral
3	sumpah bila langgar urus maha kuasa.	*I swear that if you break it, the Almighty will take care of it*	Netral
4	duit atur negara semua beli lama lama negara di jual sekarang masyarakat tengah tanggung kecil makin miskin timpang sosial jauh	*The money that regulates the state is all being bought for a long time, the state is being sold, now the community is having little responsibility and is getting poorer and socially unequal*	Negatif
5	paling arti rakyat kecil kurang makan.	*most importantly, the little people don't eat enough*	Positif
6	indonesia bubar rakyat miskin tengah jadi sapi perah gaji	*in Indonesia, the poor people have become salary cash cows*	Negatif

Pembobotan Kata (TF-IDF)

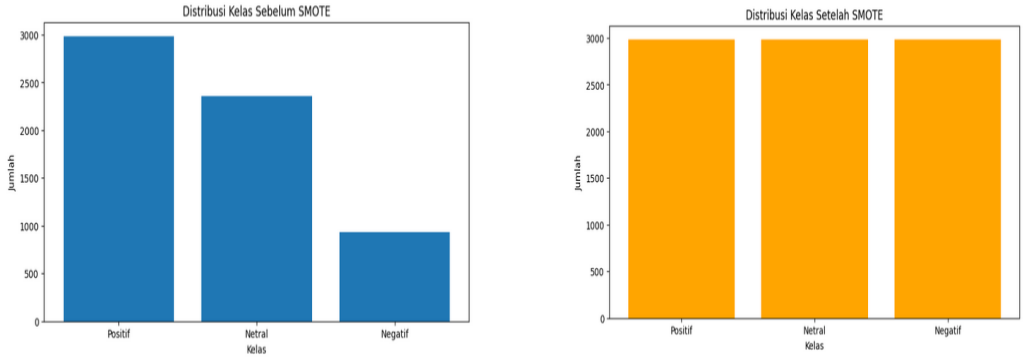
Algoritma *TF-IDF* menentukan nilai bobot untuk setiap kata dalam dataset. *TF* mengukur seberapa sering sebuah kata muncul dalam dokumen, sedangkan *IDF* merujuk pada jumlah dokumen (*DF)* yang yang memiliki kata tersebut. Implementasi *TfidfVectorizer* dan *Scikit-learn* dilakukan pada proses ini. Contoh kata dengan nilai *IDF*nya dapat dilihat pada Tabel 4.

Tabel 4. Contoh Penggalan Data Hasil Pembobotan Kata TF-IDF

No	Kata	IDF
1	blok	8.645716
2	larang	8.645716
3	selamat	9.051182

Penyeimbang Data

Setelah dilakukan ekstrasi *fitur* dengan *TF-IDF* karena data set yang tidak seimbang, kita perlu menyeimbangkan data set menggunakan *SMOTE* agar pelabelan antara sentimen positif, netral dan sentimen negatif bisa diseimbangkan.

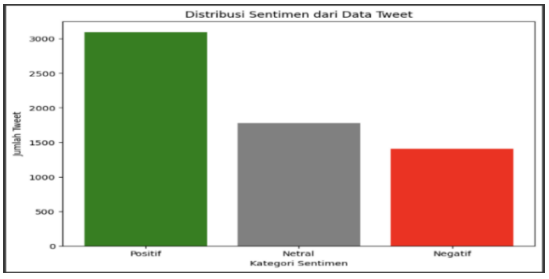


Gambar 2. Perbandingan grafik sentimen sebelum dan sesudah SMOTE

Seperti pada Gambar 2 sebelum menggunakan *SMOTE*, perbandingan data antara sentimen positif,netral dan negatif rasio datanya 3:2:1, setelah penerapan metode *SMOTE*, distribusi kelas menjadi 1:1:1 karena penambahan data sintetis pada kelas minoritas (Netral dan Negatif) hingga jumlahnya setara dengan kelas mayoritas (Positif). Metode ini bekerja dengan membandingkan jumlah data antar kelas, di mana kelas minoritas diseimbangkan terhadap kelas mayoritas menggunakan pendekatan 1 *against* label yang membandingkan kelas minoritas dengan kelas mayoritas. Hasilnya, semua kelas memiliki jumlah sampel yang sama, sehingga model dapat bekerja lebih adil dan tidak bias terhadap kelas tertentu.

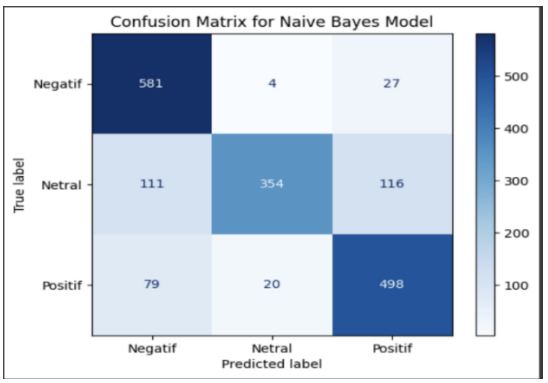
Klasifikasi Naïve Baiyes

Setelah penyeimbangan data set, data set dipisah menjadi data latih dan data uji dengan komposisi 80:20. Sehingga menghasilkan klasifikasi *naive baiyes* pada Gambar 3.



Gambar 3. Grafik Naive Baiyes

Pada Gambar 3 menunjukan hasil Analisis data klasifikasi *Naïve Baiyes* yaitu data positif 3096, data netral 1776, data negatif 1402 dari total data sebanyak 6274 yang memberikan gambaran bahwa mayoritas masyarakat mungkin lebih menerima atau tidak terlalu terpengaruh oleh isu ini. Selanjutnya dilakukan pengujian dengan *confusion matrix* dengan hasil pada Gambar 4.



Gambar 4. *Confusion Matrix Naive Baiyes*

Gambar 4 adalah *confusion matrix* untuk model *Naive Bayes* yang memiliki tiga kategori sentimen: negatif, netral, dan positif. Berikut adalah penjelasan dari hasil yang ditampilkan pada matriks tersebut:

1. **Sentimen Negatif:** Model *Naïve Bayes* dengan akurat memprediksi 581 sampel sebagai negatif. Akan tetapi, ditemukan 4 sampel masuk ke kelas netral dan 27 sampel yang masuk ke kelas positif. Sehingga model cukup optimal untuk kelas negatif.
2. **Sentimen Netral:** Model berhasill memprediksi 354 sampel sebagai netral. Akan tetapi, terdapat 111 sampel masuk ke kelas negatif, serta 116 sampel masuk ke kelas positif.Sehingga model cukup kesulitan menganali kelas netral ditandai dengan jumlah data meleset yang cukup besar.
3. **Sentimen Positif:** Model berhasil memprediksi 498 sampel sebagai positif. Akan tetapi, terdapat 79 sampel masuk ke kelas negatif dan 20 sampel masuk ke kelas positif netral.Sehingga model cukup optimal untuk kelas positif.

Setelah melihat hasil *confusion matrix* ini, langkah berikutnya adalah memanfaatkan metode *classification report* untuk mengevaluasi hasil model agar mendapatkan nilai presisi,recall dan skor F1 setiap kelas, yang membantu dalam memahami performa model secara lebih mendetail.

Classification Report:

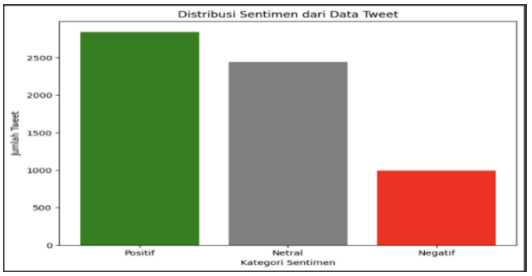
	precision	recall	f1-score	support
Negatif	0.75	0.95	0.84	612
Netral	0.94	0.61	0.74	581
Positif	0.78	0.83	0.80	597
accuracy			0.80	1790
macro avg	0.82	0.80	0.79	1790
weighted avg	0.82	0.80	0.80	1790

Gambar 5. Evaluasi *Naive Baiyes*

Algoritma *Naive Bayes* menunjukkan efektivitas sesuai pada Gambar 5 dengan akurasi keseluruhan sebesar 80%, di mana sentimen negatif dan positif terklasifikasi lebih baik dibanding sentimen netral. Hasil klasifikasi berhasil mencatat presisi 0.75, *recall* 0.95, dan skor F1 0.84 pada kategori negatif, serta presisi 0.78, *recall* 0.83, dan skor F1 0.80 pada kategori positif.. Namun, model mengalami kesulitan dalam mengenali sentimen netral, yang terlihat dari nilai *recall* yang lebih rendah (0.61), meskipun presisinya tinggi (0.94). Evaluasi ini menunjukkan bahwa model lebih efektif dalam mendeteksi sentimen ekstrem (negatif dan positif) daripada sentimen netral.

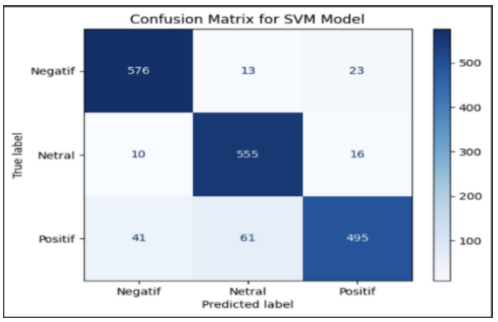
Klasifikasi Support Vector Machine

Setelah peneyeimbangan data set, sama seperti sebelumnya pada klasiifikasi naïve baiyes data set kembali dipisah menjadi 80% data digunakan untuk melatih model, sementara 20% sisanya digunakan untuk pengujian. Proses klasifikasi dengan *SVM* menghasilkan data pada grafik di bawah:



Gambar 6. Grafik *Support Vector Machine*

Pada Gambar 6 menunjukan hasil Analisis data klasifikasi *Support Vector Machine* yaitu data positif 2843, data netral 2438, data negatif 993 dari total data sebanyak 6274 yang memberikan gambaran bahwa mayoritas masyarakat mungkin lebih menerima atau tidak terlalu terpengaruh oleh isu ini. Selanjutnya dilakukan pengujian dengan confusion matrix dengan hasil bisa dilihat pada Gambar 6.



Gambar 7. *Confusion Matrix Support Vector Machine*

Gambar 7 adalah *confusion matrix* untuk model *Support Vector Machine* yang memiliki tiga kategori sentimen: negatif, netral, dan positif. Berikut adalah penjelasan dari hasil yang ditampilkan pada matriks tersebut:

1. **Sentimen Negatif:**
   Model *SVM* berhasil memprediksi 576 sampel sebagai negatif. Namun, terdapat 13 sampel yang masuk ke kelas netral dan 23 sampel masuk ke kelas positif. Sehingga model cukup optimal untuk kategori negatif.
2. **Sentimen Netral:**
   Model berhasil memprediksi 555 sampel sebagai netral dengan benar. Namun, 10 sampel yang sebenarnya netral diprediksi sebagai negatif dan 16 sampel diprediksi sebagai positif padahal netral. Hal ini menunjukkan kemampuan yang baik dalam mengidentifikasi sentimen netral, tetapi masih ada kesalahan prediksi ke kelas negatif dan positif.
3. **Sentimen Positif:**
   Model memprediksi 495 sampel sebagai positif. Akan tetapi, terdapat 41 sampel masuk ke kelas negatif dan 61 sampel masuk ke kelas netral. Sehingga performa model *SVM* baik untuk mengenali kelas positif.

Setelah melihat hasil *confusion matrix* ini, langkah berikutnya adalah memanfaatkan metode *classification report* untuk mengevaluasi hasil model agar mendapatkan nilai presisi,recall dan skor F1setiap kelas, yang membantu dalam memahami performa model secara lebih mendetail.

Classification Report:

	precision	recall	f1-score	support
Negatif	0.92	0.94	0.93	612
Netral	0.88	0.96	0.92	581
Positif	0.93	0.83	0.88	597
accuracy			0.91	1790
macro avg	0.91	0.91	0.91	1790
weighted avg	0.91	0.91	0.91	1790

Gambar 8. Evaluasi *Support Vector Machine*

Algoritma *Support Vector Machine (SVM)* ini menunjukkan kinerja yang sangat baik sesuai pada gambar 8 dengan akurasi keseluruhan sebesar 91%. Model berhasil mengklasifikasikan sentimen negatif dan netral lebih akurat dibandingkan sentimen positif. Model mencapai presisi sebesar 0.92, *recall* 0.94, dan skor F1 0.93 untuk sentimen negatif. Adapun sentimen netral memperoleh presisi 0.88, *recall* 0.96, dan skor F1 0.92. Namun, model sedikit mengalami kesulitan dalam mengenali sentimen positif, yang terlihat dari nilai *recall* yang lebih rendah (0.83), meskipun precision-nya cukup tinggi (0.93). Evaluasi ini menegaskan bahwa model SVM lebih unggul dalam klasifikasi sentimen negatif dan netral daripada sentimen positif. Meskipun performa keseluruhan cukup konsisten, ada ruang untuk peningkatan dalam mendeteksi sentimen positif agar model menjadi lebih seimbang.

Visualisasi Kata

Selanjutnya, tahapan pelengkap dilakukan dengan memvisualisasikan hasil pengklasifikasian sentimen komentar pada video di channel YouTube yang berkaitan dengan pelantikan artis sebagai anggota DPR RI tahun 2024.

Gambar 9. Visualisasi Kata

Proses ini diterapkan pada seluruh kata dan hasil pengklasifikasian untuk kategori positif, netral, dan negatif, yang dapat dilihat pada Gambar 9. Visualisasi menggunakan *wordcloud* menampilkan kata-kata yang paling sering

muncul dalam komentar pada berbagai video di *YouTube*. Dari visualisasi ini, dapat disimpulkan bahwa kata-kata yang muncul cenderung serupa, meskipun dengan frekuensi kemunculan yang bervariasi.

4. SIMPULAN

Pada penelitian ini menjelaskan tentang analisis sentimen pada komentar video *YouTube* yang terkait dengan pelantikan artis sebagai anggota DPR RI tahun 2024. Penggunaan duua model klasifikasi, yaitu Naive Bayes dan Support Vector Machine (SVM), digunakan untuk mengklasifikasikan sentimen komentar ke dalam tiga kategori: positif, netral, dan negatif. Setelah melalui proses preprocessing data seperti mencakup penghilangan stopword, pembersihan teks, *case folding*, tokenisasi, dan *stemming*, kedua model diuji dan dievaluasi dengan menggunakan confusion matrix dan classification report.

Berdasarkan hasil analisis, hasil klasifikasi *SVM* menunjukkan kinerja yang lebih tinggi secara keseluruhan dari hasil klasifikasi Naive Bayes, dengan tingkat akurasi sebesar 91% untuk SVM dan 80% untuk Naive Bayes. Model SVM lebih akurat dalam mengenali sentimen negatif dan netral dibandingkan dengan sentimen positif, yang ditunjukkan oleh nilai precision dan recall yang tinggi untuk kelas negatif dan netral. Sementara itu, model Naive Bayes lebih baik dalam mendeteksi sentimen negatif dan positif, meskipun performanya kurang optimal dalam membedakan sentimen netral. Selain itu penelitiaan ini menunjukan perlunya peningkatan pada kelas postif dan netral untuk hasil klasifikasi jauh lebih optimal. Persepsi masyarakat berdasarkan hasil kedua metode menunjukan kelas negatif cenderung lebih rendah yang memberikan gambaran bahwa mayoritas masyarakat mungkin lebih menerima atau tidak terlalu terpengaruh oleh isu tersebut.

Setelah melaui setiap proses yang ada, menunjukkan hasil klasifikasi *SVM* lebih tinggi dalam pengelompokan sentimen yang lebih akurat, terutama untuk sentimen yang bersifat ekstrem (negatif dan positif). Namun, baik Naive Bayes maupun SVM masih memerlukan peningkatan dalam mengidentifikasi sentimen positif untuk mencapai keseimbangan klasifikasi yang lebih optimal. Kesimpulan ini dapat menjadi panduan untuk pengembangan lebih lanjut dalam analisis sentimen, khususnya dalam meningkatkan akurasi deteksi sentimen netral dan positif pada model klasifikasi.

DAFTAR PUSTAKA

[1] N. W. Yunita, “24 Artis Dilantik Jadi Anggota DPR,” detik.com. Accessed: Sep. 08, 2024. [Online]. Available: https://www.detik.com/edu/edutainment/d-7566147/24-artis-dilantik-jadi-anggota-dpr.

[2] Tempo, “Semakin Banyak Orang Meengakses Berita darri Tiktok, Bagaimana Nasib Bismis Media Massa?,” tempo.com. Accessed: Sep. 08, 2024. [Online]. Available: https://bisnis.tempo.co/read/1925173/semakin-banyak-orang-mengakses-berita-dari-tiktok-bagaimana-nasib-bisnis-media-massa?tracking_page_direct.

[3] A. R. Isnain, A. I. Sakti, D. Alita, and N. S. Marga, “Sentimen Analisis Publik Terhadap Kebijakan Lockdown Pemerintah Jakarta Menggunakan Algoritma Svm,” *J. Data Min. dan Sist. Inf.*, vol. 2, no. 1, p. 31, 2021, doi: 10.33365/jdmsi.v2i1.1021.

[4] D. Atmajaya, A. Febrianti, and H. Darwis, “Metode SVM dan Naive Bayes untuk Analisis Sentimen ChatGPT di Twitter,” *Indones. J. Comput. Sci.*, vol. 12, no. 4, pp. 2173–2181, 2023, doi: 10.33022/ijcs.v12i4.3341.

[5] B. A. Maulana, M. J. Fahmi, A. M. Imran, and N. Hidayati, “Analisis Sentimen Terhadap Aplikasi Pluang Menggunakan Algoritma Naive Bayes dan Support Vector Machine (SVM),” *MALCOM Indones. J. Mach. Learn. Comput. Sci.*, vol. 4, no. 2, pp. 375–384, 2024, doi: 10.57152/malcom.v4i2.1206.

[6] D. Alita and R. A. Shodiqin, “Sentimen Analisis Vaksin Covid-19 Menggunakan Naive Bayes Dan Support Vector Machine,” *J. Artif. Intell. Technol. Inf.*, vol. 1, no. 1, pp. 1–12, 2023, doi: 10.58602/jaiti.v1i1.20.

[7] F. N. Hidayat and S. Sugiyono, “Analisis Sentimen Masyarakat Terhadap Perekrutan Pppk Pada Twitter Dengan Metode Naive Bayes Dan Support Vector Machine,” *J. Sains dan Teknol.*, vol. 5, no. 2, pp. 665–672, 2023, doi: 10.55338/saintek.v5i2.1359.

[8] M. Muslimin, V. Lusiana, F. Teknologi, P. Studi, T. Informatika, and U. Stikubank, “Analisis Sentimen Terhadap Kenaikan Harga Bahan Pokok Menggunakan Metode Naive Bayes Classifier,” vol. 7, pp. 1200–1209, 2023, doi: 10.30865/mib.v7i3.6418.

[9] E. Fitri, “Sentiment Analysis of the Ruangguru Application Using Naive Bayes, Random Forest and Support Vector Machine Algorithms,” *J. Transform.*, vol. 18, no. 1, p. 71, 2020.

[10] M. Knn and D. Smote, “Analisis Sentimen Program Mbkm Pada Media Sosial Twitter,” vol. 6, no. 2, pp. 89–98, 2023.

[11] I. Afdhal, R. Kurniawan, I. Iskandar, R. Salambue, E. Budianita, and F. Syafria, “Penerapan Algoritma Random Forest Untuk Analisis Sentimen Komentar Di YouTube Tentang Islamofobia,” *J. Nas. Komputasi dan Teknol. Inf.*, vol. 5, no. 1, pp. 122–130, 2022, [Online]. Available: http://ojs.serambimekkah.ac.id/jnkti/article/view/4004/pdf

[12] H. Setiawan, E. Utami, and S. Sudarmawan, “Analisis Sentimen Twitter Kuliah Online Pasca Covid-19 Menggunakan Algoritma Support Vector Machine dan Naive Bayes,” *J. Komtika (Komputasi dan Inform.*, vol. 5, no. 1, pp. 43–51, 2021, doi: 10.31603/komtika.v5i1.5189.

[13] R. A. Rizal, I. S. Girsang, and S. A. Prasetiyo, “Klasifikasi Wajah Menggunakan Support Vector Machine (SVM),” vol. 3, no. 2, pp. 1330–1333, 2019.

[14] H. Tuhuteru *et al.*, “Analisis Sentimen Perusahaan Listrik Negara Cabang Ambon Menggunakan Metode Support Vector Machine dan Naive Bayes Classifier,” vol. 03, no. 03, pp. 394–401, 2018, doi: 10.30591/jpit.v3i3.977.
[15] D. D. Putri, G. F. Nama, and W. E. Sulistiono, “Analisis Sentimen Kinerja Dewan Perwakilan Rakyat (DPR) pada Twitter Menggunakan Metode Naive Bayes Classifier,” vol. 10, no. 1, pp. 34–40, 2022.
[16] T. Thiraviyam, *Artificial Intelligence Marketing*, vol. 19, no. 4. 2018.
[17] F. N. Rozi and D. H. Sulistyawati, “Klasifikasi Berita Hoax Pilpres Menggunakan Metode Modified K-Nearest Neighbor Dan Pembobotan Menggunakan Tf-Idf,” *Konvergensi*, vol. 15, no. 1, 2019, doi: 10.30996/konv.v15i1.2828.