3rd Seminar Nasional Mahasiswa Fakultas Teknologi Informasi (SENAFTI)
30 Agustus 2023 – Jakarta, Indonesia
Volume 2, Nomor 2, September 2023 - ISSN 2962-8628 (*online*)

IMPLEMENTASI NAÏVE BAYES CLASSIFIER UNTUK KLASIFIKASI SENTIMEN PENGGUNA TWITTER TERHADAP KINERJA DPR

Faraz Septarian Adi Nugroho[1]

[1]Program Studi Teknik Informatika, Fakultas Teknologi Informasi, Universitas Budi Luhur, Jakarta, Indonesia

Email: [1]faraz.spt08@gmail.com

Abstrak- Sebagai wakil rakyat, Dewan Perwakilan Rakyat Republik Indonesia (DPR RI) bertanggung jawab dalam mengatasi permasalahan sosial masyarakat secara demokratis dan responsif. Namun, dalam beberapa kasus, terlihat lebih mengutamakan kepentingan partai politik daripada kepentingan masyarakat, maka dari itu banyak masyarakat yang masih merasa tidak puas dengan kinerja DPR. Dalam penelitian ini, akan dilakukan klasifikasi sentimen pengguna *twitter* Indonesia untuk mengukur tingkat kepuasan masyarakat terhadap kinerja DPR saat ini dengan menggunakan metode *Naïve Bayes Classifier*. Dalam proses klasifikasi sentimen terdapat beberapa tahap yaitu pengumpulan data, *preprocessing* yang meliputi data *cleaning*, *case folding*, *tokenization*, *filtering stopwords*, *stemming*, yang kemudian akan dilakukan pembagian untuk data dengan tiga rasio perbandingan, yaitu 70% data latih dan 30% data uji, 80% data latih dan 20% data uji, serta 90% data latih dan 10% data uji. Hasil dari pengumpulan data didapatkan sebanyak 586 *tweet*, setelah itu diberikan label dan didapatkan 104 *tweet* positif, 223 *tweet* netral dan 259 *tweet* negatif. Hasil dari pengujian klasifikasi *Naïve Bayes Classifier* menggunakan tiga rasio pembagian data yang berbeda-beda, berhasil mendapatkan nilai akurasi 36% untuk rasio 70:30, 43% untuk rasio 80:20 dan 36% untuk rasio 90:10. Perhitungan akurasi dilakukan dengan bantuan *Confusion Matrix*.

Kata Kunci: Klasifikasi Sentimen, *Text Mining, Naïve Bayes Classifier*

IMPLEMENTATION OF NAÏVE BAYES CLASSIFIER FOR CLASSIFICATION OF TWITTER USERS' SENTIMENT ON DPR PERFORMANCE

Abstract- - *As representatives of the people, the People's Consultative Assembly of the Republic of Indonesia (DPR RI) is responsible for addressing societal issues in a democratic and responsive manner. However, in some cases, it appears to prioritize the interests of political parties over the interests of the public. Therefore, many citizens still express dissatisfaction with the performance of the DPR. This research aims to classify the sentiment of Indonesian Twitter users to gauge the level of public satisfaction with the current performance of the DPR, using the Naïve Bayes Classifier method. The sentiment classification process involves several stages, including data collection, preprocessing which encompasses data cleaning, case folding, tokenization, stopword filtering, and stemming. Subsequently, the data will be divided into three different ratio combinations: 70% training data and 30% testing data, 80% training data and 20% testing data, and 90% training data and 10% testing data. The data collection resulted in a total of 586 tweets, which were then labeled, yielding 104 positive tweets, 223 neutral tweets, and 259 negative tweets. The results of testing the Naïve Bayes Classifier using the three different data division ratios achieved accuracy scores of 36% for the 70:30 ratio, 43% for the 80:20 ratio, and 36% for the 90:10 ratio. Accuracy calculations were performed with the assistance of a Confusion Matrix.*

Keywords: *Sentiment Classification, Text Mining, Naïve Bayes Classifier*

1. PENDAHULUAN

Dewan Perwakilan Rakyat (DPR) merupakan Lembaga legislatif di Indonesia yang bertanggung jawab dalam pembuatan undang-undang, pengawasan pemerintahan, dan mewakili suara rakyat. Sebagai perwakilan rakyat, kinerja DPR menjadi subjek perhatian publik yang penting. Dalam era digital seperti saat ini, media sosial terutama Twitter telah menjadi platform populer yang memungkinkan masyarakat untuk berbagi pendapat, salah satunya tentang kinerja DPR. *Twitter* merupakan platform media sosial yang populer dan gratis yang menyediakan layanan jaringan bagi pengguna untuk berbagi pendapat melalui pesan singkat, yang lebih dikenal dengan sebutan *tweet*. [1].

Twitter telah menjadi sumber data yang berharga dalam mempelajari sentimen publik. Dalam konteks ini, klasifikasi sentimen dapat digunakan untuk menganalisis dan memahami sikap dan pendapat pengguna Twitter terkait dengan kinerja DPR. Klasifikasi Sentimen adalah bidang penelitian yang bertujuan untuk mengkaji pendapat, sentimen, sikap, penilaian, dan emosi seseorang terhadap suatu topik tertentu dengan tujuan

menghasilkan penilaian positif, negatif, atau netral terkait topik tersebut.[2]. Melalui analisis sentimen, penelitian ini akan mencoba mengklasifikasikan mana saja *tweet* yang dianggap sebagai pandangan positif, mana *tweet* yang dianggap netral dan mana yang dianggap sebagai pandangan negatif.

Sebelumnya telah dilakukan penelitian menggunakan metode Support Vector Machine (SVM) untuk menganalisa sentimen dari pengguna *twitter* terhadap DPR RI. Pada penelitian tersebut, algoritma SVM dioptimasi dengan Particle Swarm Object untuk mendapatkan nilai akurasi yang lebih baik [3]. Berdasarkan penelitian sebelumnya, penelitian ini berfokus pada analisis pendapat masyarakat dengan menggunakan metode klasifikasi menggunakan *Naïve Bayes Classifier*. NBC (*Naïve Bayes Classifier*) adalah salah satu metode klasifikasi data yang menggunakan probabilitas sederhana. Metode ini menerapkan teorema Bayes dengan asumsi bahwa fitur-fitur yang digunakan dalam klasifikasi adalah independen satu sama lain [4].

Klasifikasi Sentimen merupakan salah satu cabang dari *text mining. Text mining* berfokus pada pengolahan kumpulan teks yang tidak terstruktur dengan tujuan mengidentifikasi pola-pola unik dengan cara mengekstraksi informasi berharga dari kumpulan teks tersebut [5]. Dalam *text mining*, langkah pertama adalah mengumpulkan data, yang kemudian perlu melalui tahap *preprocessing* sebelum dilakukan proses klasifikasi [6]. Tahap pengumpulan data adalah proses mengumpulkan data dari sumber tertentu yang di mana data tersebut nantinya akan dilakukan pemrosesan lebih lanjut [7]. Data yang sudah dikumpulkan selanjutnya masuk tahap *preprocessing*, *preprocessing* dilakukan untuk menghilangkan *noise* atau gangguan pada dokumen atau kalimat dengan tujuan mempermudah proses pengolahan data. Proses ini juga bertujuan untuk mengatasi data yang tidak sempurna, gangguan pada data, dan ketidaksesuaian data [8]. *Preprocessing* meliputi *data cleaning*, *case folding*, *tokenizing*, *filtering stopwords* dan *stemming.* Data yang sudah melalui tahap preprocessing akan diberi label sesuai dengan isi dari data yang ada, yang di mana Label positif diberikan pada *tweet* yang menunjukkan kecenderungan setuju dengan kasus yang dibicarakan, label negatif menunjukkan kecenderungan menyangkal atau menolak perdebatan tersebut [9]. Data yang sudah memiliki label sudah siap untuk masuk proses klasifikasi dan dilakukan prediksi berdasarkan pemodelan yang sudah dilakukan.

Hasil dari klasifikasi akan menjadi tujuan dari penelitian ini, yaitu mengetahui pandangan masyarakat terhadap kinerja Dewan Perwakilan Rakyat Republik Indonesia (DPR RI) serta mengukur performa nilai akurasi yang didapatkan untuk proses klasifikasi dengan metode *Naive Bayes Classifier*.

2. METODE PENELITIAN

2.1 Data Penelitian

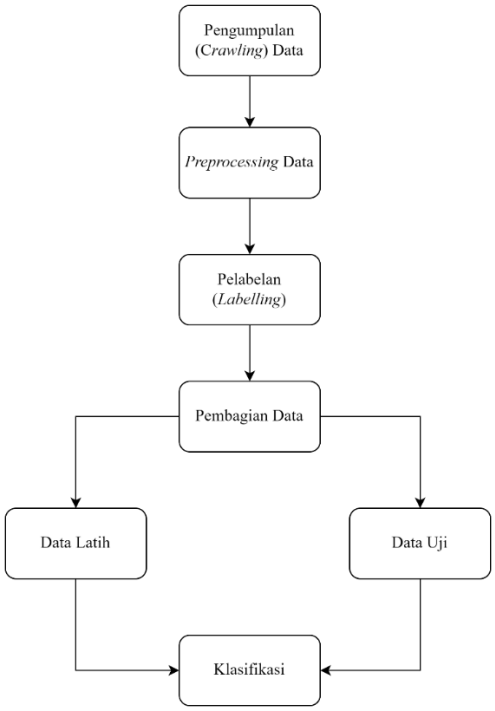
Data pada penelitian berasal dari *tweet* yang diperoleh melalui proses pengumpulan data di media sosial Twitter. Proses pengumpulan data dilakukan dengan *scraping twitter* menggunakan bahasa pemrograman *Python* dan dilakukan dalam rentang waktu 1 Januari 2022 hingga 31 Mei 2023. Terdapat total 586 data tweet yang berhasil diambil melalui proses *scraping* dengan menggunakan parameter kata kunci "kinerja dpr" dan "kerja dpr". Data yang diambil meliputi informasi *id*, teks (*tweet*), nama pengguna (*username*), dan tanggal dibuatnya *tweet* (*created_at*). Untuk *dataset* penelitian yang digunakan dapat dilihat pada tabel di bawah ini.

Tabel 1. Tabel Data Penelitian

No	*Tweet*	Waktu *Tweet* Dibuat
1	@cnbcindonesia Saya bingung dengan kerja DPR selama ini, masa aturan udah di sahin baru di kritik seharusnya dari bentuk rancangannya sudah tau dong 🙃	14 Februari 2022
2	Kerja DPR top lanjutkan bang Dasco Suami Dasco Ahmad DPR Ingatkan Mendag Minyak Goreng Penting @Gerindra @Don_dasco	16 Maret 2022
…	…	…
586	@TyaIvana6 @puanmaharani_ri Kinerja DPR makin OK	15 Juli 2022

2.2 Penerapan Metode

Terdapat beberapa tahapan yang dilakukan pada penelitian ini yang merepresentasikan setiap proses dan rancangan dalam penelitian, dari awal hingga akhir aplikasi berjalan.



Gambar 1. *Flowchart* Penerapan Metode

2.2.1 Pengumpulan Data

Dalam pengumpulan data ini, digunakan teknik *Scraping* pada Twitter dengan judul *dataset* "KinerjaDPR". *Dataset* ini terdiri dari 586 baris data yang akan dibagi menjadi data latih (*training*) dan data uji (*testing*). *Dataset* ini memiliki empat atribut, yaitu *id, tweet, username*, dan *created_at*.

2.2.2 *Preprocessing* Data

Text Preprocessing merupakan bagian dari *text mining* yang memiliki tujuan untuk menghilangkan *noise* atau gangguan pada kalimat. Proses ini bertujuan untuk mengatasi data yang kurang sempurna, mengatasi gangguan pada data, dan memperbaiki data yang tidak konsisten. Tahapan ini sangat penting karena bertujuan untuk mengolah data yang masih mentah menjadi data bersih yang siap untuk dianalisis. *Preprocessing* terdiri dari lima tahap yang mencakup:

a. *Data Cleaning*

Tahap Pembersihan Data bertujuan untuk mengatasi masalah dalam *dataset*, seperti mengganti nilai yang hilang (*missing value*), menormalisasi data yang tidak konsisten (*noisy*), mengidentifikasi dan menghapus data yang tidak konsisten dan berulang (*redundancy*) yang muncul dari penggabungan data, serta menyelesaikan masalah inkonsistensi data.

b. *Case Folding*

Konversi Huruf Kecil (*Case Folding*) adalah proses mengubah teks menjadi huruf kecil (*lowercase*). Tujuannya adalah untuk memberikan format standar pada teks agar tidak ada perbedaan yang disebabkan oleh huruf kapital atau non-kapital.

c. *Tokenizing*

Pemenggalan Teks atau Tokenizing adalah proses membagi teks menjadi bagian-bagian yang lebih kecil, yang disebut token. Pada tahap ini, dilakukan juga penghilangan angka, tanda baca, dan karakter lain yang dianggap tidak memiliki pengaruh terhadap pemrosesan teks.

d. *Filtering/Stopwords*

Tahap Penghapusan Kata-kunci (*Filtering/Stopword Removal*) adalah proses pemilihan kata-kata yang dianggap tidak relevan. Kata-kunci yang sering muncul dan dianggap tidak memberikan kontribusi signifikan, seperti kata penghubung ("dan", "akan", "atau", "yang", dll) akan dihapus.

e. *Stemming*

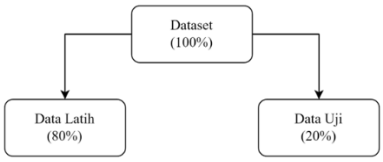
Stemming merupakan proses yang mengubah kata-kata dalam sebuah dokumen menjadi kata-kata dasar dengan menerapkan aturan tertentu. Dalam penelitian ini, digunakan pustaka Sastrawi sebagai acuan untuk melakukan proses *stemming*. Tujuannya adalah untuk menyederhanakan kata-kata dalam teks sehingga hanya tersisa kata dasar atau akar kata.

2.2.3 Pelabelan Data

Pada tahap ini, setiap data akan diberikan label atau kelas berdasarkan karakteristik dari kalimat yang terdapat pada dokumen. Data/*Tweet* yang telah dilakukan *preprocessing* akan mendapatkan label positif untuk unggahan yang mencerminkan kepuasan terhadap kinerja DPR RI, label netral diberikan untuk unggahan yang tidak terkait dengan konteks kasus, sementara label negatif diberikan pada unggahan yang mengungkap ketidakpuasan atau ketidaksetujuan terhadap kinerja DPR RI. Ahli yang memiliki keahlian relevan dalam domain penelitian ini akan membantu dalam memberikan label pada data.

2.2.4 Pembagian Data

Dalam tahap pembagian data, *dataset* yang telah diberi label akan dibagi menjadi data latih dan data uji. Data latih digunakan untuk melatih dan mengembangkan model agar dapat melakukan prediksi atau tugas lainnya dengan akurat. Sementara itu, data uji akan digunakan untuk mengukur performa atau kualitas model setelah melalui proses pelatihan menggunakan data latih [10]. Pada tahapan ini, data *tweet* yang telah diberikan label positif, netral dan negatif akan dibagi dengan tiga rasio perbandingan, yaitu 70% dan 30%, 80% dan 20%, serta 90% dan 10% masing-masing untuk data latih dan data uji.



Gambar 2. Tahapan Pembagian Data

a. Data Latih

Data latih memiliki peran penting sebagai pembangun pengetahuan dalam proses klasifikasi. Proses pembangunan pengetahuan melibatkan pemodelan untuk menghasilkan model latih menggunakan data latih yang telah tersedia.

b. Data Uji

Data uji merupakan data yang disiapkan untuk menguji tingkat keakuratan algoritme klasifikasi berdasarkan model latih.

2.2.5 Klasifikasi Data

Tahapan ini melakukan penentuan kelas untuk data uji. Yang di mana proses klasifikasi data akan menggunakan algoritme *Naïve Bayes Classifier* yang telah diberikan pengetahuan dari data latih sebelumnya. Hasil dari data klasifikasi akan diuji oleh data uji untuk menghasilkan label prediksi. Berikut persamaan yang digunakan untuk mengklasifikasi data uji.

$$\mathrm{Log}(\mathrm{P}(\mathrm{H})) + \sum \mathrm{Log}(\mathrm{P}(\mathrm{XI}|\mathrm{VJ})) \quad (1)$$

2.3 Rancangan Basis Data

Untuk membuat sistem klasifikasi sentimen terhadap kinerja DPR RI menggunakan data yang diperoleh dari Twitter, diperlukan perancangan basis data yang dapat menyimpan data yang dibutuhkan saat aplikasi berjalan.

Berikut rancangan basis data yang akan digunakan dalam sistem. Berikut adalah spesifikasi basis data yang digunakan oleh sistem ini.

Tabel 2. Tabel *Dataset*

No	Nama *Field*	Tipe Data	Lebar	Keterangan
1	*id*	*varchar*	25	ID *Tweet*
2	*username*	*varchar*	100	Nama pengguna
3	*text*	*text*	-	*Tweet*
4	*created_at*	*text*	-	Waktu *tweet* dibuat

Tabel 3. Tabel *Preprocessing*

No	Nama *Field*	Tipe Data	Lebar	Keterangan
1	*id*	*varchar*	25	ID *Tweet*
2	*username*	*varchar*	100	Nama pengguna
3	*text*	*text*	-	*Tweet*
4	*text_clean*	*text*	-	*Tweet* yang sudah melalui proses *cleaning*
5	*text_stem*	*text*	-	*Tweet* yang sudah melalui proses *stemming*
6	*text_stopwords*	*text*	-	*Tweet* yang sudah melalui proses penghilangan kata yang tidak memiliki makna
7	*created_at*	*text*	-	Waktu *tweet* dibuat
8	label	*varchar*	25	Kelas *tweet* (Aktual)

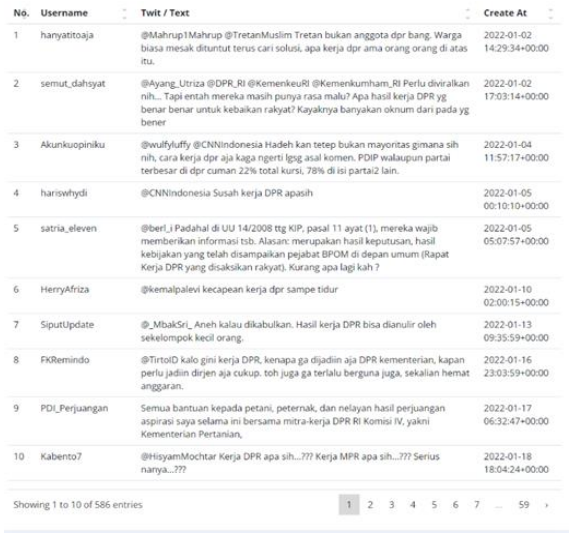
Tabel 4. Tabel Klasifikasi

No	Nama *Field*	Tipe Data	Lebar	Keterangan
1	*id*	*varchar*	25	ID *Tweet*
3	*text*	*text*	-	*Tweet* yang sudah melalui proses *stemming*
4	*label_manual*	*varchar*	25	Kelas *tweet* (Aktual)
5	*label_system*	*varchar*	25	Kelas *tweet* (Hasil Prediksi)

3. HASIL DAN PEMBAHASAN

3.1 Tahap Pengumpulan Data

Data pada penelitian ini diperoleh melalui proses *scrapping* dengan bahasa pemrograman *Python* dengan kata kunci "kinerja DPR" dan "kerja DPR". Data yang berhasil dikumpulkan akan diambil informasinya berupa *id, username, tweet dan created_at*, yang kemudian disimpan dalam *file* CSV. *File* ini yang nantinya akan diimpor ke sistem untuk diproses ke tahap berikutnya.



No.	Username	Twit / Text	Create At
1	hanyatitoaja	@Mahrup1Mahrup @TretanMuslim Tretan bukan anggota dpr bang. Warga biasa mesak dituntut terus cari solusi, apa kerja dpr ama orang orang di atas itu.	2022-01-02 14:29:34+00:00
2	semut_dahsyat	@Ayang_Utriza @DPR_RI @KemenkeuRI @Kemenkumham_RI Perlu diviralkan nih... Tapi entah mereka masih punya rasa malu? Apa hasil kerja DPR yg benar benar untuk kebaikan rakyat? Kayaknya banyakan oknum dari pada yg bener	2022-01-02 17:03:14+00:00
3	Akunkuopiniku	@wulfyluffy @CNNIndonesia Hadeh kan tetep bukan mayoritas gimana sih nih, cara kerja dpr aja kaga ngerti lgsg asal komen. PDIP walaupun partai terbesar di dpr cuman 22% total kursi, 78% di isi partai2 lain.	2022-01-04 11:57:17+00:00
4	hariswhydi	@CNNIndonesia Susah kerja DPR apasih	2022-01-05 00:10:10+00:00
5	satria_eleven	@berl_i Padahal di UU 14/2008 ttg KIP, pasal 11 ayat (1), mereka wajib memberikan informasi tsb. Alasan: merupakan hasil keputusan, hasil kebijakan yang telah disampaikan pejabat BPOM di depan umum (Rapat Kerja DPR yang disaksikan rakyat). Kurang apa lagi kah ?	2022-01-05 05:07:57+00:00
6	HerryAfriza	@kemalpalevi kecapean kerja dpr sampe tidur	2022-01-10 02:00:15+00:00
7	SiputUpdate	@_MbakSri_ Aneh kalau dikabulkan. Hasil kerja DPR bisa dianulir oleh sekelompok kecil orang.	2022-01-13 09:35:59+00:00
8	FKRemindo	@TirtoID kalo gini kerja DPR, kenapa ga dijadiin aja DPR kementerian, kapan perlu jadiin dirjen aja cukup. toh juga ga terlalu berguna juga, sekalian hemat anggaran.	2022-01-16 23:03:59+00:00
9	PDI_Perjuangan	Semua bantuan kepada petani, peternak, dan nelayan hasil perjuangan aspirasi saya selama ini bersama mitra-kerja DPR RI Komisi IV, yakni Kementerian Pertanian,	2022-01-17 06:32:47+00:00
10	Kabento7	@HisyamMochtar Kerja DPR apa sih...??? Kerja MPR apa sih...??? Serius nanya...???	2022-01-18 18:04:24+00:00

Showing 1 to 10 of 586 entries 1 2 3 4 5 6 7 ... 59 ›

Gambar 3. Tahap Pengumpulan Data

3.2 Tahap *Preprocessing*

Tahap *preprocessing* merupakan tahapan untuk membersihkan data yang tidak terstruktur menjadi lebih terstruktur dan lebih bersih dari berbagai macam *noise*. Tahap *preprocessing* meliputi *case folding, cleaning, tokenizing, filtering* (*stopwords*) dan *stemming*.

a. *Case Folding*

Case Folding merupakan proses untuk mengubah seluruh karakter pada dokumen menjadi huruf kecil (*lower case*), misalnya: 'Pemerintah' akan diubah menjadi 'pemerintah', kata 'DPR' akan diubah menjadi 'dpr' dan seterusnya.

Tabel 5. Tabel *Case Folding*

Sebelum	Sesudah
Kinerja DPR Dinilai Belum Memuaskan	kinerja dpr dinilai belum memuaskan

b. *Cleaning*

Pada proses *Cleaning* dilakukan penyaringan dan pembuangan tanda baca, *mention* akun, *hashtag*, serta berbagai *noise* lainnya.

Tabel 6. Tabel *Cleaning*

Sebelum	Sesudah
@HisyamMochtar Kerja DPR apa sih…??? Kerja MPR apa sih…??? Serius nanya…???	Kerja DPR apa sih Kerja MPR apa sih Serius nanya

c. *Tokenizing*

Proses *Tokenizing* adalah proses untuk memecah kalimat menjadi kata yang berdiri sendiri. Seperti pada contoh berikut ini: kalimat 'dukung kinerja dpr terus mendengar aspirasi rakyat' akan diubah menjadi ' dukung, kinerja, dpr, terus, mendengar, aspirasi, rakyat'.

Tabel 7. Tabel *Tokenizing*

Sebelum	Sesudah
kerja dpr apa sih kerja mpr apa sih serius nanya	kerja, dpr, apa, sih, kerja, mpr, apa, sih, serius, nanya

d. *Filtering/Stopwords*
Pada proses ini dilakukan penghapusan *stopwords* atau kata-kata yang kurang memiliki makna namun sering dijumpai pada teks.

Tabel 8. Tabel *Filtering/Stopwords*

Sebelum	Sesudah
justru di kwalitas semua hasil nya merugikan rakyat dan etika kerja dpr bobrok	kwalitas hasil merugikan rakyat etika kerja dpr bobrok

e. *Stemming*
Pada tahapan ini dilakukan proses mengubah setiap kata pada dokumen menjadi kata dasar. Seperti pada kata 'memuaskan' akan diubah menjadi 'puas'.

Tabel 9. Tabel *Stemming*

Sebelum	Sesudah
kwalitas hasil merugikan rakyat etika kerja dpr bobrok	kwalitas hasil rugi rakyat etika kerja dpr bobrok

3.3 Tahap Pelabelan

Dalam penelitian ini, tahap pelabelan dilakukan secara manual oleh peneliti untuk memberi label pada seluruh data yang terdiri dari 586 *tweet*. Setelah pelabelan, data akan memiliki label negatif dan positif. Terdapat 104 *tweet* yang memiliki label positif, 223 *tweet* 223 *tweet* yang diberi label netral, dan label negatif terdiri dari 259 *tweet*. Untuk memastikan keakuratan pelabelan, hasil data yang telah dilabeli juga diverifikasi oleh Dr. Masrizal, M.A., dosen dari Program Studi Sosiologi Fakultas Sosial dan Politik di Universitas Syiah Kuala. Beliau dipilih sebagai ahli untuk tahap pemberian label data karena posisinya sebagai seorang pengajar di bidang ilmu pemerintahan dan politik, yang sangat relevan dengan konteks penelitian ini.

3.4 Tahap Pembagian Data

Pada tahapan pembagian data, *dataset* yang telah diberi label akan dilakukan proses pembagian menjadi data latih dan data uji. Pada penelitian ini, proses pembagian data latih dan data uji dilakukan dengan 3 rasio perbandingan, yaitu masing-masing untuk data latih dan data uji 70:30, 80:20 dan 90:10.

Tabel 10. Tabel Pembagian Data Rasio 80:20

Jenis Data	Jumlah
Data Latih	469
Data Uji	117
Total *Dataset*	586

3.5 Tahap Klasifikasi *Naïve Bayes Classifier*

Tahap ini melibatkan proses klasifikasi data dengan menggunakan Algoritme *Naïve Bayes*. Algoritme ini digunakan untuk menghitung dan mengklasifikasikan data, sehingga menghasilkan prediksi pada *dataset* yang sudah memiliki label sebelumnya. Proses pertama yang dilakukan pada tahapan ini adalah mencari nilai probabilitas tiap kelas label, data yang digunakan adalah data yang sudah melewati tahap *preprocessing* dan sudah diberikan label.

Tabel 11. Tabel Jumlah Data Latih

Jumlah Data Latih Positif	Jumlah Data Latih Netral	Jumlah Data Latih Negatif	Jumlah Seluruh Data Latih
84	178	207	469

Untuk mendapatkan nilai *prior probability*, bisa dilakukan dengan menggunakan persamaan berikut.

$$P(H) = \frac{\text{jumlah data latih setiap kelas}}{\text{Jumlah seluruh data latih}} \quad (2)$$

Berikut perhitungan *prior probability* untuk label positif.
P(Positif) = $\frac{84}{469}$
P(Positif) = 0,179

Berikutnya perhitungan *prior probability* untuk label netral.
P(Netral) =$\frac{178}{469}$
P(Netral) = 0,379

Terakhir perhitungan *prior probability* untuk label negatif.
P(Negatif) = $\frac{207}{469}$
P(Negatif) = 0,442

Setelah mendapatkan nilai *prior probability* untuk setiap label, proses berikutnya adalah proses klasifikasi dari contoh data uji.

Jumlah kemunculan kata = 4312
Kemunculan kata (Positif) = 914
Kemunculan kata (Negatif) = 1928
Kemunculan kata (Netral) = 1480

Tabel 12. Tabel Data Uji

Tweet	Label Aktual
Gak puas lihat kerja dpr sangat lamban	Negatif

$$P(X_I|V_J) = \frac{\text{jumlah kemunculan kata di setiap kelas}}{\text{kemunculan kata kelas}} \quad (3)$$

Tabel 13. Tabel Perhitungan Data Uji Pada Label Positif

Kata	Frekuensi Kemunculan Kata	$P(X_I\|V_J)$	Hasil
gak	4	4 / 914	0,004
puas	19	19 / 914	0,020
lihat	5	5 / 914	0,005
kerja	114	114 / 914	0,124
dpr	109	109 / 914	0,119
sangat	1	1 / 914	0,001
lamban	0	0 / 914	0

Tabel 14. Tabel Perhitungan Data Uji Pada Label Negatif

Kata	Frekuensi Kemunculan Kata	$P(X_I\|V_J)$	Hasil
gak	167	167 / 1928	0,086
puas	5	5 / 1928	0,003
lihat	19	19 / 1928	0,010
kerja	259	259 / 1928	0,134
dpr	253	253 / 1928	0,131
sangat	0	0 / 1928	0
lamban	89	89 / 1928	0,046

Tabel 15. Tabel Perhitungan Data Uji Pada Label Netral

Kata	Frekuensi Kemunculan Kata	$P(X_I\|V_J)$	Hasil
gak	14	14 / 1480	0,009
puas	1	1 / 1480	0,001
lihat	1	1 / 1480	0,001
kerja	187	187 / 1480	0,126
dpr	152	152 / 1480	0,102
sangat	1	1 / 1480	0,001
lamban	0	0 / 1480	0

Tabel 16. Tabel Perhitungan Klasifikasi Data Uji

Label	$\text{Log}(P(H)) + \sum\text{Log}(P(X_I\|V_J))$	Hasil (*max Score*)
Positif	Log(0,179) + (Log(0,004) + Log(0,020) + Log(0,005) + Log(0,124) + Log(0,119) + Log(0,001) + 0)	-11,974
Negatif	Log(0,442) + (Log(0,086) + Log(0,003) + Log(0,010) + Log(0,134) + Log(0,131) + 0 + Log(0,046))	**-9,072**
Netral	Log(0,379) + (Log(0,009) + Log(0,001) + Log(0,001) + Log(0,126) + Log(0,102) + Log(0,001) + 0)	-13,177

Berdasarkan hasil perhitungan di atas, skor untuk klasifikasi negatif memiliki nilai tertinggi. Oleh karena itu, tweet tersebut diklasifikasikan sebagai tweet dengan sentimen negatif.

3.6 Tahap Perhitungan Akurasi

Pada penelitian ini, tahap penghitungan akurasi menggunakan Confusion Matrix. Confusion matrix merupakan tabel yang membandingkan hasil klasifikasi yang dilakukan oleh sistem (prediksi) dengan hasil klasifikasi yang sebenarnya. Tabel ini menunjukkan jumlah data uji yang diklasifikasikan dengan benar dan jumlah data uji yang diklasifikasikan dengan salah.[11]. Pengujian dilakukan dengan tiga rasio pembagian data latih, 70% : 30%, 80% : 20%, 90% : 10%. Berikut tahap perhitungan akurasi dengan rasio pembagian data 80% : 20%. Berikut persamaan yang digunakan dalam perhitungan pengujian.

$$\text{Akurasi} = \frac{\text{TPos} + \text{TNeu} + \text{TNeg}}{\text{TPos} + \text{TNeu} + \text{TNeg} + \text{FPos} + \text{FNeu} + \text{FNeg}} \quad (4)$$

$$\text{Presisi} = \frac{\text{Presisi Positif} + \text{Presisi Netral} + \text{Presisi Negatif}}{\text{Jumlah Kelas}} \quad (5)$$

$$\textit{Recall} = \frac{\textit{Recall}\ \text{Positif} + \textit{Recall}\ \text{Netral} + \textit{Recall}\ \text{Negatif}}{\text{Jumlah Kelas}} \quad (6)$$

$$\textit{F-1 Score} = \frac{2 * \text{Presisi} * \textit{Recall}}{\text{Presisi} + \textit{Recall}} \quad (7)$$

Tabel 17. Tabel Hasil Pengujian

Proporsi *Dataset*		Hasil Pengujian		
Training	*Testing*	Akurasi	*Presisi*	*Recall*
70% (410)	30% (176)	36%	47%	53%
80% (469)	20% (117)	43%	47%	61%
90% (527)	10% (59)	36%	47%	55%

Berdasarkan hasil penelitian yang dilakukan, nilai akurasi dengan rasio perbandingan data 80:20 menjadi yang paling tinggi yaitu 43%, sedangkan rasio pembagian data 70:30 dan 90:10 mendapatkan hasil yang lebih rendah di angka 36%. Rasio pembagian data 80:20 bisa mendapatkan nilai akurasi yang lebih tinggi dikarenakan kualitas data latih yang digunakan lebih sempurna dibandingkan rasio pembagian data 70:30 dan 90:10.

4. KESIMPULAN

Dari hasil pengujian sistem yang telah dibuat menggunakan dataset dan algoritme yang diajukan, terlihat bahwa dari 586 *tweet* terkait dengan kinerja DPR, dengan tiga pengujian klasifikasi menggunakan metode *Naïve Bayes Classifier* dan menggunakan tiga rasio pembagian data yang berbeda-beda, berhasil mendapatkan nilai akurasi 36% untuk rasio 70:30, 43% untuk rasio 80:20 dan 36% untuk rasio 90:10. Penelitian ini melibatkan beberapa tahap utama, termasuk pengumpulan data, *preprocessing*, pelabelan, pembagian data, dan klasifikasi. Tahap *preprocessing* memiliki peran penting dalam penelitian ini, karena hasil yang baik dari *preprocessing* dapat mempengaruhi hasil klasifikasi yang optimal.

UCAPAN TERIMA KASIH

Penulis berterima kasih sebesar-besarnya kepada Bapak Dr, Masrizal, M.A. yang telah bersedia meluangkan waktunya untuk membantu penulis dalam proses pelabelan data dalam penelitian ini, sehingga data pada penelitian ini dapat digunakan dengan baik untuk proses klasifikasi.

DAFTAR PUSTAKA

[1] Styawati, N. Hendrastuty, A. R. Isnain, and A. Y. Rahmadhani, "Analisis Sentimen Masyarakat Terhadap Program Kartu Prakerja Pada Twitter Dengan Metode Support Vector Machine," *Jurnal Informatika: Jurnal pengembangan IT (JPIT)*, pp. 150–155, 2021.

[2] A. P. Natasuwarna, "Analisis Sentimen Keputusan Pemindahan Ibukota Negara Menggunakan Klasifikasi Naive Bayes," *Seminar Nasional Sistem Informasi dan Teknik Informatika*, pp. 47–53, 2019.

[3] A. Faisal, Y. Alkhalifi, A. Rifai, and W. Gata, "Analisis Sentimen Dewan Perwakilan Rakyat Dengan Algoritma Klasifikasi Berbasis Particle Swarm Optimization," *JOINTECS (Journal of Information Technology and Computer Science)*, pp. 61–70, May 2020.

[4] D. Normawati and S. A. Prayogi, "Implementasi Naïve Bayes Classifier Dan Confusion Matrix Pada Analisis Sentimen Berbasis Teks Pada Twitter," *Jurnal Sains Komputer & Informatika (J-SAKTI)*, pp. 697–711, 2019.

[5] A. Sabrani, I. W. Wedashwara W., and F. Bimantoro, "Multinomial Naïve Bayes untuk Klasifikasi Artikel Online tentang Gempa di Indonesia," *Jurnal Teknologi Informasi, Komputer, dan Aplikasinya*, pp. 80–100, 2020.

[6] B. M. Pintoko and K. Muslim L., "Analisis Sentimen Jasa Transportasi Online pada Twitter Menggunakan Metode Naïve Bayes Classifier," *e-Proceeding of Engineering*, pp. 8121–8130, 2018.

[7] M. Ghifari, R. R. Santika, and S. Waluyo, "ANALISIS SENTIMEN TWITTER TERHADAP KENAIKAN BAHAN BAKAR MINYAK MENGGUNAKAN ALGORITMA NAÏVE BAYES," *Seminar Nasional Mahasiswa Fakultas Teknologi Informasi (SENAFTI)*, pp. 219–226, 2023.

[8] F. V. Sari and A. Wibowo, "Analisis Sentimen Pelanggan Toko Online JD.ID Menggunakan Metode Naïve bayes Clasifier Berbasis Konversi Ikon Emosi," *Jurnal Teknik Industri, Mesin, Elektro dan Ilmu Komputer*, pp. 681–686, 2019.

[9] M. Priandi and Painem., "Analisis Sentimen Masyarakat Terhadap Pembelajaran Daring di Era Pandemi Covid-19 pada Media Sosial Twitter Menggunakan Ekstraksi Fitur Countvectorizer dan Algoritme K-Nearest Neighbor," *Seminar Nasional Mahasiswa Ilmu Komputer dan Aplikasinya (SENAMIKA)*, pp. 311–319, 2021.

[10] M. P. Wibowo, S. Amini, Indra, and D. Kusumaningsih, "ANALISIS SENTIMEN MASYARAKAT INDONESIA PADA TWITTER TERHADAP ISU RESESI 2023 MENGGUNAKAN METODE NAIVE BAYES," *Seminar Nasional Mahasiswa Fakultas Teknologi Informasi (SENAFTI)*, vol. 2, pp. 201–210, Apr. 2023.

[11] M. I. Fikri, T. S. Sabrila, and Y. Azhar, "Perbandingan Metode Naïve Bayes dan Support Vector Machine pada Analisis Sentimen Twitter," *SMATIKA JURNAL*, vol. 10, pp. 71–76, Dec. 2020.