

Implementasi *Naïve Bayes Classifier* Dan *Confusion Matrix* Pada Analisis Sentimen Berbasis Teks Pada *Twitter*

Dwi Normawati^{1*}, Surya Allit Prayogi²

Program Studi Teknik Informatika Universitas Ahmad Dahlan

Ringroad Selatan, Kampus 4 UAD, Yogyakarta 55191

e-mail: dwinorma@tif.uad.ac.id^{1*}, surya1500018124@webmail.uad.ac.id²

Abstract

Twitter is one of the social media that is currently in great demand by internet users. The number of tweets circulating on Twitter is not yet known whether these tweets contain more positive, negative, and neutral opinions. For that we need a system that can process data by applying sentiment analysis. This study uses the Naïve Bayes Classifier (NBC) method to analyze the level of sentiment towards data carried out by crawling on Twitter. The data studied as a simple case study uses only 8 tweet data which is divided into 5 training data and 3 test data. The data is processed using the preprocessing stage, then classified using the NBC method, the calculation of performance uses confusion matrix techniques. This study resulted in a structured exposure to the process and results of NBC implementation and performance testing using the confusion matrix which obtained 82% accuracy, 93% precision, and 52% recall. However, these results are more focused on ease explaining for each stage and process in more detail, not on the numbers obtained. Research with larger data will be carried out later by developing a computer-based application system.

Keywords: *Naïve Bayes Classifier, Confusion Matrix, Text Clasification, Twitter.*

Abstrak

Twitter adalah salah satu media sosial yang saat ini sangat diminati oleh pengguna internet. Banyaknya cuitan yang beredar di Twitter belum dapat diketahui apakah cuitan tersebut lebih banyak mengandung opini positif, negatif, dan netral. Untuk itu diperlukan sistem yang dapat mengolah data dengan menerapkan analisis sentimen. Penelitian ini menggunakan metode Naïve Bayes Classifier (NBC) untuk menganalisis tingkat sentimen terhadap data yang dilakukan dengan teknik crawling pada Twitter. Data yang diteliti sebagai studi kasus sederhana hanya menggunakan sebanyak 8 data cuitan yang dibagi menjadi 5 data latih dan 3 data uji. Data tersebut diolah dengan tahap preprocessing, lalu diklasifikasi menggunakan metode NBC, perhitungan performa menggunakan teknik confusion matrix. Penelitian ini menghasilkan pemaparan yang terstruktur pada proses dan hasil implementasi NBC dan pengujian performa menggunakan confusion matrix yang didapatkan akurasi sebesar 82%, presisi 93%, dan recall sebesar 52%. Namun demikian hasil ini lebih difokuskan untuk kemudahan penjelasan setiap tahapan dan proses secara lebih detil, bukan fokus pada angka-angka yang didapatkan. Penelitian dengan data yang lebih besar dilakukan kemudian menggunakan sistem aplikasi berbasis komputer.

Kata Kunci: *Naïve Bayes Classifier, Confusion Matrix, Klasifikasi Teks, Twitter*

1. PENDAHULUAN

Media sosial *Twitter* saat ini menjadi aplikasi yang diminati oleh pengguna internet. Riset per November 2019 menyebutkan bahwa di Indonesia tercatat sebanyak 78 juta pengguna *Twitter* [1]. Banyaknya *netizen* memberikan opini melalui cuitan, komentar, atau kritikan terhadap Isu-isu

yang sedang hangat menjadi berita yang *trending*. Salah satu berita yang sering muncul menjadi *trending topic* di *Twitter* adalah kontroversi tentang Ahok. Ahok terbukti bersalah melanggar Pasal 156a KUHP tentang penodaan agama [2]. Ahok dibebaskan dari penjara pada tanggal 24 Januari 2019 namun tak lama kemudian masyarakat Indonesia dihebohkan dengan berita Ahok diangkat sebagai Komisaris Utama Pertamina [3]. Berita tersebut menjadi *viral* di berbagai media sosial khususnya *Twitter*. Banyak cuitan dari netizen namun belum dapat diketahui apakah cuitan tersebut lebih banyak mengandung opini positif, negatif, ataupun netral. Untuk itu diperlukan sistem yang dapat mengolah data dengan menerapkan analisis sentimen.

Analisis sentimen adalah penambangan teks yang menganalisis dan mengekstraksi informasi yang bersifat subjektif dan dapat membantu pihak yang membutuhkan [4]. Penelitian tentang analisis sentimen ini sebelumnya pernah dilakukan seperti Analisis Sentimen Calon Gubernur DKI Jakarta 2017 di *Twitter* Menggunakan Metode *Lexicon Based*, *Support Vector Machine* (SVM), dan *Naive Bayes Classifier NBC* [5]. Penelitian ini melakukan riset tentang opini masyarakat yang mengandung sentimen positif, netral, atau negatif Calon Gubernur DKI Jakarta 2017 di *Twitter*. Akurasi paling tinggi ketika menggunakan metode klasifikasi NBC dengan rata-rata akurasi mencapai 95%, presisi 95%, *recall* 95%, *True Positive (TP) rate* 96,8%, dan *True Negative (TN) rate* 84,6%. Pada SVM didapatkan rata-rata akurasi 90%, presisi 89,9%, *recall* 90%, *TP rate* 98,4%, dan *TN rate* 38,5%.

Penelitian ini mengambil contoh studi kasus yang bertujuan untuk mengetahui opini masyarakat terhadap berita kontroversi Ahok di *Twitter* berdasarkan tiga klasifikasi, yaitu komentar positif, netral, dan negatif. Metode NBC dipilih karena tingkat akurasi yang tinggi, mudah untuk dipahami, dan memiliki cepat dalam mengklasifikasi data [6]. Penelitian ini diharapkan mampu menjelaskan langkah-langkah secara terstruktur dalam mengimplementasikan metode NBC dalam mengklasifikasi teks. Analisis Sentimen Masyarakat terhadap Calon Presiden Indonesia 2014 berdasarkan Opini dari *Twitter* Menggunakan Metode *Naive Bayes Classifier* (NBC) melakukan klasifikasi sentimen analisis terhadap komentar atau cuitan dari masyarakat terhadap Calon Presiden Indonesia Tahun 2014. Komentar tersebut diambil melalui *Twitter* dengan beberapa tahapan, yaitu pengumpulan data, *preprocessing* data, *POS Tagging*, ekstraksi opini menggunakan *rule based*, dan klasifikasi opini. Jumlah percakapan pada pasangan Prabowo–Hatta Rajasa lebih unggul sebesar 53% dibandingkan Jokowi–Jusuf Kalla sebesar 47%. Hasil dari pengamatan polaritas sentimen masyarakat menunjukkan Prabowo–Hatta Rajasa mendapatkan 47,7% sentimen positif, 26,4% negatif, dan 25,9% netral. Sedangkan pasangan Jokowi–Jusuf Kalla mendapatkan 37,6% sentimen positif, 34,4% negatif, dan 27,9% netral. Dari hal ini disimpulkan bahwa Prabowo–Hatta Rajasa lebih unggul dalam hal jumlah percakapan dan sentimen positif pada *Twitter* [6].

Analisis sentimen data presiden Jokowi dengan *Preprocessing* Normalisasi dan *Stemming* Menggunakan Metode NBC dan *Support Vector*

Machine (SVM) [7] melakukan klasifikasi sentimen analisis terhadap komentar masyarakat terhadap Presiden Jokowi. Komentar tersebut diambil melalui media sosial dan beberapa blog politik menggunakan teknik *search techniques*. Data yang telah didapat dievaluasi melalui *preprocessing*, yaitu normalisasi dan *stemming* dengan bantuan aplikasi Sastrawi Master. Tahap selanjutnya adalah tokenisasi *N-Gram*, *Unigram*, *Bigram*, dan *Trigram* terhadap kalimat, kemudian menghilangkan kata-kata yang umum digunakan dan tidak memiliki informasi yang berharga pada konteks atau disebut *stopword removal*, dan mempertahankan *emoticon* sebagai simbol ekspresi dalam tulisan. Akurasi yang terbaik pada penelitian ini adalah dengan melakukan normalisasi dan *stemming* pada data sebesar 89,2655% menggunakan SVM.

Analisis Sentimen Opini Publik Berita Kebakaran Hutan melalui Komparasi Algoritma *Support Vector Machine* dan *K-Nearest Neighbor* berbasis *Particle Swarm Optimization* [8] melakukan klasifikasi sentimen pada pengguna sosial media yang memberikan komentar positif dan negatif terhadap berita kebakaran hutan yang berdampak kabut asap. Pengujian dilakukan terhadap *review* opini publik berita kebakaran hutan yang dikumpulkan melalui *online news* didapatkan 360 data (180 positif dan 180 negatif) menggunakan metode SVM, SVM berbasis PSO (SVM+PSO), k-NN, dan k-NN berbasis PSO (k-NN+PSO). SVM memiliki akurasi 80,83% dan *Area Under The Curve* (AUC) sebesar 0,947, sedangkan SVM+PSO menghasilkan akurasi 86,11% dan AUC sebesar 0,922.

Text Mining Dalam Analisis Sentimen Asuransi pada Media Sosial Facebook Menggunakan Metode *Naïve Bayes Classifier* [9] melakukan klasifikasi analisis sentimen pada pengguna yang memberikan tanggapan terhadap kepuasan pelayanan pada sebuah perusahaan Asuransi melalui media sosial opini publik yang ada pada komentar media sosial Facebook. Penelitian ini telah menghasilkan sistem analisis sentimen asuransi yang mampu mentransformasi sentimen berupa teks ulasan pada *fanpage* sehingga dapat menampilkan informasi sentimen masyarakat yang bersifat positif dan negatif. Akurasi pada sistem yang dibangun mencapai 95%.

2. METODOLOGI PENELITIAN

Penelitian ini lebih fokus pada tahapan-tahapan proses implementasi dan hasil pada *preprocessing*, NBC, dan *confusion matrix*. Sebagai studi kasus hanya disampaikan data terbatas untuk kemudahan pemahaman sehingga secara konsep lebih mudah dipahami. Implementasi dengan data yang besar akan dilakukan pada penelitian selanjutnya.

2.1. *Preprocessing*

Text mining merupakan teknik penambangan data untuk melakukan analisis pada opini di dalam teks tertulis. Media sosial adalah salah satu media yang dapat diterapkan teknik ini dengan tujuan untuk evaluasi dan menunjang organisasi untuk penyampaian informasi [9]. Analisis sentimen

adalah penambahan teks dengan menganalisis, memahami, mengolah, dan mengekstrak data tekstual yang bersifat opini, emosi, dan sikap terhadap suatu objek. Teks yang ditambang adalah data yang akan diteliti. Data diambil dari komentar, ulasan, atau testimoni dari produk yang akan diteliti dengan tujuan untuk mengetahui secara umum opini masyarakat terhadap objek yang dianalisis. Klasifikasi analisis sentimen dibagi menjadi tiga kelas, yaitu positif, negatif, dan netral. Tahap ini merupakan proses yang penting untuk menentukan dokumen yang memiliki dan menyimpulkan opini bernilai positif, negatif, atau netral [6].

Crawling merupakan pengunduhan atau pengambilan data dengan tujuan untuk mengumpulkan data dari suatu *database*, misalnya data yang diunduh dari *server Twitter* berupa *user* dan *tweet* termasuk atributnya menggunakan fasilitas *add-one Twitter archiver* yang disediakan oleh *Google* pada *Spreadsheet* [10]. *Pre-processing* bertujuan untuk menghilangkan *noise* dan menyeragamkan bentuk kata untuk mengurangi *volume* kosakata [11]. Langkah-langkah *preprocessing* adalah sebagai berikut.

- 1) *Cleaning* adalah menghilangkan karakter pada teks berupa *hashtag*, *url*, *mention*, dan *symbol* sehingga menghasilkan data cuitan yang asli.
- 2) *Tokenizing* adalah memotong setiap kata pada dokumen yang bertujuan untuk mengubah huruf kapital menjadi huruf kecil. Tahap ini hanya memproses huruf sehingga karakter atau tanda baca lainnya dihapus. Hanya huruf 'a' sampai dengan 'z' yang diterima.
- 3) *Filtering* atau *stopword* adalah proses mengambil kata penting dari hasil *tokenizing*. Terdapat algoritma *stoplist* didalamnya yang bertujuan untuk membuang kata-kata yang tidak penting dan *wordlist* yang berfungsi sebagai tempat menyimpan kata-kata penting. Kata yang dibuang adalah kata yang tidak deskriptif, contohnya yang, dan, di, dan dari.
- 4) *Stemming* adalah proses mengubah kata hasil dari *filtering* ke bentuk dasarnya dengan cara menghilangkan imbuhan-imbuhan pada kata dalam dokumen, misal membela menjadi bela, menguatkan menjadi menguat, menjaga menjadi jaga, dikatakan menjadi kata, dan lain-lain.

*TF*ID* merupakan metode untuk mencari bobot suatu kata dalam dokumen kunci di setiap kategori dan mencari kata kunci yang hampir mirip dengan kategori yang tersedia. Pembobotan ini menggunakan teknik *Term Frequency and Inverse Document Frequency* (TF*IDF) untuk mengekstraksi ciri dari suatu teks dengan cara menggabungkan metode TF dan IDF [12] dengan persamaan berikut.

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \quad (1)$$

Kelemahan metode TF adalah jika suatu kata muncul pada semua dokumen maka tidak bisa dipastikan bahwa kata tersebut memiliki makna khusus yang relevan sehingga digunakan IDF untuk menghitung jumlah dokumen yang mengandung kata yang dimaksud, kemudian dibagi dengan total dokumen yang ada [12].

$$idf_i = \log \frac{|D|}{|\{D_j: t_i \in d_j\}|} \quad (2)$$

Dengan tujuan mendapatkan pembobotan yang sesuai untuk tiap *term* dalam tiap dokumen, maka dilakukan kombinasi metode TF dan IDF [12].

$$tf \times idf_{i,j} = tf_{i,j}idf_i \quad (3)$$

2.1 Klasifikasi Menggunakan Naïve Bayes Classifier

NBC adalah salah satu metode untuk mengklasifikasi data dengan probabilitas sederhana yang mengaplikasikan *teorema bayes* dengan karakter independen yang tinggi. Metode ini sesuai untuk banyak dataset dengan performa yang cepat dalam mengklasifikasi data dan memiliki akurasi tinggi [6].

Dasar dari *Naïve Bayes* yang dipakai dalam pemrograman adalah rumus Bayes, yaitu peluang kejadian A sebagai B ditentukan dari peluang B saat A, peluang A, dan peluang B seperti persamaan berikut.

$$P(A|B) = \frac{P(B|A) \times P(A)}{P(B)} \quad (4)$$

Pada aplikasinya rumus ini berubah menjadi:

$$P(C_i|D) = \frac{P(D|C_i) \times P(C_i)}{P(D)} \quad (5)$$

NBC bisa disebut sebagai *Multinomial Naïve Bayes* merupakan model penyederhanaan Metode Bayes yang cocok dalam klasifikasi teks atau dokumen dengan persamaan berikut.

$$V_{MAP} = \arg \max P(v_j | a_1, a_2, a_3 \dots a_n) \quad (6)$$

Menurut persamaan (3) maka persamaan (6) dapat ditulis:

$$V_{MAP} = \arg \max_{v_j \in V} \frac{P(a_1, a_2, a_3 \dots a_n | v_j) P(v_j)}{P(a_1, a_2, a_3 \dots a_n)} \quad (7)$$

dan dapat diturunkan menjadi:

$$V_{MAP} = \arg \max_{v_j \in V} P(a_1, a_2, a_3 \dots a_n | v_j) P(v_j) \quad (8)$$

Karena $P(a_1, a_2, \dots, a_n | v_j)$ sulit untuk dihitung maka diasumsikan bahwa setiap kata di dalam dokumen tidak memiliki keterkaitan.

Rumus pengujian sentiment adalah sebagai berikut.

$$V_{MAP} = \arg \max_{v_j \in V} P(a_i | v_j) \times \prod P(v_j) \quad (9)$$

Rumus hitung data uji:

$$P(v_j) = \frac{|Dok_i|}{|training|} \quad (10)$$

$$P(a_i | v_j) = \frac{n_i + 1}{n + |kosakata|} \quad (11)$$

Keterangan:

$P(v_j)$ = probabilitas setiap dokumen pada sekumpulan dokumen

$P(a_i | v_j)$ = probabilitas kemunculan kata a_i pada suatu dokumen dengan kategori kelas v_j

- $|Dok_i|$ = frekuensi dokumen pada setiap kategori
- $|training|$ = jumlah dokumen training yang ada
- n_i = frekuensi kata ke-k pada setiap kategori
- $|kosakata|$ = jumlah kosakata yang ada pada dokumen uji

2.2. Evaluasi Performa Menggunakan Confusion Matrix

Confusion matrix adalah tabel yang menyatakan klasifikasi jumlah data uji yang benar dan jumlah data uji yang salah. Contoh *confusion matrix* untuk klasifikasi biner ditunjukkan pada Tabel 1 [13].

Tabel 1. *Confusion matrix*

		Kelas Prediksi	
		1	0
Kelas Sebenarnya	1	TP	FN
	0	FP	TN

Keterangan:

- TP (*True Positive*) = jumlah dokumen dari kelas 1 yang benar diklasifikasikan sebagai kelas 1
- TN (*True Negative*) = jumlah dokumen dari kelas 0 yang benar diklasifikasikan sebagai kelas 0
- FP (*False Positive*) = jumlah dokumen dari kelas 0 yang salah diklasifikasikan sebagai kelas 1
- FN (*False Negative*) = jumlah dokumen dari kelas 1 yang salah diklasifikasikan sebagai kelas 0

Rumus *confusion matrix* untuk menghitung *accuracy*, *precision*, dan *recall* seperti berikut.

$$accuracy = \frac{TP+TN}{Total} \quad (12)$$

$$precision = \frac{TP}{TP+FP} \quad (13)$$

$$recall = \frac{TP}{TP+FN} \quad (14)$$

3. HASIL DAN PEMBAHASAN

Setelah mendapatkan kephahaman pada landasan teori maka selanjutnya melakukan perhitungan pada studi kasus dengan lima data cuitan yang diambil dari *crawling* pada *Twitter* yang akan diklasifikasi dengan metode *Naïve Bayes Classifier* (NBC). Untuk pembahasan yang lebih terstruktur dan mudah dipahami, pembahasan dilakukan dengan beberapa sub bab dibuat berurutan dari proses dan hasil *preprocessing*, implementasi dan hasil NBC, serta evaluasi performa.

3.1. Proses dan Hasil Preprocessing

Data penelitian sebagai studi kasus diambil lima data cuitan yang bersifat mentah dan terdapat karakter, simbol, singkatan, dan bahasa asing yang perlu diolah terlebih dahulu seperti pada Tabel 2 berupa cuitan atau data mentah hasil *crawling*.

Tabel 2. Data mentah hasil *crawling*

Dokumen	Cuitan
1	@hanny_tiwa Selamat Bekerja pak ahok @basuki_btp #AhokisBack
2	@bintangku206 AHOK itu penista Tapi Ahok Dilindungi
3	@jihyunpetzie Selamat bertugas ya pak @basuki_btp #AhokisBack
4	@vivanewscom Ahok Tidak Pantas Dikasih Jabatan #vivanews
5	@muit1924 Penista bisa jadi pejabat.

Tabel 3. Proses dan hasil *preprocessing*

<i>Case folding</i>	<i>Cleaning</i>	<i>Tokenizing</i>	<i>Filtering</i>	<i>Stemming</i>
@hanny_tiwa selamat bekerja pak ahok @basuki_btp #ahokisback	selamat bekerja pak ahok	selamat, bekerja, pak, ahok	selamat, bekerja, ahok	selamat, kerja, ahok
@bintangku206 ahok itu penista tapi ahok dilindungi	ahok itu penista tapi ahok dilindungi	ahok, itu, penista, tapi, ahok, dilindungi	ahok, penista, lindung	ahok, nista, lindung
@jihyunpetzie selamat bertugas ya pak @basuki_btp #AhokisBack	Selamat bertugas ya pak	selamat, bertugas, ya, pak	selamat, bertugas	selamat, tugas
@vivanewscom ahok tidak pantas dikasih jabatan #vivanews	jawara bekasi ahok tidak pantas dikasih jabatan	ahok, tidak, pantas, dikasih, jabatan	ahok, tidak, pantas, dikasih, jabatan	jawara, bekasi, ahok, tidak, pantas, kasih, jabatan
@muit1924 penista jadi pejabat.	penista jadi pejabat	penista, jadi, pejabat	penista, jadi, pejabat	nista, jadi, pejabat

Data mentah hasil *crawling* yang memiliki huruf kapital diubah menjadi huruf kecil melalui proses *case folding*, kemudian dilakukan *cleaning* atau penghilangan karakter, URL, *mention*, dan *hashtag* sehingga didapatkan cuitan yang sebenarnya. Selanjutnya proses *tokenizing* dilakukan pemotongan atau pemenggalan kata pada data cuitan yang selanjutnya dipisahkan dengan tanda koma (,). *Filtering* kemudian dilakukan untuk pengambilan kata-kata

penting setelah proses *tokenizing* *Stemming* sebagai proses akhir *preprocessing* memastikan data yang masih mengandung kata imbuhan dihilangkan untuk mendapatkan kumpulan data yang berisi kata dasar (*root*). Proses dan hasil dari *case folding* hingga *stemming* dapat dilihat pada Tabel 3.

3.2. Implementasi dan Hasil NBC

Setelah melakukan tahapan *preprocessing* selanjutnya adalah implementasi NBC. Proses diawali dengan memberikan label pada setiap hasil *stemming* dengan tiga kelas sentimen yang diberikan kode 0 adalah positif, 1 adalah negatif, dan 2 adalah netral seperti pada Tabel 4. Kemunculan setiap kata dalam dokumen diidentifikasi kemunculannya seperti pada Tabel 5.

Tabel 4. Hasil *labelling*

Dokumen	Cuitan	Kelas Sentimen
1	selamat, kerja, ahok	0
2	ahok, nista, lindung	1
3	selamat, tugas	0
4	ahok, tidak, pantas, kasih, jabatan	1
5	nista, jadi, pejabat	1

Tabel 5. Kata yang muncul dalam dokumen

Dokumen	selamat	kerja	ahok	nista	lindung	tugas	tidak	pantas	kasih	jabatan	jadi	pejabat	Kelas sentimen
1	1	1	1	0	0	0	0	0	0	0	0	0	+
2	0	0	1	1	1	0	0	0	0	0	0	0	-
3	1	0	0	0	0	1	0	0	0	0	0	0	+
4	0	0	1	0	0	0	1	1	1	1	0	0	-
5	0	0	0	1	0	0	0	0	0	0	1	1	-

Dokumen kelas positif seperti pada Tabel 6 terdapat pada Dokumen 1 dan 3. Selanjutnya probabilitas pada kata yang mengandung sentimen negatif seperti pada Tabel 7 yang terdapat pada Dokumen 2, 4, dan 5.

Tabel 6. Dokumen kelas sentimen positif

Dok	selamat	Kerja	Ahok	Nista	Lindung	Tugas	Tidak	Pantas	Kasih	Jabatan	Jadi	Pejabat	Kelas sentimen
1	1	1	1	0	0	0	0	0	0	0	0	0	+
3	1	0	0	0	0	1	0	0	0	0	0	0	+
sum	2	1	1	0	0	1	0	0	0	0	0	0	5

Tabel 7. Dokumen kelas sentimen negatif

Dokumen	selamat	kerja	ahok	nista	lindung	tugas	tidak	pantas	kasih	jabatan	jadi	pejabat	Kelas sentimen
2	0	0	1	1	1	0	0	0	0	0	0	0	-
4	0	0	1	0	0	0	1	1	1	1	0	0	-
5	0	0	0	1	0	0	0	0	0	0	1	1	-
sum	0	0	2	2	1	0	1	1	1	1	1	1	11

Nilai probabilitas kelas sentimen positif dan probabilitas kelas sentimen negative dihitung dengan persamaan (10).

$$P(v_+) = \frac{|2|}{|5|} = 0,4$$

$$P(v_-) = \frac{|3|}{|5|} = 0,6$$

Nilai probabilitas setiap kata pada kelas sentimen positif maupun negatif dihitung menggunakan persamaan (11) didapatkan seperti pada Tabel 8 dan Tabel 9.

Tabel 8. Probabilitas kata pada kelas sentimen positif

Kata	Probabilitas	Probabilitas
selamat	$P(a_{selamat} v_+) = \frac{2 + 1}{5 + 12 } = \frac{3}{17} = 0,1764$	$P(a_{selamat} v_-) = \frac{0 + 1}{11 + 12 } = \frac{1}{23} = 0,0434$
kerja	$P(a_{kerja} v_+) = \frac{1 + 1}{5 + 12 } = \frac{2}{17} = 0,1176$	$P(a_{kerja} v_-) = \frac{0 + 1}{11 + 12 } = \frac{1}{23} = 0,0434$
ahok	$P(a_{ahok} v_+) = \frac{1 + 1}{5 + 12 } = \frac{2}{17} = 0,1176$	$P(a_{ahok} v_-) = \frac{0 + 1}{11 + 12 } = \frac{1}{23} = 0,0434$
tugas	$P(a_{tugas} v_+) = \frac{1 + 1}{5 + 12 } = \frac{2}{17} = 0,1176$	$P(a_{tugas} v_-) = \frac{0 + 1}{11 + 12 } = \frac{1}{23} = 0,0434$

Tabel 9. Probabilitas kata pada kelas sentimen negatif

Kata	Probabilitas	Probabilitas
ahok	$P(a_{ahok} v_-) = \frac{2+1}{11+ 12 } = \frac{3}{23} = 0,1304$	$P(a_{ahok} v_+) = \frac{0+1}{5+ 12 } = \frac{1}{17} = 0,058$
nista	$P(a_{nista} v_-) = \frac{2+1}{11+ 12 } = \frac{3}{23} = 0,1304$	$P(a_{nista} v_+) = \frac{0+1}{5+ 12 } = \frac{1}{17} = 0,058$
lindung	$P(a_{lindung} v_-) = \frac{1+1}{11+ 12 } = \frac{2}{23} = 0,0869$	$P(a_{lindung} v_+) = \frac{0+1}{5+ 12 } = \frac{1}{17} = 0,058$
tidak	$P(a_{tidak} v_-) = \frac{1+1}{11+ 12 } = \frac{2}{23} = 0,0869$	$P(a_{tidak} v_+) = \frac{0+1}{5+ 12 } = \frac{1}{17} = 0,058$
pantas	$P(a_{pantas} v_-) = \frac{1+1}{11+ 12 } = \frac{2}{23} = 0,0869$	$P(a_{pantas} v_+) = \frac{0+1}{5+ 12 } = \frac{1}{17} = 0,058$
kasih	$P(a_{kasih} v_-) = \frac{1+1}{11+ 12 } = \frac{2}{23} = 0,0869$	$P(a_{kasih} v_+) = \frac{0+1}{5+ 12 } = \frac{1}{17} = 0,058$
jabatan	$P(a_{jabatan} v_-) = \frac{1+1}{11+ 12 } = \frac{2}{23} = 0,0869$	$P(a_{jabatan} v_+) = \frac{0+1}{5+ 12 } = \frac{1}{17} = 0,058$
jadi	$P(a_{jadi} v_-) = \frac{1+1}{11+ 12 } = \frac{2}{23} = 0,0869$	$P(a_{jadi} v_+) = \frac{0+1}{5+ 12 } = \frac{1}{17} = 0,058$
pejabat	$P(a_{pejabat} v_-) = \frac{1+1}{11+ 12 } = \frac{2}{23} = 0,0869$	$P(a_{pejabat} v_+) = \frac{0+1}{5+ 12 } = \frac{1}{17} = 0,058$

3.3. Implementasi Data Uji dan Evaluasi Performa

Tabel 10 adalah data yang akan dilakukan pengujian untuk menentukan kelas sentimen setiap dokumen.

Tabel 10. Data uji

Dokumen	Cuitan	Kelas Sentimen
6	selamat, tugas, ahok, komut, pertamina	?
7	mantap	?
8	fadli, zon, sebut, ahok, tidak, pantas, jadi,	?

Dokumen	Cuitan	Kelas Sentimen
	komisaris, utama pertamina	

Data uji dihitung probabilitas dan dicari probabilitas tertinggi menggunakan persamaan (9).

$$\begin{aligned}
 P(Dok6|v_+) &= P(a_{selamat} | v_+) \times P(a_{tugas} | v_+) \times \\
 &P(a_{ahok} | v_+) \times P(a_{komut} | v_+) \times \\
 &P(a_{pertamina} | v_+) \times P(v_+) \\
 &= \frac{3}{17} \times \frac{2}{17} \times \frac{2}{17} \times 1 \times 1 \times \frac{2}{5} \\
 &= 0,00097
 \end{aligned}$$

$$\begin{aligned}
 P(Dok6|v_-) &= P(a_{selamat} | v_-) \times P(a_{tugas} | v_-) \times \\
 &P(a_{ahok} | v_-) \times P(a_{komut} | v_-) \times \\
 &P(a_{pertamina} | v_-) \times P(v_-) \\
 &= \frac{1}{23} \times 1 \times \frac{3}{23} \times 1 \times 1 \times \frac{3}{5} \\
 &= 0,0034
 \end{aligned}$$

$$\begin{aligned}
 P(Dok7|v_+) &= P(a_{mantap} | v_-) \times P(v_+) \\
 &= 1 \times \frac{2}{5} = 0,4
 \end{aligned}$$

$$\begin{aligned}
 P(Dok7|v_-) &= P(a_{mantap} | v_-) \times P(v_-) \\
 &= 1 \times \frac{3}{5} = 0,6
 \end{aligned}$$

$$\begin{aligned}
 P(Dok8|v_+) &= P(a_{fadli} | v_+) \times P(a_{zon} | v_+) \times P(a_{sebut} | v_+) \times P(a_{ahok} | v_+) \times P(a_{tidak} | v_+) \times \\
 &P(a_{pantas} | v_+) \times P(a_{jadi} | v_+) \times P(a_{komisaris} | v_+) \times P(a_{utama} | v_+) \times P(v_+) \\
 &= 1 \times 1 \times 1 \times \frac{2}{17} \times \frac{1}{17} \times \frac{1}{17} \times \frac{1}{17} \times 1 \times 1 \times \frac{2}{5} = 0,000009578
 \end{aligned}$$

$$\begin{aligned}
 P(Dok8|v_-) &= P(a_{fadli} | v_-) \times P(a_{zon} | v_-) \times \\
 &P(a_{sebut} | v_-) \times P(a_{ahok} | v_-) \times \\
 &P(a_{tidak} | v_-) \times P(a_{pantas} | v_-) \times \\
 &P(a_{jadi} | v_-) \times P(a_{komisaris} | v_-) \times \\
 &P(a_{utama} | v_-) \times P(v_-) \\
 &= 1 \times 1 \times 1 \times \frac{3}{23} \times \frac{2}{23} \times \frac{2}{23} \times \frac{2}{23} \times 1 \times 1 \times \frac{3}{5} \\
 &= 0,00005145
 \end{aligned}$$

Nilai probabilitas dari setiap data uji yang diperoleh dirangkum pada Tabel 11. Dokumen 6 menghasilkan sentimen positif, sedangkan Dokumen 7 dan Dokumen 8 menghasilkan sentimen negatif.

Tabel 11. Hasil dari perhitungan probabilitas

Dokumen	Probabilitas		Kelas sentimen
	Positif	Negatif	
6	0,00097	0,0034	+
7	0,4	0,6	-
8	0,000009578	0,00005145	-

Setelah melakukan pengujian data maka selanjutnya melakukan pengujian evaluasi performa menggunakan *confusion matrix* untuk mengukur performa berupa akurasi, presisi, dan *recall*. Hasil klasifikasi menggunakan metode NBC ditunjukkan pada Tabel 12 sehingga menghasilkan perhitungan yang telah diklasifikasi dan diberi kode untuk kelas Positif (0), kelas Negatif (1), dan kelas Netral (2) seperti ditunjukkan pada Tabel 12 dan Tabel 13.

Tabel 12. Hasil klasifikasi data uji

Dokumen	Kelas Sebenarnya	Kelas Prediksi
6	(+)	(+)
7	(+)	(-)
8	(-)	(-)

Tabel 13. Perhitungan *confusion matrix*

Kelas Sebenarnya	Kelas Prediksi			Jumlah
	0	1	2	
0	1 TPP	1 PFNeg	0 PFNet	2 J4
1	0 NegFP	1 TNegNeg	0 NegFNet	1 J5
2	0 NetFP	0 NetFNeg	0 TNetNet	0 J6
Jumlah	1 J1	2 J2	0 J3	Total = 3

Keterangan:

- TPP (True Positive Positive) = jumlah dokumen dari kelas 0 yang benar diklasifikasikan sebagai kelas 0
- TNegNeg (True Negative Negative) = jumlah dokumen dari kelas 1 yang benar diklasifikasikan sebagai kelas 1
- TNetNet (True Netral Netral) = jumlah dokumen dari kelas 2 yang benar diklasifikasikan sebagai kelas 2
- PFNeg (Positive False Negatif) = jumlah dokumen dari kelas 0 yang salah diklasifikasikan sebagai kelas 1
- NegFP (Negatif False) = jumlah dokumen dari kelas 1 yang

<i>Positive)</i>		=	salah diklasifikasikan sebagai kelas 0
PFNet (<i>Positive False</i> <i>Netral)</i>		=	jumlah dokumen dari kelas 0 yang salah diklasifikasikan sebagai kelas 2
NetFP (<i>Netral False</i> <i>Positive)</i>		=	jumlah dokumen dari kelas 2 yang salah diklasifikasikan sebagai kelas 0
NetFNeg (<i>Netral False</i> <i>Negatif)</i>		=	jumlah dokumen dari kelas 2 yang salah diklasifikasikan sebagai kelas 1
NegFNet (<i>Negatif False</i> <i>Netral)</i>		=	jumlah dokumen dari kelas 1 yang salah diklasifikasikan sebagai kelas 2

Pada Tabel 12 dan 13 terdapat kelas prediksi dan kelas sebenarnya. Kelas prediksi adalah kelas hasil prediksi dari proses klasifikasi menggunakan NBC, sedangkan kelas sebenarnya adalah kelas yang sebelumnya sudah ditentukan sentimennya. Setelah itu diperoleh nilai TPP = 1, TNegNeg = 1, TNetNet = 0 dan PFNeg = 1, NegFP = 0, dan NetFP = 0. *Accuracy*, *precision*, dan *recall* masing-masing dihitung menggunakan persamaan (12), (13), dan (14).

$$\begin{aligned}
 accuracy &= \frac{TPP + TNegNeg + TNetNet}{Total} \\
 &= \frac{1+1+0}{2+1+0} = \frac{2}{3} = 60\% \\
 precision &= \frac{TPP}{TPP + NegFP + NetFP} \\
 &= \frac{1}{1 + 0 + 0} = \frac{1}{1} = 100\% \\
 recall &= \frac{TPP}{TPP + PFNeg + PFNet} \\
 &= \frac{1}{1+1+0} = \frac{1}{2} = 50\%
 \end{aligned}$$

Jadi pada pengujian akurasi yang diperoleh dengan menggunakan 8 data yang terdiri dari 5 data latih dan 3 data uji menghasilkan akurasi 60%, presisi 100%, dan *recall* 50%.

4. SIMPULAN

Berdasarkan hasil dari penelitian yang telah dilakukan maka sistem analisis sentimen pada Twitter menggunakan metode NBC telah berhasil dibangun dan mampu memberikan informasi yang sangat bermakna untuk mengetahui pendapat atau opini masyarakat. Performa yang dievaluasi menggunakan metode *confusion matrix* memiliki akurasi 60%, presisi 100%, dan *recall* 50%. Hasil tersebut didapat dari hasil analisis 5 data latih dan 3

data uji sebagai studi kasus untuk kemudahan dalam pemahaman implementasi dan perhitungan secara terstruktur. Metode NBC pada penelitian ini menghasilkan klasifikasi yang baik dengan nilai akurasi yang tinggi sehingga cocok untuk memprediksi tingkat sentimen analisis. Selanjutnya akan dikembangkan dan diimplementasikan dengan data yang lebih besar menggunakan sistem aplikasi berbasis komputer sehingga lebih mudah dan cepat.

DAFTAR PUSTAKA

- [1] Muhammad Sufyan Abdurrahman (2020) Penetrasi Internet Indonesia pada 2020, *alinea.id*. Tersedia pada: <https://www.alinea.id/kolom/tantangan-penetrasi-internet-indonesia-pada-2020-b1ZJC9smS> (Diakses: 7 Agustus 2020).
- [2] Priska Sari Pratiwi (2017) Ahok Divonis Dua Tahun Penjara, *CNN Indonesia*. Tersedia pada: <https://www.cnnindonesia.com/nasional/20170509080949-12-213328/ahok-divonis-dua-tahun-penjara> (Diakses: 13 Maret 2020).
- [3] Dewan Komisaris | PT Pertamina (Persero) (2019). Tersedia pada: <https://www.pertamina.com/id/dewan-komisaris> (Diakses: 25 November 2019).
- [4] Rahmat Burhanudin (2018) Mengenal Sentiment Analysis, <https://mamat.co/>. Tersedia pada: <https://mamat.co/mengenal-sentiment-analysis/> (Diakses: 24 September 2019).
- [5] Buntoro, G. A. (2017) "Analisis Sentimen Calon Gubernur DKI Jakarta 2017 Di Twitter," *Integer Journal*, 2(1), hal. 32-41.
- [6] Nurhuda, F., Widya Sihwi, S. dan Doewes, A. (2016) "Analisis Sentimen Masyarakat terhadap Calon Presiden Indonesia 2014 berdasarkan Opini dari Twitter Menggunakan Metode *Naive Bayes Classifier*," *Jurnal Teknologi & Informasi ITSmart*, 2(2), hal. 35.
- [7] Nurirwan Saputra, Teguh Bharata Adji, A. E. P. (2015) "Analisis Sentimen Data Presiden Jokowi dengan *Pre-Processing* Normalisasi dan *Stemming* Menggunakan Metode *Naive Bayes* dan SVM," *Jurnal Dinamika Informatika*, 5(November), hal. 12.
- [8] Utami, L. A. (2017) "Melalui Komparasi Algoritma *Support Vector Machine* dan *K-Nearest Neighbor* Berbasis *Particle Swarm Optimization*," *Jurnal Pilar Nusa Mandiri*, 13(1), hal. 103-112.
- [9] Oktasari, L., Chrisnanto, Y. H. dan Yuniarti, R. (2016) "*Text Mining* dalam Analisis Sentimen Asuransi Menggunakan Metode *Naive Bayes Classifier*," *Prosiding SNST*, 7, hal. 37-42.
- [10] Eka Sembodo, J., Budi Setiawan, E. dan Abdurrahman Baizal, Z. (2016) "*Data Crawling* Otomatis pada Twitter," (September), hal. 11-16. doi: 10.21108/indosc.2016.111.
- [11] Yusnitasari, T. et al. (2017) "Analisis Sentimen Terhadap *Review* Restoran Fish Streat pada Aplikasi Zomato Menggunakan *Stemming* Nazief Adriani dan *Naive Bayes Classifier*," *Prosiding Setrinov*, 3, hal.

163-174.

- [12] Saputro, P. H., Aristin, M. dan Tyas, Dy. L. (2017) "Klasifikasi Lagu Daerah Indonesia Berdasarkan Lirik Menggunakan Metode TF-IDF dan *Naive Bayes Classifier*," Jurnal Teknologi Informatika dan Terapan, 4(1), hal. 45-50.
- [13] Putra, D. dan Wibowo, A. (2020) "Prediksi Keputusan Minat Penjurusan Siswa SMA Yadika 5 Menggunakan Algoritma *Naive Bayes*," 2, hal. 84-92.