Implementasi Metode K-Means dan Naïve Bayes *Classifier* untuk Analisis Sentimen Pemilihan Presiden (Pilpres) 2019

Imam Kurniawan[1], Ajib Susanto[2]
Jurusan Teknik Informatika, Fakultas Ilmu Komputer
Universitas Dian Nuswantoro Semarang
Semarang, Indonesia
e-mail: [1]kurniawanz.seven32@gmail.com, [2]ajib.susanto@dsn.dinus.ac.id

Diajukan: 7 Maret 2019; Direvisi: 12 Juni 2019; Diterima: 4 Juli 2019

Abstrak

Pemilihan umum presiden yang diselenggarakan setiap lima tahun sekali merupakan momen yang penting untuk mewujudkan demokrasi dalam Negara Kesatuan Republik Indonesia. Penyampaian dukungan dilakukan baik tim sukses, buser maupun pendukung untuk mencitrakan positif calon masing-masing. Berbagai media digunakan salah satunya adalah Twitter, masyarakat menyampaikan komentar positif dan negatif bahkan cenderung "kampanye hitam" dan hoax sebelum pemilu dilaksanakan maupun saat pemilu sedang berlangsung mengenai pemilu yang diadakan, komentar di Twitter saat ini belum dapat ditentukan lebih ke arah positif atau negatif, oleh karena itu perlu dilakukan analisis sentimen untuk mengetahui kecenderungan opini masyarakat terhadap pemilu. Tujuan dari penelitian ini memperoleh analisis dokumen text untuk mendapatkan sentimen positif atau negatif. Metode yang digunakan K-Means untuk melakukan klastering pada data latih dan Naive Bayes classifier untuk mengklasifikasi pada data testing. Hasil dari pembobotan ini berupa sentimen positif dan negatif. Data diambil dari Twitter mengenai pemilu presiden 2019 sebanyak 500 data tweet. Dari hasil pengujian 100 dan 150 data uji diperoleh akurasi rata-rata 93.35% dan error rate sebesar 6.66%.

Kata kunci: *Pemilu 2019, K-Means, Naïve Bayes, Analisis sentimen, Twitter.*

Abstract

The presidential election held every five years is an important moment to realize democracy in the Unitary State of the Republic of Indonesia. Submission of support is done both successful team, buzzers and supporters to positively portray each candidate. One of the various media used is Twitter, the public conveys positive and negative comments and even tends to "black campaigns" and hoaxes before elections are held and when elections are taking place on elections held, comments on Twitter today cannot be determined more positively or negatively. Therefore, it is necessary to conduct sentiment analysis to determine the tendency of public opinion on the election. The purpose of this study is to obtain an analysis of text documents to get positive or negative sentiments. The method used by K-means is to cluster the training data and Naive Bayes classifier to classify the testing data. The results of this weighting are positive and negative sentiments. Data was taken from Twitter regarding the 2019 presidential election as many as 500 data tweets. From the results of testing 100 and 150 test data obtained an average accuracy of 93.35% and an error rate of 6.66%.

Keywords*: 2019 president election, K-Means, Naïve Bayes, Sentiment Analysis, Twitter.*

1. Pendahuluan

Pemilihan umum di dalam sejarah nasional Indonesia telah dilaksanakan selama beberapa kali, namun pemilihan umum yang dilakukan langsung oleh masyarakat Indonesia baru pertama kali dimulai pada era reformasi setelah era orde baru runtuh yaitu tahun 2004. Pemilihan umum presiden yang akan diselenggarakan pada tahun 2019 merupakan momen yang penting untuk mewujudkan demokrasi dalam Negara Kesatuan Republik Indonesia. Kandidat dan tim sukses pada pemilihan presiden tahun 2019 ini dapat memanfaatkan media sosial untuk menyampaikan pesan kampanye [1], salah satu media yang aktif digunakan untuk kampanye adalah Twitter. Pada pemilihan presiden 2019 yang akan datang sebanyak 40% atau sekitar 90 juta orang merupakan pemilih pemula atau generasi *millennial*, sehingga kekuatan sosial

media tidak dapat dianggap remeh untuk elektabilitas kedua pasang calon [2] hal ini disampaikan oleh Direktur Eksekutif *Voxpol Center Research and Consulting* Pangi Syarwi Chaniago.

Jejaring sosial seperti Twitter sekarang menjadi perangkat komunikasi yang sangat populer di kalangan pengguna dunia maya dan dapat digunakan sebagai media kampanye peserta pemilihan presiden dalam menyampaikan citra positif bagi pasangan masing-masing pada calon pemilih dan pendukungnya. Para kandidat yang berlaga dalam berbagai pemilihan umum di seluruh Asia pada tahun 2019 terlihat memanfaatkan Twitter dan kanal media sosial lainnya untuk membagikan slogan dan kebijakan, mengguncang popularitas saingan, dan membangun massa menjelang kampanye. Hal itu terutama mencolok di Indonesia sebagai salah satu dari lima besar pengguna media sosial di dunia, menjelang Pilpres 2019 bulan April mendatang [3]. Menurut laporan yang dirilis oleh *Wearesocial* pada awal tahun 2018 dijelaskan bahwa pengguna internet di Indonesia mencapai 132 juta orang dengan 60% persennya mengakses internet menggunakan ponsel pintar menjadikan Indonesia sebagai peringkat keempat terbesar negara pengakses internet, untuk penggunaan *social media* sendiri Indonesia memperoleh peringkat ke tiga dengan 53 juta pengguna [4]. Menurut data yang dirilis oleh Twitter Indonesia pada akhir tahun 2016 lalu, dikatakan bahwa 77% pengguna Twitter di Indonesia adalah pengguna aktif. Hal ini dapat dilihat dari banyaknya jumlah *tweet* yang dihasilkan selama tahun 2016 yang mencapai 4.1 miliar *tweet* [5].

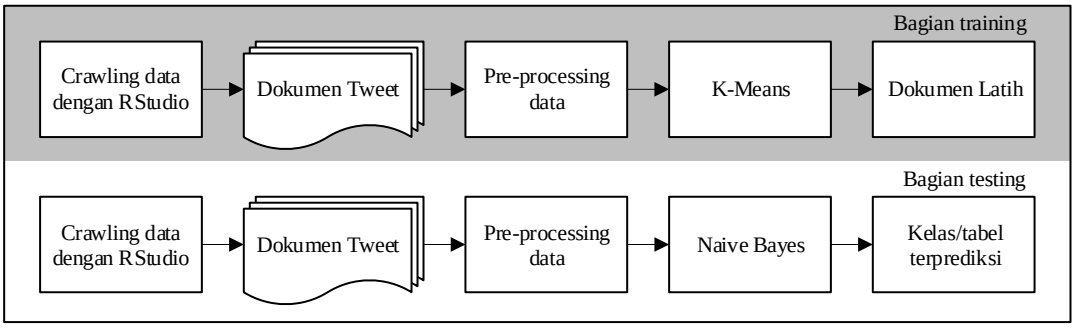
Terdapat banyak komentar positif dan negatif masyarakat saat sebelum pemilu dilaksanakan maupun saat pemilu sedang berlangsung mengenai pemilu yang diadakan. Twitter menyediakan sumber-sumber opini yang banyak jumlahnya, sebagai contoh untuk tagar #2019GantiPresiden mempunyai jumlah 1.189.714 *tweet* dan #2019TetapJokowi dengan 345.849 *tweet*, hasil riset dan *monitoring* Binokular kurun waktu 1 Juli-31 Desember 2018 [6]. Opini di Twitter belum dapat diidentifikasi secara langsung merupakan opini positif atau opini negatif. Informasi yang diterima langsung dari media *Twitter* jika dipahami apa adanya tanpa melakukan pengecekan terlebih dahulu sumber asli dan yang menyebarkan terpercaya atau tidak cenderung akan menjadi berita palsu atau *hoax* bahkan menjurus "kampanye hitam" kepada calon lawan di pemilihan presiden. Di Twitter ditemukan opini-opini, aspirasi atau komentar dari masyarakat yang dapat digunakan untuk mengekspresikan peristiwa yang sedang terjadi dalam hal ini adalah yang berhubungan dengan pemilihan presiden dengan *hashtag*(#) #2019GantiPresiden dan #2019TetapJokowi. Agar opini-opini tersebut dapat dimanfaatkan dan berguna, dibutuhkan berbagai proses sehingga didapatkan suatu informasi yang penting melalui analisis sentimen.

Analisis sentimen disebut juga dengan *opinion mining* (penambangan opini) yaitu proses untuk mengekstrak suatu opini atau pendapat dari dokumen untuk topik tertentu [7]. Analisis sentimen dilakukan untuk mengetahui kecenderungan opini seseorang terhadap sebuah peristiwa atau masalah, apakah cenderung beropini positif atau negatif. Teknik yang digunakan yaitu *text mining*. *Text mining* merupakan sebuah teknik yang digunakan untuk mengekstraksi informasi yang berguna dari data teks yang tidak terstruktur. *Text mining* mengekstraksi kata kunci atau mengekstraksi pendapat dan ulasan analisis *text* sehingga dapat mendukung untuk memahami pendapat masyarakat dalam data text [8].

Dalam proses analisis sentimen terdapat beberapa metode *Naïve Bayes, Maximum Entropy, SVM* [9], *Neural Network* [10], penelitian sebelumnya data *Twitter* digunakan untuk analisis sentimen opini Pilkada DKI 2017 dengan menggunakan metode Naïve Bayes menghasilkan nilai akurasi tertinggi sebesar 74.81% [11]. Penelitian berjudul *"Indonesian News Classification based on NaBaNA"* penelitian ini mengategorikan berita berbahasa Indonesia pada aplikasi berita Jawa Tengah menggunakan *naive bayes* dan *stemming* dari Naizief-Adriani. Penelitian ini menghasilkan nilai akurasi sebanyak 94% [12]. Selain akurasinya yang cukup tinggi algoritma *stemming* dan kualitas data yang digunakan juga berpengaruh terhadap nilai akurasi dari hasil pengujian. Algoritma *K-Means* juga digunakan untuk analisis sentimen r*eview* film dengan menguji hasil akurasi dari metode *K-Means* dengan menggunakan data *review* film berbahasa Inggris [13]. Pada penelitian ini diusulkan penggunaan dua metode yaitu metode *K-Means* dan *Naïve Bayes Classifier* untuk menghasilkan analisis sentimen opini data *Twitter* yang cenderung beropini positif atau negatif.

2. Metode Penelitian

Penelitian ini dilakukan dengan menggunakan beberapa tahap sesuai yang dijelaskan pada gambar 1, seperti: pengumpulan data, *pre-processing* data, pembangunan sistem menggunakan metode *K-Means* dan *Naive Bayes*, serta evaluasi sistem.



Gambar 1. Desain sistem klastering dan klasifikasi data opini Pilpres 2019.

2.1. Pengumpulan Data

Proses pengumpulan data yang telah dilakukan berasal dari *tweet* yang dikirimkan oleh masyarakat di *twitter*. Data yang diambil berupa *tweet* yang menggunakan *hashtag*(#) #2019GantiPresiden dan #2019TetapJokowi. Data diperoleh dengan memanfaatkan *tools Rstudio* dan *twitter API*. Langkah pertama yang dilakukan adalah dengan memperoleh akses ke *API twitter* dengan cara login ke akun twitter yang dimiliki ke http://developer.twitter.com. Setelah memperoleh *API key*, *API Secret*, *Access token* serta *Access token secret*, *API Twitter* baru dapat diakses.

Dari hasil data yang dikumpulkan sebanyak 500 data, kemudian dilakukan klastering untuk membagi data menjadi dua kelas yaitu positif dan negatif. Pengelompokan menjadi dua kelas tersebut dilakukan secara otomatis oleh *k-means* dengan menggunakan *database* kata positif dan kata negatif. Hasil pengelompokan tersebut adalah 175 dokumen positif dan 375 dokumen negatif.

Kalimat dalam kelompok data yang beropini positif mengandung kata-kata seperti : semangat, menang, jujur, baik, sabar, bahagia, tenang, hebat, cinta, bangga dan sebagainya. Sedangkan kalimat dalam kelompok data beropini negatif mengandung kata-kata seperti: bohong, waspada, salah, hancur, buta, fitnah, panik, boneka, hutang, ngeri, mati, jahat, bocor, neraka dan sebagainya.

2.2. Pre-Processing Data

Data *Tweet* yang sudah diambil masih berupa data mentah yang belum siap diolah, oleh karena itu dilakukan tahap *pre-processing* untuk mendapatkan data yang siap untuk diolah pada proses selanjutnya. *Text pre-processing* digunakan untuk segmentasi *text*, dan hanya melalui segmentasi *text* karakteristik dapat dinilai, dianalisis, dan di klasifikasi [14]. Tahapan *pre-processing* yang dilakukan adalah:

1. Normalisasi: Adalah proses yang dilakukan untuk membersihkan fitur-fitur yang tidak diperlukan dalam pengambilan data yang ada pada Twitter, seperti URL, Username, dan lain-lain.
2. *Case Folding*: dalam sebuah *tweet* sering kali memiliki banyak perbedaan penggunaan pada bentuk huruf, pada bagian ini dilakukan perubahan seluruh huruf kapital (*uppercase*) dikembalikan menjadi huruf kecil (*lowercase*) agar seragam.
3. *Tokenizing*: Tokenisasi adalah proses yang dilakukan untuk memenggal kalimat menjadi beberapa bagian atau kata berdasarkan tanda bacanya seperti koma, titik, dan tanda baca lainnya.
4. *Stopword Removal:* Merupakan proses yang dilakukan dengan cara menghilangkan kata yang tidak diperlukan. Jika kata tersebut dibuang maka tidak akan mengubah atau menghilangkan informasi yang berada dalam kalimat tersebut. Seperti kata hubung yang, akan, di , pada dan lain-lain.
5. *Stemming:* Adalah proses yang memiliki tujuan untuk menghilangkan imbuhan-imbuhan yang terdapat pada sebuah kata. Proses ini mengubah kembali semua kata menjadi kata dasar. Sebagi contoh kata “pembohong” menjadi kata “bohong”.

2.3. Klastering Data Latih Menggunakan Algoritma *K-Means*

Klastering merupakan metode yang digunakan membagi sekumpulan data ke dalam sejumlah kelompok tertentu. Pengelompokan data adalah langkah penting yang berkembang dalam berbagai masalah pengenalan pola dan aplikasi pengambilan keputusan [15]. Algoritma pengelompokan berdasarkan *centroid* yang paling populer termasuk K-Means [15], *hierarchical clustering*, *spectral clustering* dan *Gaussian mixture* model.

Algoritma *K-Means* terdiri dari dua fase yang terpisah. Fase pertama menghitung k *centroid* dan fase kedua mengambil setiap titik ke kluster yang memiliki titik pusat terdekat dari titik data masing-masing [16]. Terdapat beberapa metode untuk menghitung *centroid* terdekat salah satunya adalah *Euclidian Distance*. Tahapan pengelompokan data menggunakan metode *K-Means* adalah [17]:

1. Menentukan jumlah *cluster*
2. Mengelompokkan data sehingga terbentuk K buah *cluster* dengan titik *centroid* dari setiap cluster merupakan titik *centroid* yang telah dipilih sebelumnya.
3. Hitung pusat *centroid* dari data yang ada di masing-masing kelompok. Tempat *centroid* setiap kelompok diambil dari rata-rata (*mean*) semua nilai data pada setiap fiturnya.
4. Alokasikan masing-masing data pada *centroid* terdekat. Untuk mengukur jarak data ke pusat *centroid* dapat dilakukan dengan rumus *Eucilidean distance* sebagai berikut:

$$d = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2} \quad (1)$$

Data dialokasikan kembali secara tegas ke dalam kelompok yang memiliki *centroid* dengan jarak yang paling dekat dari data tersebut. Proses pengalokasian kembali data ini menggunakan persamaan:

$$a_{ij} = \begin{cases} 1\ d = \min\{D(x_i, C_i)\} \\ \\ 0\ lainnya \end{cases} \quad (2)$$

a_{ij} adalah nilai keanggotaan titik X_i ke *centroid* C_1, d adalah jarak terpendek dari data Xi ke k kelompok setelah dibandingkan, dan C1 adalah *centroid* ke-1.

5. Ulangi langkah 2-4 hingga nilai dari titik *centroid* tidak lagi mengalami perubahan.

2.4. Klasifikasi Data Uji Menggunakan Naive Bayes

Klasifikasi merupakan proses untuk menemukan model dari sebuah data. Tujuan proses klasifikasi adalah untuk mengambil suatu keputusan dengan melakukan prediksi suatu kasus berdasarkan hasil dari klasifikasi yang telah diperoleh. Naive Bayes merupakan metode pengklasifikasian yang sering digunakan dalam sentimen analisis karena sederhana dan mudah dalam melakukan pengklasifikasian dokumen.

Teori Bayes secara umum dapat dinotasikan dengan persamaan:

$$P(A|B) = \frac{P(B|A) \times P(A)}{P(B)} \quad (3)$$

Di mana :

A : Hipotesis data merupakan suatu *class* spesifik.
B : Data dengan kelas yang masih belum diketahui.
P(A|B) : Probabilitas hipotesis berdasar kondisi.
P(A) : Probabilitas hipotesis.
P(B|A) : Probabilitas berdasarkan kondisi pada hipotesis.
P(B) : Probabilitas B.

2.5. Evaluasi Sistem

Evaluasi sistem dilakukan dengan cara menghitung tingkat keakuratan suatu metode untuk menganalisis hasil opini dari Twitter. *Confussion matrix* dipilih sebagai metode pengujian untuk menguji keakuratan metode *Naive Bayes* sehingga dapat diketahui hasil akurasinya. Dalam perhitungan *confusion matrix* terdiri dari *True Positive* (TP) yaitu jumlah data kelas positif yang diklasifikasikan sebagai kelas positif dan *True Negative* (TN) merupakan jumlah data kelas negatif yang diklasifikasikan sebagai kelas negatif. Sedangkan *False Positive* (FP) merupakan jumlah data kelas negatif yang diklasifikasikan sebagai kelas positif, dan *False Negati*ve (FN) adalah jumlah data kelas positif yang diklasifikasikan sebagai kelas negatif.

Dari *confussion matrix* diperoleh nilai *accuracy* dan *error rate*. Rumus masing-masing ada di bawah ini:

$$Akurasi = \frac{TP+TN}{TP+FP+TN+FN} \times 100\% \quad (4)$$

$$error\ rate = \frac{FP+FN}{TP+FP+TN+FN} \times 100\% \quad (5)$$

3. Hasil dan Pembahasan

Data *tweet* hasil *crawling* yang digunakan dalam penelitian ini adalah *tweet* yang berhubungan dengan masa kampanye Pilpres 2019 dan menggunakan tagar #2019GantiPresiden dan #2019TetapJokowi. Data tersebut kemudian akan dibagi menjadi 2 bagian, yaitu data *training* dan data *testing*. Data *training* merupakan data yang telah dilakukan *preprocessing* data dan telah diberi label menggunakan metode *K-Means*. Untuk masing-masing *hashtag* akan diambil sebanyak 250 data yang akan digunakan sebagai data latih.

3.1. Preprocessing Data

a. *Case Folding*

Proses ini merupakan proses yang mengubah huruf besar (*uppercase*) menjadi huruf kecil (*lowercase*). Proses ini dilakukan untuk kemudian mempermudah dalam melakukan proses selanjutnya.

Input : Yakin kalo #2019GantiPresiden akan membawa Indonesia ini menjadi lebih baik. Yang sudah sudah juga begitu ..

Output : yakin kalo #2019gantipresiden akan membawa indonesia ini menjadi lebih baik. yang sudah sudah juga begitu ..

Gambar 2. *Case Folding*

b. Normalisasi

Dalam data *tweet*, terdapat beberapa fitur yang tidak memiliki pengaruh pada proses selanjutnya, maka komponen-komponen tersebut di hilangkan, beberapa komponen yang perlu dihilangkan seperti, *username*, dan URL. Pada komponen *username* biasanya akan diawali dengan karakter "@". Sedangkan untuk URL ditandai dengan munculnya format *URL* seperti *http*, *https*, atau *www*.

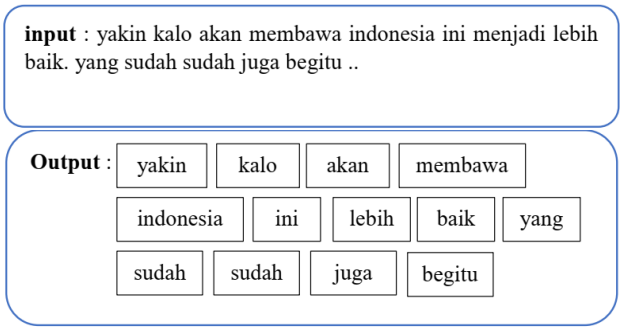
Input : yakin kalo #2019gantipresiden akan membawa indonesia ini menjadi lebih baik. yang sudah sudah juga begitu ..

Output : yakin kalo akan membawa indonesia ini menjadi lebih baik. yang sudah sudah juga begitu ..

Gambar 3. Normalisasi.

c. *Tokenizing*

Pada tahapan ini kalimat pada *tweet* akan diperiksa secara menyeluruh. Kemudian dilakukan pemenggalan kata berdasarkan karakter pemisahnya, sehingga kata yang bukan karakter pemisah akan digabungkan dengan karakter selanjutnya.



Gambar 4. *Tokenizing*.

d. *Stopword Removal*

Setelah dilakukan proses *case folding*, normalisasi dan *tokenizing*, proses selanjutnya adalah dengan melakukan *stopword removal*. Pada proses ini setiap kata akan diperiksa, jika di dalam kata tersebut terdapat kata hubung, kata depan, atau kata ganti, maka kata tersebut akan dihilangkan. Kata depan, hubung, atau ganti tersebut terlebih dahulu didefinisikan dalam Tabel 1.

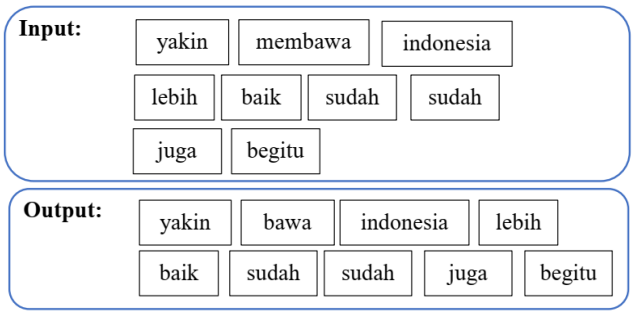
Tabel 1. *Stopword Removal List*.

No	*Stopword*
1	Dan
2	Dengan
3	Yang
4	Akan
5	Kalo
6	Ini
7	Itu
...	...

Gambar 5. *Stopword Removal*.

e. *Stemming*

Tweet yang sudah melalui proses *stopword removal* akan dilanjutkan ke tahap *stemming*, pada tahap ini *stemming* yang digunakan adalah sastrawi *stemming* dengan algoritma dari Nazief dan Adrian. Proses *stemming* akan menghapus imbuhan baik awalan, akhiran, maupun awalan akhiran yang kemudian akan diubah menjadi kata dasar.



Gambar 6. *Stemming*

3.2. Pembobotan

Setelah melewati tahap *pre-processing*, tahap berikutnya dilakukan pembobotan kata dengan menggunakan *opinion word*, berikut contoh tabel pembobotan untuk 6 (enam) *tweet*:

Tabel 2. Hasil Pembobotan Kata.

***Tweet* ke**	**Hasil *Preprocessing***	**Bobot**	
		Positif	**Negatif**
1.	sudah jelas akan dukung umat	1 dukung	0
2.	tambah yakin pokok	1 yakin	0
3.	saat kader rusuh debat pilkada antipati masyarakat jokowi	0	2 rusuh, antipati
4.	sombong hancur tau sombong tolak benar remeh orang korupsi	1 benar	4 sombong, sombong, remeh,korupsi
5.	cerdas politik	1 cerdas	0
6.	pimpin berani independen berani	3 berani, independen, berani	0

3.3. Klastering

Pemberian label dilakukan secara otomatis, data dibagi menjadi 2 kelas, yaitu kelas positif (C1) dan kelas negative(C2). Proses klastering dilakukan menggunakan *centroid* yang dipilih secara acak pada *data training*. *Data training* yang digunakan sebagai *centroid* adalah *tweet* ke 3= {0,2} dan ke 6 = {3,0} yang kemudian dihitung menggunakan rumus *Euclidian Distance.*

Tabel 3. Inisiasi *Centroid.*

Tweet	**Jarak C1**	**Jarak C2**	**Centroid Terdekat**
1	3,04	0,5	C2
2	3,04	0,5	C2
3	1,12	2,5	C1
4	1,12	4,03	C1
5	3,04	0,5	C2
6	3,9	1,5	C2

Setelah dilakukan iterasi sebanyak 2 (dua) kali diperoleh nilai seperti Tabel 3 di atas dan dikarenakan anggota *centroid* 1 dan *centroid* 2 pada kedua iterasi tidak mengalami perubahan, maka iterasi dihentikan dengan hasil anggota C1 *tweet* 3,4 dan pada anggota C2 *tweet* 1,2,5,6.

3.4. Skenario Pengujian

Pada bagian ini dipaparkan pengujian yang dilakukan dan hasil yang didapat. Setelah data dipersiapkan, kemudian dilakukan *pre-processing* terhadap data latih dan dilakukan pelabelan untuk tiap dokumennya menggunakan K-Means. Data yang digunakan sebagai data latih sebanyak 500 data dengan hasil pelabelan oleh K-Means masing-masing menjadi 175 yang memiliki kelas positif dan 325 data memiliki kelas negatif. Selanjutnya peneliti melakukan proses pengujian data uji menggunakan algoritma *naive bayes*.

a. Skenario Pertama: Pengujian model dengan data set sebanyak 100 data uji

Proses Naive Bayes yang telah dilakukan dengan data pelatihan sebanyak 500 data dan data pengujian sebanyak 100 data, menghasilkan tabel *confussion matrix* seperti terlihat di Tabel 4.

Tabel 4. *Confussion Matrix* dari *Naive Bayes* dengan 100 data uji.

		Aktual	
		Positif	**Negatif**
Prediksi	**Positif**	16	3
	Negatif	3	78

$$Akurasi = \frac{TP+TN}{TP+FP+TN+FN} \times 100\%$$

Berdasarkan 100 data uji penelitian dapat diketahui *matrix* sebagai berikut :
TP (*True Positive*) : 16
TN (*True Negative*) : 78
FP (*False Positive*) : 3
FN (*False Negative*) : 3
Sehingga dapat dihitung menjadi :

$$Akurasi = \frac{16+78}{16+3+78+3} \times 100\%$$

$$Akurasi = \frac{94}{100} \times 100\%$$

$$Akurasi = 94\%$$

Dan *error rate* :

$$error\ rate = \frac{FP+FN}{TP+FP+TN+FN} \times 100\%$$

$$error\ rate = \frac{3+3}{16+3+78+3} \times 100\%$$

$$error\ rate = \frac{6}{100} \times 100\%$$

$$error\ rate = 6\%$$

Tabel 5. Akurasi dan *ErrorRate* dari 100 data uji.

Dokumen	***Accuracy***	***Error Rate***
100	94%	6%

b. Skenario Kedua: Pengujian model dengan data set sebanyak 150 data uji

Proses Naive Bayes yang telah dilakukan dengan data pelatihan sebanyak 500 data dan data pengujian sebanyak 150 data, menghasilkan tabel *confussion matrix* seperti terlihat di Tabel 6.

Tabel 6. *Confussion Matrix* dari *Naive Bayes* dengan 150 data uji.

		Aktual	
		Positif	**Negatif**
Prediksi	**Positif**	51	8
	Negatif	3	88

$$Akurasi = \frac{TP+TN}{TP+FP+TN+FN} \times 100\%$$

Berdasarkan 150 data uji penelitian dapat diketahui *matrix* sebagai berikut :
TP (*True Positive*) : 51
TN (*True Negative*) : 88

FP (*False Positive*) : 3
FN (*False Negative*) : 8
Sehingga dapat dihitung menjadi :

$$Akurasi = \frac{51 + 88}{51 + 3 + 88 + 8} \times 100\%$$

$$Akurasi = \frac{139}{150} \times 100\%$$

$$Akurasi = 92.7\%$$

Dan *error rate* :

$$error\ rate = \frac{FP + FN}{TP + FP + TN + FN} \times 100\%$$

$$error\ rate = \frac{3 + 8}{51 + 3 + 88 + 8} \times 100\%$$

$$error\ rate = \frac{11}{150} \times 100\%$$

$$error\ rate = 7.33\%$$

Tabel 7. Akurasi dan *ErrorRate* dari 150 data uji.

Dokumen	*Accuracy*	*Error Rate*
150	92.7%	7.33%

Tabel 8. Rata-rata Akurasi dan *Error Rate.*

Dokumen	*Accuracy*	*Error Rate*
100	94%	6%
150	92.7%	7.33%
Rata-rata	93.35	6.66%

4. Kesimpulan

Penelitian ini mempresentasikan analisis sentimen terhadap data opini pada *Twitter* mengenai pemilihan umum presiden tahun 2019. Metode untuk melakukan analisis sentimen menggunakan *K-Means* untuk melakukan klastering pada data latih dan menghasilkan bobot positif atau negatif pada tiap dokumen latih, kemudian metode Naive Bayes digunakan untuk melakukan klasifikasi pada dokumen uji. Untuk menentukan kinerja mesin pengklasifikasian Naive Bayes dalam proses klasifikasi dilakukan percobaan selama dua kali menggunakan *confussion matrix*. Dari hasil pengujian 100 dan 150 data uji tersebut didapatkan akurasi rata-rata sebesar 93.35% dan *error rate* rata-rata sebesar 6.66%.

Daftar Pustaka

[1] KPU, "Peraturan KPU Nomor 23 Tahun 2018 Tentang Kampanye Pemilihan Umum," KPU, Jakarta, 2018.

[2] M. Ridwan, "Pemilihan Presiden 2019, Jangan Remehkan Kekuatan Media Sosial," Bisnis.Com, 18 10 2018. [Online]. Available: https://kabar24.bisnis.com/read/20181018/15/850844/pemilihan-presiden-2019-jangan-remehkan-kekuatan-media-sosial. [Diakses 29 03 2019].

[3] S. Roughneen, "Dipenuhi Akun Palsu, Bisakah Twitter Goyang Pilpres 2019?," Nikkei Asia, 25 2 2019. [Online]. Available: https://www.matamatapolitik.com/analisis-x-mengapa-media-sosial-miliki-kemungkinan-kecil-pengaruhi-pemilu-di-asia/. [Diakses 29 03 2019].

[4] We Are Social, "Digital in 2018 in Southeast Asia Part 2 - South-East," 29 January 2018. [Online]. Available: https://www.slideshare.net/wearesocial/digital-in-2018-in-southeast-asia-part-2-southeast-86866464. [Diakses 25 04 2018].

[5] Herman, "Indonesia Masuk Lima Besar Pengguna Twitter," 03 05 2017. [Online]. Available: http://www.beritasatu.com/iptek/428591-indonesia-masuk-lima-besar-pengguna-twitter.html. [Diakses 2018 04 15].

[6] R. Nurmansyah dan T. Rahmat, "Suara.Com," Suara.Com, 05 01 2019. [Online]. Available: https://www.suara.com/tekno/2019/01/05/220500/riset-binokular-jokowi-capres-terbanyak-yang-diberitakan-pada-2018. [Diakses 29 03 2019].

[7] H. Kaur, V. Mangat dan N. , "A Survey of Sentiment Analysis Techniques," *2017 International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC)*, pp. 921–925, 2017.

[8] T. Matsumoto, W. Sunayama, Y. Hatanaka dan K. Ogohara, "Data Analysis Support by Combining Data Mining and Text Mining," *2017 6th IIAI International Congress on Advanced Applied Informatics*, pp. 313-318, 2017.

[9] M. Fachrurrozi dan N. Yusliani, "Analisis Sentimen Pengguna Jejaring Sosial Menggunakan Metode Support Vector Machine," *Konferensi Nasional Sistem Informasi*, vol. 1, no. Konferensi Nasional Sistem Informasi, 2015.

[10] Binus University, "MTI, Binus University," Maste Of Information Technology, 04 10 2017. [Online]. Available: https://mti.binus.ac.id/2017/10/04/1900/. [Diakses 30 03 2019].

[11] A. R. T. Lestari, R. S. Perdana dan M. A. Fauzi, "Analisis Sentimen Tentang Opini Pilkada Dki 2017 Pada Dokumen Twitter Berbahasa Indonesia Menggunakkan Naive Bayes dan Pembobotan Emoji," *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, vol. 1, pp. 1718-1724, 2017.

[12] G. Septian, A. Susanto dan G. F. Shidik, "Indonesian News Classification based on NaBaNA," *International Seminar on Application for Technology of Information and Communication (iSemantic)*, 2017.

[13] S. Budi, "Text Mining Untuk Analisis Sentimen Review Film," *Techno.COM*, vol. 16, pp. 1-8, 2017.

[14] Y. Y. Yang dan F. Zhon, "Microblog Sentiment Analysis Algorithm Research and Implementation," *14th International Symposium on Distributed Computing and Applications for Business Engineering and Science*, pp. 288-291, 2015.

[15] A. B. Ayed, M. B. Halima dan A. M. Alimi, "Adaptive fuzzy exponent cluster ensemble system based feature selection and spectral clustering," *2017 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, 2017.

[16] N. Dhanachandra, K. Manglem dan Y. J. Chanu, "Image Segmentation Using K -means Clustering Algorithm and Subtractive Clustering Algorithm," *Procedia Computer Science*, vol. 54, pp. 764-771, 2015.

[17] S. dan D. A. Diartono, "Analisa Jejaring Sosial Twitter Menggunakan Klastering Kmeans dan Hirarki Agglomeratif," *Prosiding SINTAK 2017*, pp. 404-413, 2017.