

## ANALISIS SENTIMEN TWITTER MENGGUNAKAN TEXT MINING DENGAN ALGORITMA NAÏVE BAYES CLASSIFIER

*Adhi Viky Sudiantoro<sup>1</sup>, Eri Zuliarso<sup>2</sup>*

<sup>1,2</sup>Program Studi Teknik Informatika, Fakultas Teknologi Informasi, Universitas Stikubank  
e-mail: <sup>1</sup>adhi.viky@gmail.com, <sup>2</sup>eri299@edu.unisbank.ac.id

### **ABSTRAK**

*Analisis sentimen atau bisa di sebut juga opinion mining merupakan proses memahami, mengekstrak dan mengolah data tekstual secara otomatis untuk mendapatkan informasi sentimen yang terkandung dalam suatu kalimat opini terhadap sebuah masalah atau objek oleh seseorang, apakah cenderung beropini negatif atau positif. Penelitian ini bertujuan untuk mengklasifikasi data tweet menjadi 2 sentimen yaitu positif dan negatif. Dalam penelitian ini menggunakan teks Bahasa Indonesia yang terdapat di sosial media twitter berupa tweet. Opini masyarakat yang terdapat di tweet tersebut dapat dimanfaatkan sebagai bahan analisis sentimen untuk mengetahui masyarakat terhadap pilkada jawa barat apakah positif atau negatif. Data yang digunakan terdiri 300 data tweet yang dibagi menjadi 2 yaitu untuk data latih sebanyak 200 data dan 100 data untuk data uji. Pengklasifikasian data tweet menggunakan text mining dengan Naïve Bayes Classifier. Sebelum klasifikasi, dilakukan beberapa tahap pemrosesan teks seperti case folding, normalisasi, tokenisasi dan stopwords removal. Hasil dari 100 data uji yang klasifikasi menghasilkan 32 data bersentimen positif dan sebanyak 68 data sentimen negative. Dapat diartikan bahwa 100 data uji yang diklasifikasi masuk dalam kategori bersentimen negatif dikarenakan data positif lebih kecil daripada data yang bersentimen negatif. Adapun nilai akurasi algoritma Naïve Bayes Classifier memberikan nilai akurasi sebesar 84%.*

**Kata kunci:** Analisis sentimen, Text Mining, Naïve Bayes Classifier, klasifikasi, twitter.

### **1. PENDAHULUAN**

#### *1.1 Latar Belakang*

Pada saat ini perkembangan jejaring sosial sebagai alat komunikasi yang sangat populer, masyarakat menggunakan jejaring sosial sebagai salah satu media untuk berkomunikasi. Salah satunya jejaring sosial twitter, twitter digunakan sebagai sarana promosi produk, iklan, kampanye politik maupun sebagai sarana menyampaikan pendapat terkait kritik, saran, isu-isu dan opini-opini publik. Pilkada atau yang biasa disebut pemilihan kepala daerah yang akan dilangsungkan pada tahun 2018. Terdapat isu-isu yang terkait pada pilkada yang memunculkan banyak opini, tidak hanya opini yang positif tapi juga yang negatif. Oleh karena itu, Twitter di anggap lebih diminati oleh para masyarakat indonesia karna dirasa lebih mudah dan simpel dalam merepresentasikan opininya.

Text mining adalah salah satu teknik yang dapat digunakan untuk melakukan klasifikasi dokumen, clustering, information extraction, analisis sentimen dan information retrieval dimana text mining merupakan variasi dari data mining yang berusaha menemukan pola yang menarik dari sekumpulan data tekstual yang berjumlah besar [1].

Analisis sentimen atau bisa di sebut juga opinion mining merupakan proses memahami, mengekstrak dan mengolah data tekstual secara otomatis untuk mendapatkan informasi sentimen yang terkandung dalam suatu kalimat apakah beopini positif atau negatif. Analisis sentimen dapat sebagai solusi dari masalah yang sudah dipaparkan sebelumnya. Pada penelitian ini algoritma yang akan digunakan adalah NBC (Naive Bayes Classifier). Algoritma Naïve Bayes Classifier dirasa cocok digunakan pada analisis sentimen dikarenakan algoritma ini bertujuan sebagai metode klasifikasi kedalam kategori positif dan negatif.[2].

Berdasarkan latar belakang yang dikemukakan di atas maka dapat dirumuskan masalah penelitian sebagai berikut: Bagaimana implementasi algoritma NBC (Naive Bayes Classifier) dalam mengklasifikasikan data tweet pada media sosial twitter apakah data tweet tersebut termasuk ke dalam kelas bersentimen positif atau negatif ?

Selain itu agar masalah yang sedang di tinjau lebih terarah dan tercapai sasaran yang telah ditentukan, maka harus diberikan batasan-batasan masalah antara lain:

- a. Objek yang digunakan pada penelitian ini adalah data tweet dari pengguna twitter terkait pilkada jawa barat yang akan di langsupkan pada tahun 2018.
- b. Tweet yang akan digunkan yaitu kalimat tweet yang menggunakan bahasa Indonesia saja.
- c. Algoritma yang akan di gunakan untuk pengklasifikasian dalam penelitian ini adalah NBC (Naive Bayes Classifier).
- d. Pada penelitian ini proses stemming dan stopword hanya untuk kata-kata berbahasa indonesia saja.
- e. Bahasa pemograman yang digunakan dalam penelitian ini yaitu R.

Adapun tujuan pada penelitian ini adalah untuk mengklasifikasi sentiment pada data tweet dan menggunakan text mining dengan metode NBC (Naive Bayes Classifier) untuk analisis sentiment di media sosial twitter.

Adapun manfaat pada penelitian ini adalah sebagai berikut :

- a. Penelitian ini diharapkan dapat memberi tambahan pengetahuan tentang implementasi algoritma NBC (Naive Bayes Classifier) dan dapat memberikan gambaran mengenai bagaimana analisis sentimen pada jejaring sosial twitter untuk mengklasifikasi opini untuk berbagai kepentingan serta pengoptimalan informasi jejaring sosial untuk kepentingan publik.
- b. Selain itu juga bermanfaat untuk mencari sebuah informasi tentang suatu produk atau merek untuk mengetahui respon pasar terhadap produk atau merek tersebut apakah mempunyai respon positif atau negatif dari pasar dan juga bermanfaat bagi tokoh publik untuk melakukan pengukuran respon masyarakat terhadap dirinya.
- c. Penelitian ini juga bermanfaat untuk mengetahui berapa keberhasilan algoritma NBC (Naive Bayes Classifier) pada pengklasifikasian teks berbahasa Indonesia.

Penelitian sebelumnya yang pernah dilakukan berkaitan dengan masalah yang dihadapi penulis yaitu dengan judul “Analisis Sentimen pada Twitter untuk Mengenai Penggunaan Transportasi Umum dalam Kota dengan Metode Support Vector Machine”. penelitian dilakukan oleh Novantirani, Anita.,dkk pada tahun 2015. Pada penelitian ini hasil pengujian dan analisis yang dilakukan oleh penulis dengan metode support vector machine , dengan akurasi mencapai 78,12 %. Hasil akurasi terhadap penggunaan metode support vector machine di pengaruhi oleh beberapa hal yaitu dari komposisi jumlah data training dan testing, jumlah dataset yang digunakan dan komposisi jumlah data positif dan negatif.[3]

Penelitian lainnya adalah “Analisis Sentimen Hatespeech pada Twitter dengan Metode Naive Bayes Classifier dan Support Vector Machine”. Penelitian ini dilakukan oleh Asrofi Buntoro, Ghulam pada tahun 2016. Pada penelitian ini diperoleh hasil nilai akurasi tertinggi didapat dengan metode klasifikasi Support Vector Machine (SVM) dengan tokenisasi unigram, stopword list Bahasa Indonesia dan emoticons, dengan nilai rata-rata akurasi mencapai 66,6%, nilai presisi 67,1%, nilai recall 66,7% nilai TP rate 66,7% dan nilai TN rate 75,8%. [4]

Penelitian lainnya adalah “Analisis Sentimen Calon Gubernur DKI JAKARTA 2017 di Twitter”. Penelitian ini dilakukan oleh Asrofi Buntoro, Ghulam, pada tahun 2016. Data yang di gunakan dalam penelitian ini, yaitu 100 tweet dengan kata kunci AHY, 100 tweet dengan kata kunci Ahok, dan 100 tweet dengan kata kunci Anies. Pada penelitian ini diperoleh hasil nilai akurasi tertinggi didapat saat menggunakan metode klasifikasi Naive Bayes Classifier (NBC) untuk klasifikasi data AHY, dengan nilai rata-rata akurasi mencapai 95%, nilai presisi 95%, nilai recall 95% nilai TP rate 96,8% dan nilai TN rate 84,6%. [5]

## 2. METODE PENELITIAN

### 2.1 Analisis Data

Pada penelitian ini, data yang digunakan ialah tweet berbahasa Indonesia yang terdapat pada Twitter. Tweet yang digunakan ialah tweet-tweet yang mengandung opini masyarakat Indonesia terhadap pilkada Jawa Barat. Total tweet yang digunakan sebagai data berjumlah 300. Pemilihan data secara manual yaitu memilih kalimat-kalimat tweet yang berbahasa Indonesia dan tidak mengandung gambar. Tweet yang telah dipilih kemudian di simpan ke file excel. Data yang dibutuhkan dalam penelitian ini terdiri dari dua jenis, yaitu data latih dan data uji. Untuk kebutuhan latih, data yang berhasil dikumpulkan tersebut akan dikategorikan secara manual yang dilakukan oleh Penulis dan menilai sentimen yang terkandung di dalam Tweet tersebut dan menandai Tweet tersebut ke dalam 2 kategori sentimen yaitu Tweet yang mengandung sentimen negatif dan positif.. Data Tweet yang diambil untuk data latih adalah sebesar 200 data, dimana data ini terbagi menjadi beberapa bagian yaitu data yang mengandung kata positif dan data yang mengandung kata negatif.

Tabel 1. Tabel Jenis Sentimen

Jenis Sentimen Tweet	Positif	Negatif
Tweet Pilkada Jawa Barat	92	208
Latih	50	150
Uji	42	58

Dari tabel di atas yaitu data yang mengandung kata positif berjumlah 92 data dan data yang mengandung kata negatif berjumlah 208 data. Kemudian data tersebut dibagi untuk data training sebesar 200 data dan untuk testing sebesar 100 data.

### 2.2 Analisis Pemrosesan Teks

Pemrosesan teks merupakan proses menggali, mengolah, mengatur informasi dengan cara menganalisis hubungannya, aturan-aturan yang ada di data tekstual semi terstruktur atau tidak terstruktur. Untuk lebih efektif dalam proses pemrosesan dilakukan langkah transformasi data ke dalam suatu format yang memudahkan untuk kebutuhan pemakai. Proses ini disebut text processing. Setelah dalam bentuk yang lebih terstruktur dengan adanya

proses diatas data dapat dijadikan sumber data yang dapat diolah lebih lanjut. Tahapan untuk Pemrosesan teks terdiri dari tokenizing, normalisasi fitur, case folding dan stopword removal.

2.3 Analisis Implementasi Algoritma Naïve Bayes Classifier dan Hitung Akurasi

2.3.1. Implementasi Algoritma Naïve Bayes Classifier

Metode Naïve Bayes Classifier digunakan untuk melakukan pengkategorian, yaitu untuk melihat pendapat atau kecenderungan opini terhadap sebuah masalah atau objek oleh seseorang, apakah cenderung berkategori negatif atau positif. Data yang sudah melalui proses pemrosesan teks data kemudian akan melalui tahap klasifikasi menggunakan Naïve Bayes Classifier untuk mengetahui data tersebut apakah berkategori positif atau berkategori negatif.

2.3.2. Hitung Akurasi

Untuk menghitung tingkat akurasi dari klasifikasi sistem menggunakan confusion matrix. Data hasil klasifikasi akan dihitung berapa banyak data yang benar dan berapa banyak data yang salah.

3. HASIL DAN PEMBAHASAN

3.1 Pembahasan

- a. Menyiapkan data yang akan diolah diaplikasi RStudio, dalam penelitian ini data yang akan diolah yaitu data teks dari tweet yang ada di sosial media twitter yang berhubungan dengan pilkada Jawa Barat di ambil pada bulan Mei hingga Juni 2018. Data tersebut dimasukkan ke dalam file excel yang diberi nama newdataset3.csv. kemudian data tersebut diload kedalam aplikasi RStudio untuk diolah lebih lanjut.

Berikut sebagian data tweet yang berhubungan dengan pilkada Jawa Barat :

class	text
1 Neg	Nonton potongan video debat pilgub Jabar, ada pasang...
2 Neg	Saat2 Kader PDIP Buat Rusuh Di Debat Pilkada #Jabar Se...
3 Neg	Hobi kok bikin ricuh. Tanda-tanda kekalahan paslon no...
4 Neg	Kok bisa di acara debat gini ngomong ganti presiden. S...
5 Neg	Detik2 bludernya paslon #3 Pilkada Jabar yang akan b...
6 Neg	Mau tahu program utama yg ditawarkan paslon no.3 pil...
7 Neg	Brilian????...saya warga Jabar sudah pasti tdk akan mem...
8 Neg	Malu..lihat Debat Pilkada JABAR ini. 'KALAH DEBAT' Gant...
9 Neg	Debat pilkada Jabar malam ini : Kenapa pesorak pdip ga...
10 Neg	Kericuhan semalem di Debat Pilkada Jabar, gegara Paslo...
11 Neg	Belum jadi gubernur saja sudah pertontonkan ketidak p...
12 Neg	rupanya dr acara debat2 yang ada, debat pilkada Jabar y...
13 Neg	Ridwan Kamil menyalahkan Dedy Mizwar tp tidak berkac...
14 Neg	Menang pilkada Jabar lalu mau mengganti presiden 201...
15 Neg	pak.jendral. Pak malu kali kalau bapak pakai lembaga su...

Gambar 2. Data Tweet

- b. Selanjutnya tahap klasifikasi 100 data tweet yang sudah melalui tahap pemrosesan teks, data akan masuk ketahap klasifikasi untuk mengetahui dari 100 data tersebut masuk ke dalam kategori sentiment positif atau sentiment negatif. Setelah melalui tahap klasifikasi di dapat hasil sebagai berikut :

```
> predict(classifier, newdata=testNB)
[1] Pos Pos
Neg Pos
[24] Neg Pos Neg Pos Pos Neg Neg Pos Pos Neg Neg Neg Neg Neg Pos Neg Neg Neg
Neg Neg
[47] Neg Neg Neg Neg Neg Neg Neg Pos Neg Neg Neg Neg Neg Neg Neg Neg Neg
Pos Neg
[70] Pos Neg Neg
Neg Neg
[93] Neg Neg Neg Neg Neg Neg Neg
Levels: Neg Pos
```

Gambar 3. Hasil Klasifikasi

Dari gambar tersebut data yang mengandung sentiment positif sebanyak 32 data dan data yang mengandung sentiment negatif sebanyak 68 data. Dapat diartikan bahwa 100 data uji yang sudah melalui tahap klasifikasi masuk dalam kategori data yang bersentimen negatif dikarenakan data positif lebih kecil daripada data yang bersentimen negatif.

- c. Dari hasil klasifikasi akan dilanjutkan ketahap perhitungan akurasi. Tahap akurasi berguna untuk mengetahui kinerja dari algoritma naïve bayes classifier dalam mengklasifikasi data teks apakah mendapatkan akurasi yang tinggi atau malah mendapatkan akurasi yang rendah. Berikut gambar yang menunjukkan hasil uji akurasi dari algoritma naïve bayes classifier :

```

Accuracy : 0.84
95% CI : (0.7532, 0.9057)
No Information Rate : 0.58
P-Value [Acc > NIR] : 2.27e-08

Kappa : 0.6604
McNemar's Test P-Value : 0.02445

Sensitivity : 0.9483
Specificity : 0.6905
Pos Pred Value : 0.8088
Neg Pred Value : 0.9063
Prevalence : 0.5800
Detection Rate : 0.5500
Detection Prevalence : 0.6800
Balanced Accuracy : 0.8194

'Positive' Class : Neg

```

Gambar 4. Hasil Akurasi

Dari gambar tersebut di dapat hasil uji akurasi sebanyak 84 %. Dapat disimpulkan bahwa algoritma Naive Bayes classifier cukup berhasil memprediksi kategori sentimen yang benar karena dari hasil uji akurasi algoritma naïve bayes classifier mendapatkan akurasi yang tinggi yaitu 84% yang artinya kinerja algoritma naïve bayes classifier dapat mengklasifikasi data teks dengan sangat baik.

#### 4. KESIMPULAN

Berdasarkan hasil analisis dan pengujian yang telah dilakukan pada, maka dapat diambil kesimpulan sebagai berikut :

- Algoritma Naïve Bayes Classifier sangat efektif untuk digunakan sebagai proses klasifikasi tweet yang dibutuhkan dalam sistem analisis sentiment ini dimana nilai yang di dapatkan dalam pengujian sampai 84%.
- Metode Naïve Bayes Classifier dapat digunakan untuk melakukan klasifikasi tweets dengan cukup baik pada sistem analisis sentimen.
- 300 data yang dibagi menjadi 2 yaitu data latih sebanyak 200 data dan data uji berjumlah 100 data. Hasil dari klasifikasi diketahui bahwa 100 data yang diuji masuk dalam kategori sentimen negatif.

#### 5. SARAN

Penulis menyarankan pengembangan penelitian lebih lanjut sistem pengklasifikasian Tweet sebagai berikut:

- Peneliti menyarankan untuk penelitian selanjutnya untuk penambahan jumlah data latih dan data uji untuk mendapatkan hasil yang lebih baik saat klasifikasi tweets.
- Peneliti menyarankan untuk penelitian selanjutnya bahasa yang digunakan juga tidak hanya bahasa Indonesia tetapi dapat menggunakan bahasa daerah atau bahasa asing seperti bahasa inggris dan bahasa lainnya.
- Peneliti menyarankan untuk penelitian selanjutnya pada tahap pemrosesan teks ditambah fitur stemming untuk mendapatkan hasil yang lebih baik
- Peneliti menyarankan untuk penelitian selanjutnya tidak hanya menggunakan algoritma Naïve Bayes Classifier saja tetapi juga menggunakan algoritma support vector machine dan algoritma klasifikasi teks yang lainnya.

#### DAFTAR PUSTAKA

- Manalu, Boy Utomo. (2014). Analisis Sentimen Pada Twitter Menggunakan Text Mining. Universitas Sumatera Utara Medan.
- Muslim, Alif Akbarul. (2016). Rancang Bangun Aplikasi Berbasis Web Untuk Analisa Sentimen Pada Mikroblog Twitter Dengan Metode Naive Bayes. Universitas Islam Negeri Maulana Malik Ibrahim Malang.
- Novantirani, Anita dkk. (2015). Analisis Sentimen pada Twitter untuk Mengenai Penggunaan Transportasi Umum Darat Dalam Kota dengan Metode Support Vector Machine. Universitas Telkom Bandung
- Buntoro, Ghulam Asrofi. (2016). Analisa Sentimen Hatespeech pada Twitter dengan Metode Naïve Bayes Classifier dan Support Vector Machine. Universitas Muhammadiyah Ponorogo.
- Buntoro, Ghulam Asrofi. (2016). Analisis Sentimen Calon Gubernur DKI Jakarta 2017 Di Twitter. Universitas Muhammadiyah Ponorogo.