Volume 4, Nomor 2, Desember 2023, hlm 224-233
INFOTECH: Jurnal Informatika Teknologi
p ISSN 2722-9378 | e ISSN 2722-9386
http://jurnal.sttmcileungsi.ac.id/index.php/infotech

# Analisis sentimen twitter terhadap pemilihan presiden menggunakan algoritma naïve bayes

## *Sentiment analysis twitter of presidential election using naïve bayes algorithm*

**Panji Al Muqsith Prasetyo*, Arief Hermawan**

* Jurusan Informatika, Fakultas Sains & Teknologi, Universitas Teknologi Yogyakarta, Jl. Siliwangi Jl. Ring Road Utara, Jombor Lor, Sendangadi, Kec. Mlati, Kabupaten Sleman, Daerah Istimewa Yogyakarta 55285

**INFORMASI ARTIKEL**

***Article History:***
*Submission: 17-10-2023*
*Revised: 15-11-2023*
*Accepted: 21-11-2023*

**Kata Kunci**:
Analisis sentiment; naïve bayes; Masyarakat; pemilihan presiden

***Keywords:***
*Sentiment analysis, naïve bayes; people; presidential election*

*** Korespondensi:**
**Panji Al Muqsith Prasetyo**
panji.5400411249@student.uty.ac.id

**ABSTRAK**

Tahun 2024 esok merupakan akhir dari masa kepemimpinan presiden Joko Widodo yang mana menjadi penentu dari presiden selanjutnya. Saat ini masyarakat indonesia sedang heboh dengan maraknya calon presiden dan calon wakil presiden yang sudah ditentukan. Capres-cawapres saat ini yang dapat dipastikan akan berkompetisi di ajang pemilihan presiden antara lain Ganjar-Mahfud, Prabowo-Gibran, dan Anies-Cak imin. Berbagai komentar dari banyaknya platform internet seperti sosial media khususnya media sosial Twitter atau X menjadi tempat menuai opini dan reaksi terhadap ketiga capres-cawapres tersebut. Ini menjadi masalah bagi masyarakat yang kurang literasi dalam kebahasaan dan literasi digital, mereka akan mudah terpancing serta tergiring oleh opini dari netizen yang memiliki literatur lebih baik. Dengan munculnya masalah tersebut, maka dikembangkan sistem sentimen analisis ini yang menjadi salah satu platform untuk menentukan memiliki sifat apa komentar tersebut secara otomatis. Sifat dari komentar umumnya dibagi menjadi 3, yaitu netral, positif, dan negatif. Ditambah, dengan metode *crawling* menggunakan API dari X (Twitter) akan mempermudah dalam mengumpulkan dataset komentar dengan lebih fleksibel. Dari hasil crawling ini, didapat komentar sebanyak 540. Metode lain yang akan digunakan untuk proses sentimen analisis ini yaitu algoritma *naive bayes* dengan tipe multinomialNB. Penelitian ini menghasilkan nilai akurasi sebesar 85% dengan precision 86.54%, *recal*l 85%, dan f1-score sebesar 85% dimana hasil ini menggunakan skenario test_size sebesar 0.2 yang menjadi skenario paling baik dalam pembangunan model.

***ABSTRACT***

*The year 2024 tomorrow is the end of President Joko Widodo's leadership period which is the determinant of the next president. The ascent of predetermined candidates for president and vice president is currently thrilling the Indonesian populace. Ganjar-Mahfud, Prabowo-Gibran, and Anies-Cak imin are the confirmed contenders for the office of president and vice president in the next presidential election. Various comments from many internet platforms such as social media, especially social media, Twitter or X became a place to reap opinions and reactions to the three vice presidential candidates. This is a problem for people who lack literacy in language and digital literacy, they will be easily provoked and led by opinions from netizens who have better literature. With the emergence of these problems, this sentiment analysis system was developed which became one of the platforms to determine what nature the comment had*

*automatically. The nature of comments is generally divided into 3, namely neutral, positive, and negative. Plus, crawling using the API from X (Twitter) will make it easier to collect comment datasets more flexibly. From the results of this crawling, 540 comments were obtained. The multinomialNB type of the Naïve Bayes algorithm will also be employed in this sentiment analysis process. This study used a test_size scenario of 0.2, which is the ideal scenario in model development, and produced an accuracy value of 85% with 86.54% precision, 85% recall, and 85% f1-score.*

## 1. PENDAHULUAN

Indonesia adalah negara Asia Tenggara dengan sistem pemerintahan presidensial. Presiden adalah gelar jabatan resmi yang memegang kepala organisasi, asosiasi, perusahaan, universitas atau negara. Di Indonesia, konsep presiden didasarkan pada UUD 1945, dimana presiden adalah jabatan tertinggi dalam sistem pemerintahan yang menguasai kekuasaan di Indonesia. Dalam UUD 1945, Pasal 7, yang berbunyi "Presiden dan Wakil Presiden memegang jabatan selama lima tahun, dan sesudahnya dapat dipilih kembali dalam jabatan yang sama, hanya untuk satu kali masa jabatan," menetapkan bahwa masa jabatan presiden di Indonesia dapat mencapai maksimal dua periode, atau 10 tahun [1].

Seiring dengan masa kepresidenan Joko Widodo akan berakhir pada beberapa waktu yang akan datang, banyak kandidat-kandidat baru dari berbagai partai politik saling mengajukan siapa yang akan menjadi presiden selanjutnya. Pada bulan November 2023 ini telah hadir 3 kandidat capres-cawapres untuk berkompetisi dalam ajang pemilihan presiden menggantikan presiden saat ini Joko Widodo yang telah memegang status presiden sebanyak 2 periode. Kandidat-kandidat yang dimaksud adalah Ganjar Pranowo yang bersanding dengan Mahfud MD, Prabowo Subianto yang bersanding dengan Gibran Rakabuming Raka, dan Anies Baswedan yang bersanding dengan Muhaimin Iskandar. Komentar-komentar dari berbagai macam platform media sosial menjadi acuan pandangan masyarakat terhadap bagaimana sistem pemerintahan Indonesia jika memilih salah satu kandidat tersebut, salah satu media sosialnya adalah Twitter. Twitter merupakan platform yang umum digunakan oleh publik untuk memposting opini, pandangan, dan komentar tentang topik tertentu. Umumnya, sebuah komentar mengandung unsur mendukung (positif), netral, dan negatif (tidak mendukung). Ini akan menjadi masalah bagi Masyarakat yang kurang dan bahkan tidak memiliki literasi kebahasaan serta digitalisasi, mereka akan sangat mudah untuk digiring opininya, menyebarkan hoax, dan hal lainnya yang dapat merugikan banyak orang.

Sentimen, opini, dan komentar dapat ditentukan secara manual, tetapi jika terdapat banyak data, pasti akan memakan waktu lama. Masalah ini dapat diatasi dengan membangun sistem analisis sentimen yang dapat mengklasifikasikan komentar menjadi positif, netral, dan negatif. Dengan demikian, pengguna dapat dengan mudah mengetahui bahwa komentar tersebut memiliki tujuan positif, netral, maupun negatif dengan lebih rinci. Metode *Naïve Bayes* umum digunakan dalam mengklasifikasikan data suatu nilai dari variabel pada data testing [2]. Dalam penelitian ini, nilai yang dimaksud adalah sentimen yang memiliki 3 nilai, yaitu -1 (tidak mendukung salah satu pihak), 0 (netral/tidak berpihak pada sisi manapun), dan 1 (sangat mendukung salah satu pihak) [3].

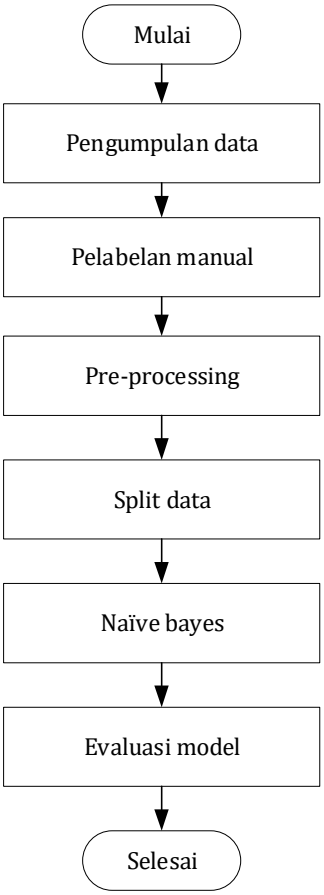
Naive Bayes banyak digunakan dalam penelitian karena menawarkan cara yang lebih baik untuk mengklasifikasikan data dalam hal akurasi dan perhitungan dibandingkan dengan metode klasifikasi lainnya [4]. Alasan mengapa metode ini banyak digunakan pada penelitian dengan tema analisis sentimen adalah dikarenakan algoritma perhitungan yang cepat serta tinggi akurasi sehingga suatu sistem dapat bekerja optimal [5]. Pada penelitian ini digunakan naïve bayes sebagai metode pembelajaran mesin, karena metode ini telah diuji dalam banyak penelitian berdasarkan *Naive Bayes* dan memberikan akurasi dan hasil evaluasi yang tinggi.

Pada penelitian ini dilaksanakan bagaimana sistem mampu mengklasifikasikan komentar yang tersebar di Twitter, yang terdiri dari komentar positif, netral, dan negatif terkait pemilihan presiden. Penelitian ini akan menggunakan data dari Twitter melalui teknik *scraping* data yang digunakan oleh API Twitter. Penelitian ini menggunakan metode machine learning "Naive Bayes Classifier" karena dapat membantu dalam klasifikasi sentimen. Tujuan dari penggunaan metode ini adalah untuk mengetahui seberapa baik metode ini bekerja dan untuk membantu sistem

melakukan analisis sentimen otomatis. Kelebihan dari penelitian ini adalah penggunaan metode normalisasi kata. Di mana, banyak kata tertentu yang disingkat, tidak lengkap, dan lainnya yang dapat menjadi masalah ketika pembangunan model dan probabilitas pada tahap TF-IDF. Metode yang digunakan untuk mengumpulkan dataset penelitian ini adalah dengan memanfaatkan API yang disediakan oleh twitter developer serta menggunakan bahasa pemrograman Python untuk proses *backend*.

## 2. METODE

Metode yang dilakukan ada 6 tahap, yaitu mengumpulkan data, pelabelan manual, *pre-processing* dataset, split data, pembangunan model dengan *naïve bayes*. Gambar 1 merupakan *flowchart* untuk melakukan penelitian.




Gambar 1. Tahap penelitian

### 2.1 Pengumpulan data

Dalam penelitian ini, sentimen yang digunakan berasal dari *teks tweet* berbahasa Indonesia yang diposting secara langsung di *platform* media sosial *Twitter*; *platform* ini adalah salah satu dari banyak platform media sosial yang memungkinkan setiap komentar dipublikasikan secara publik, sehingga setiap orang dapat melihat, memberikan tanggapan atau balasan, me-retweet, atau bahkan menyukai komentar tersebut. Dengan kata lain, *Twitter* merupakan situs yang memungkinkan penggunanya memposting meliputi suatu topik dan membahasnya lebih lanjut terkait banyaknya isu-isu yang terjadi saat ini. Terdapat 4 *keyword* yang dibutuhkan untuk mendapatkan dan mengumpulkan data yaitu, “pemilihan presiden”, “pemilihan presiden ganjar”, “pemilihan presiden prabowo”, dan “pemilihan presiden anies”.

### 2.2 Pelabelan data manual

Pelabelan manual ini dibutuhkan dikarenakan machine learning naïve bayes ini bersifat supervised, di mana sebuah data harus memiliki label atau target. Selain itu pelabelan manual ini tidak bisa dilakukan secara sembarang. Ini dikarenakan penelitian ini membutuhkan tingkat literasi tertentu, agar sebuah sentimen dapat dengan pasti merujuk pada sifat apa suatu sentimen. Tabel 1 merupakan contoh data yang sudah di *crawling* dan telah ditentukan kelas sentimennya.

Tabel 1. Contoh data

| Data ke- | Sentimen | Kelas sentimen | Deskripsi |
|---|---|---|---|
| 1 | saya dukung kerja dan janji pak ganjar pranowo | 1 | Positif |
| 2 | cara pandang paslon itu tidak buruk, tapi gimana implementasinya? | 0 | Netral |
| 3 | paslon anies aneh! masa iya gak bisa jawab pertanyaan rakyatnya? | -1 | Negatif |

Merujuk pada Tabel 1, pelabelan manual dilakukan bertujuan untuk mengetahui jenis kalimat dalam suatu komentar atau sentimen. Label dataset yang akan ditentukan terbagi menjadi 3 kelas sentimen, yaitu positif, netral, dan negatif. Tiap sentimen harus memiliki satu diantara 3 kelas yang diajukan, masing-masing kelas menunjukkan suatu komentar apakah bersifat mendukung (positif), netral, dan tidak mendukung (negatif).

### 2.3 Pre-processing

Tahapan dalam proses pre-processing adalah tahapan lanjutan dari pelabelan data manual, dimana data mentah yang sebelumnya dikumpulkan menggunakan metode *crawling* akan diubah menjadi data yang siap untuk digunakan dalam sistem. Tahap *pre-processing* ini harus dilakukan untuk sistem berbasis teks, dikarenakan suatu karakter akan dibobot yang nantinya akan dihitung dan mengantisipasi adanya kekurangan kualitas dan kinerja Naïve Bayes dalam implementasinya. Tahap pre-processing terdiri dari 6 tahap [6], yaitu:

a. *Case Folding*, adalah proses mengubah suatu kata dalam suatu data teks menjadi 1 ragam, umumnya 2 jenis yang digunakan adalah *uppercase* dan *lowercase* [7]. Sebagai contoh Saya diubah menjadi saya.
b. *Cleaning*, proses pembersihan keseluruhan teks dalam suatu sentimen dari karakter yang tidak diperlukan. Contohnya angka, *link*, karakter yang bukan huruf seperti @()*$, dan lain sebagainya [8].
c. T*okenizin*g, proses pemisahan atau pemecah suatu kata dari suatu kalimat menjadi beberapa kata. Contohnya suatu kalimat “saya cinta indonesia" diubah menjadi [“saya”, “cinta”, “indonesia”] [9].
d. Normalisasi, merupakan proses bertujuan untuk mengubah kata-kata yang dipersingkat dan kata-kata yang tidak baku.
e. S*temming*, adalah sebuah proses penghapusan kata yang memiliki kata imbuhan me-, -kan, di-, -an, dan lain sebagainya.
f. *Stopword*, yaitu proses penghapusan kata yang tidak memiliki arti.

### 2.4 Split data

Split data adalah proses pembagian data antara data tes dan data latih yang akan digunakan untuk keperluan algoritma [10]. Proses ini dibagi menjadi 3 bagian yaitu menentukan variabel x dan y, lalu besarnya pembagian data, dan menentukan *random state*. Variabel x ini nantinya akan menjadi data berupa kolom sentimen dan y akan berisi kolom jenis sentimen (positif, netral, dan negatif). Besarnya split data beragam, mulai dari 0.1, 0.2, dan yang paling umum adalah 0.3. Split data ini menggunakan *function* dari *library scikit-learn*.

Tabel 2. Besaran split data

| **Persentase Data** | | **Jumlah Data** | |
|---|---|---|---|
| **Data train** | **Data test** | **Data train** | **Data test** |
| 90% | 10% | 468 | 52 |
| 80% | 20% | 416 | 104 |
| 70% | 30% | 364 | 156 |

Pada Tabel *2*, dilakukan pengujian dengan skenario berbeda yakni dengan data 90:10 dimana data train sebesar 468 dan data test sebesar 52, kemudian dengan data 80:20 di mana data train sebesar 416 dan data test sebesar 104, dan terakhir dengan data 70:30 di mana data data train sebesar 364 dan data test sebesar 156. Setelah dilakukan proses split data, proses dilanjut ke pembobotan kata dengan TF-IDF.

2.5 Metode naïve bayes

Naïve Bayes adalah salah satu algoritma pembelajaran mesin yang digunakan untuk jenis data yang dengan label, metode ini cocok untuk sistem yang membutuhkan klasifikasi multi-class [11]. Metode ini menggunakan klasifikasi menggunakan teori probabilitas dan statistik ilmuwan Inggris Thomas Bayes. Asumsi yang sangat kuat (naif) dan independensi dari setiap situasi atau kejadian merupakan karakteristik utama pengklasifikasi Naive Bayes.

a. Pembobotan kata dengan TF-IDF

Pembobotan kata merupakan proses dimana kata akan diubah menjadi frekuensi vektor dimana nantinya hasil dari pembobotan kata ini dapat meningkatkan kemampuan pada sistem analisis sentiment [12]. TF-IDF merupakan singkatan dari Term *Frequency–Inverse Document Frequency*, dimana kata akan dimasukkan ke algoritma pembobotan kata. Kegunaan dari algoritma TF-IDF adalah untuk mencari dan mengumpulkan representasi suatu kata atau nilai dari suatu kumpulan dokumen train data (x) yang nantinya akan membentuk suatu nilai vektor [13].

Dengan menggunakan contoh data kalimat pada Tabel 1, tiap kata dalam dokumen positif akan dikumpulkan menjadi 1. Hal serupa berlaku untuk tiap kelas sentimen, dengan adanya algoritma TF-IDF akan memudahkan Naïve Bayes dalam menyeleksi fitur dalam implementasinya. TF-IDF memiliki 2 tahap, TF yaitu menghitung seberapa umum kata tersebut dalam tiap dokumen, semakin sering suatu kata muncul, maka semakin tinggi bobotnya. Kemudian ada tahap IDF, dimana metode ini memberikan bobot yang lebih tinggi pada suatu kata yang jarang muncul dalam dokumen dibandingkan kata yang lebih sering muncul. Sementara itu TF-IDF merupakan penggabungan kedua tahap, TF-IDF diperlukan untuk menentukan apakah suatu tersebut penting atau tidak dalam dokumen, penentuan ini didasari oleh frekuensi kemunculannya dalam keseluruhan dokumen tersebut dan dalam kumpulan data secara keseluruhan.

Selain itu, TF-IDF memiliki rumusnya tersendiri. Dibawah ini merupakan rumus untuk TF, IDF, dan TF-IDF [14].

$$TF(t,d) = \frac{n_{t,d}}{Total\ jumlah\ term\ dalam\ d} \quad (1)$$

$$IDF_{d=Log\left(\frac{jumlah\ keseluruha\quad dokumen}{Jumlah\ dokumen\ yang\ memiliki\ kata\ t}\right)} \quad (2)$$

$$TF - IDF = TF * IDF \quad (3)$$

b. Naïve Bayes

Algoritma ini didasarkan pada teorema Bayes dengan asumsi semua kata-kata dalam dataset adalah independen satu sama lain, yang seringkali merupakan asumsi yang cukup sederhana namun berguna dalam banyak kasus [8]. Algoritma ini umum digunakan untuk sebuah model dikarenakan akurasi yang tinggi dan perhitungan yang sederhana, Naïve Bayes memiliki rumus tersendiri yang dapat dilihat pada rumus 1 [15].

$$\mathrm{P}(A|B)\frac{P(B|A)*P(A)}{P(B)} \quad (4)$$

Keterangan:
P(A|B) : Suatu kata (B) termasuk ke dalam kelas A
P(B|A) : Probabilitas kata (B) muncul dalam kelas A
P(A) : Probabilitas prior, probabilitas suatu kata secara umum termasuk kelas A
P(B) : Probabilitas bahwa suatu kata (B) muncul bersamaan

c. Evaluasi model

Pada penelitian ini, model yang telah melalui proses learning menggunakan algoritma Naïve Bayes perlu dilakukan evaluasi model. Evaluasi model ini menggunakan *classification report*, dimana berisi hasil *accuracy*, *recall*, *precision*, dan F1 *Score*.

## 3. HASIL DAN PEMBAHASAN

Pada bab hasil dan pembahasan ini akan membahas semua hal terkait metode penelitian yang sebelumnya telah dijabarkan. Mulai dari tahap *crawling* hingga evaluasi model machine learning. Setiap gambar yang disampaikan agar disampaikan pembahasan, apa maksud dari gambar tersebut.

### 3.1 Hasil pengumpulan data

Penelitian ini menggunakan teknik *crawling* menggunakan API *Twitter Developer* yang telah terdaftar sebelumnya, serta menggunakan bahasa pemrograman Python dan *library tweepy* untuk mendapatkan data teks. Tabel *3* data yang didapatkan dari hasil *crawling* ini adalah sebanyak 540 data, dengan 200 data positif, 199 data netral, dan 141 data negatif.

Tabel 3. Hasil *crawling* data

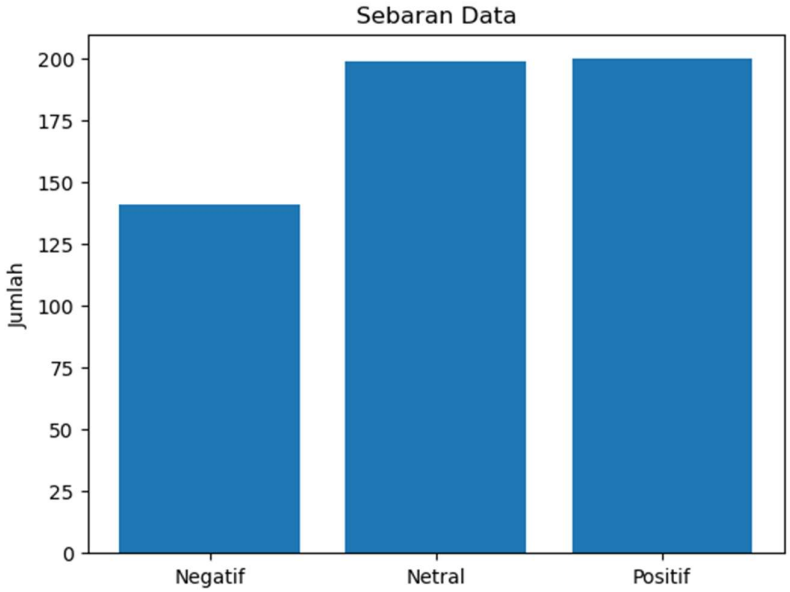
| No. | created_at | username | text |
|---|---|---|---|
| 1 | 2022-12-18 11:10:34+00:00 | Tukang Retweet | b'RT @fim_mifta: Sudah jd rahasia umum kl dunia politik Indonesia saat ini adalah politik transaksional. Itu berlaku dlm pemilihan kades, bup\xe2\x80\xa6' |
| 2 | 2022-12-16 15:41:54+00:00 | Awan Martha irawan | b'RT @__RismaWidiono_: Untaian dukungan serta Doa untuk @ganjarpranowo agar maju pada pemilihan Presiden Tahun 2024 semakin terdengar di seluruh \xe2\x80\xa6' |
| … | … | … | … |
| 540 | 2022-12-16 11:29:13+00:00 | Kejaksaan Tinggi Riau | b'RT @KejaksaanRI: Bahwa isu yang strategis untuk 2023-2024 mendatang adalah terkait dengan adanya proses pemilihan umum serentak baik itu Pe\xe2\x80\xa6' |

Dari hasil *crawling* data menggunakan *tweepy*, didapatkan 3 atribut. Created_at digunakan untuk menjelaskan kapan suatu sentimen dibuat, lalu username digunakan untuk mengetahui siapa yang membuat sentimen tersbeut, dan atribut teks merupakan poin utama dimana sebuah sentimen dapat dilakukan pelabelan manual.

### 3.2 Pelabelan manual

Pelabelan manual menggunakan software tambahan berupa Microsoft Excel untuk menentukan apakah sebuah sentimen mengandung satu sifat dari tiga sifat yang dijabarkan sebelumnya. Pelabelan manual ini sangat sulit jika dilakukan dengan otomatis, dikarenakan mesin belum cukup mampu dengan baik untuk mendeteksi kelas sentimen tertentu jika itu berbahasa

Indonesia. Gambar *2* merupakan gambar sebaran data dari proses *crawling* yang sebelumnya didapat sebanyak 540 data dan sudah dibersihkan dari data yang duplikat.




Gambar 2. Sebaran data

3.3 Pre-processing data

Data yang sebelumnya telah dikumpulkan serta dilakukan pelabelan manual akan masuk ke tahap pre-processing. Dimana data tersebut diproses serta diubah ke dalam data matang atau data yang siap pakai, tahap pre-processing ini memiliki 6 tahap. Dimana masing-masing tahap memiliki fungsinya tersendiri, hasil dari masing-masing tahap pre-processing dapat dilihat pada Tabel *4*.

Tabel 4. Contoh hasil *pre-processing*

| *Pre-processing* | Hasil | kelas_sentimen |
|---|---|---|
| Teks mentah | b'RT @fim_mifta: Sudah jd rahasia umum kl dunia politik Indonesia saat ini adalah politik transaksional. Itu berlaku dlm pemilihan kades, bup\xe2\x80\xa6' | -1 |
| *Case Folding* | b'rt @fim_mifta: sudah jd rahasia umum kl dunia politik indonesia saat ini adalah politik transaksional. itu berlaku dlm pemilihan kades, bup\xe2\x80\xa6' | |
| *Text Cleaning* | sudah jd rahasia umum kl dunia politik indonesia saat ini adalah politik transaksional itu berlaku dlm pemilihan kades bupxexxa | |
| Tokenizing | [sudah, jd, rahasia, umum, kl, dunia, politik, indonesia, saat, ini, adalah, politik, transaksional, itu, berlaku, dalam, pemilihan, kades, bupxexxa] | |
| *Normalize Text* | [sudah, jadi, rahasia, umum, kalau, dunia, politik, indonesia, saat, ini, adalah, politik, transaksional, itu, berlaku, dalam, pemilihan, kades, bupati] | |
| *Stemming* | sudah jadi rahasia umum kalau dunia politik indonesia saat ini adalah politik transaksional itu laku dalam pilih kades, bupati | |
| *Stopword* | sudah jadi rahasia umum kalau dunia politik indonesia ini politik transaksional laku pilih kades, bupati | |

Pada baris *stopword*, proses *pre-processing* selesai dilakukan dan sudah siap digunakan (matang) untuk proses selanjutnya. Data yang digunakan untuk proses selanjutnya adalah keseluruhan data yang terdapat pada kolom *stopword*.

### 3.4 Pembobotan Kata dengan TF-IDF

TF-IDF merupakan salah satu metode *bag of words* yang sederhana dimana suatu kata akan diubah menjadi suatu vektor [10]. *Bag of words* ini menghasilkan output perhitungan frekuensi kemunculan kata pada keseluruhan dokumen. TF-IDF sendiri terbagi menjadi 2 tahap yaitu TF dan IDF, TF adalah proses menghitung suatu kata dalam seluruh dokumen, sementara IDF merupakan proses menghitung dan mengukur apakah suatu kata penting atau tidak. Semakin kecil nilainya, maka semakin tidak penting kata tersebut. Di lain sisi, TF-IDF adalah proses perkalian antara TF dan IDF yang menghasilkan nilai berupa vektor serta menjadi patokan suatu algoritma machine learning dalam memahami data berupa teks. Sebagai gambaran, TF-IDF akan ditampilkan pada tabel 5.

Tabel 5. Contoh hasil perhitungan TF-IDF

| Kata ke- | Kata | TF | IDF | TF-IDF |
|---|---|---|---|---|
| 1 | sudah | 1 | 2.73239375982 | 2.73239375982 |
| 2 | jadi | 1 | 2.73239375982 | 2.73239375982 |
| 3 | rahasia | 1 | 2.73239375982 | 2.73239375982 |
| 4 | umum | 1 | 2.73239375982 | 2.73239375982 |
| 5 | kalau | 1 | 2.73239375982 | 2.73239375982 |
| 6 | dunia | 1 | 2.73239375982 | 2.73239375982 |
| 7 | politik | 2 | 2.73239375982 | 5.46478752 |
| 8 | indonesia | 1 | 2.73239375982 | 2.73239375982 |
| 9 | ini | 1 | 2.73239375982 | 2.73239375982 |
| 10 | transaksional | 1 | 2.73239375982 | 2.73239375982 |
| 11 | laku | 1 | 2.73239375982 | 2.73239375982 |
| 12 | pilih | 1 | 2.73239375982 | 2.73239375982 |
| 13 | kades | 1 | 2.73239375982 | 2.73239375982 |
| 14 | bupati | 1 | 2.73239375982 | 2.73239375982 |
| **Total** | | | | 40.9859064 |

Sebagai contoh perhitungan TF-IDF, dibawah ini menggunakan kata “politik” sebagai gambaran bagaimana TF-IDF bekerja menghitung sebuah kata. Diketahui kata “politik” muncul sebanyak 2 kali dalam keseluruhan dokumen

$$TF = 2$$

$$IDF = log10\left(\frac{5402}{2}\right) = 2.73239375982$$

$$TF - IDF = 2 * 1.87506126 = 5.46478752$$

Proses terus dilakukan pada tiap kata dan keseluruhan dokumen dalam suatu dataset. Jika semua kata sudah dihitung, kemudian keseluruhan TF-IDF tiap kata dijumlahkan. Ketika proses TF-IDF telah selesai, selanjutnya adalah tahap pembangunan model dengan *Naïve Bayes*.

### 3.5 Naïve Bayes

Pada proses *Naïve Bayes* ini terdapat 3 tahap, yaitu proses penentuan x dan y, kemudian pembagian data, proses algoritma *Naïve Bayes*, serta evaluasi model. Variabel x yang digunakan dalam *Naïve Bayes* ini adalah *data frame* kolom *Stopword*, sementara itu target atau variabel y yang digunakan adalah sentimen pada tiap-tiap kolom. Kemudian pembagian data berupa variabel test_size adalah sebesar 0.1 dan random_state adalah 0, maksud dari test_size ini adalah seberapa besar data yang akan di test dari keseluruhan data. Sebagai contoh suatu data berisi 10 baris dan ditentukan test_size sebesar 0.3, maka 30% dari 100% data (10 data) akan dibagi menjadi 30% data test dan 70% data train. Sementara itu untuk parameter random_state menentukan seberapa acak sebuah data akan dibagi, tiap angka akan merubah dalam pemilihan

datanya. Proses dilanjut ke dalam algoritma *Naïve Bayes*, jenis *Naïve Bayes* yang digunakan adalah *Multinomial Naïve Bayes*. Dimana jenis ini berfokus pada data teks yang sebelumnya telah dilakukan pembobotan kata.

### 3.5.1 Evaluasi model

Hasil evaluasi model akan menampilkan pembangunan model dengan *Naïve Bayes*, hasil ini merupakan hasil dengan skenario terbaik yaitu test_size sebesar 20% (0.2).

Tabel 6. Hasil evaluasi Model

| Hasil | Akurasi | Precision | Recall | F1-Score |
|---|---|---|---|---|
| Skor | 85% | 86.54% | 85.09% | 85% |

Pada tabel 6, skenario test size 20% memberikan hasil yang tinggi mengingat jumlah data yang sedikit digunakan untuk test dan lebih banyak data train. Sehingga algoritma naïve bayes bekerja dengan lebih optimal dan efisien.

### 3.5.2 Perbandingan hasil

Sebagai analisis tambahan, peneliti menambahkan hasil pembangunan model dengan skenario lain. Analisis yang dimaksud adalah perubahan parameter pada test_size untuk tahap split data. Perubahan jumlah split data seringkali menghasilkan nilai yang fluktuatif antara kondisi split data satu dengan lainnya.

Tabel 7. Perbandingan hasil

| **test_size** | **Akurasi** | **Precision** | **Recall** | **F1-Score** |
|---|---|---|---|---|
| 0.1 | 84% | 85.32% | 84% | 83.51% |
| 0.2 | 85% | 86.54% | 85% | 85% |
| 0.3 | 85% | 86.45% | 85% | 84.64% |

Tabel 7, perubahan parameter tiap baris memiliki hasil yang berbeda-beda. Ini dikarenakan variabel test_size tertentu memiliki cara penggunaan yang berbeda pula terhadap machine learning. Dalam kondisi test_size 0.2, performa naïve bayes dapat mencapai nilai pair classification. Dalam kondisi lain, akurasi menjadi rendah dikarenakan test_size tersebut lebih tinggi sehingga pemilihan data untuk machine learning tidak dapat membuat model dengan optimal atau random_state yang terlalu acak sehingga pemilihan data rentan kurang bagus.

## 4. SIMPULAN

Berdasarkan hasil dan pembahasan dari penelitian yang telah dilakukan, metode klasifikasi sentimen menggunakan algoritma Naive Bayes dapat dilakukan dengan performa baik. Data sebanyak 540 dari teknik *crawling* berhasil berjalan dengan baik serta menghasilkan output yang diharapkan berdasarkan metodologi yang sebelumnya telah dijabarkan. Data ini lebih dominan pada kata yang bersifat positif sehingga dalam beberapa percobaan hasil dari klasifikasi dengan data baru cenderung lebih banyak mengeluarkan ouput positif, dengan akurasi yang cukup tinggi dengan nilai mencapai 77%. Sistem ini perlu dikembangkan lebih lanjut agar akurasi yang didapatkan lebih tinggi. Dengan implementasi sistem analisis sentimen menggunakan Python, diharapkan pengguna internet menggunakan aplikasi ini dengan bijak sehingga suatu sentimen tidak lagi terjebak atau teriring oleh komentar-komentar yang merujuk ke dalam konteks yang merujuk pada penggiringan opini.

## REFERENSI

[1] Pemerintah Indonesia, "UNDANG-UNDANG DASAR NEGARA REPUBLIK INDONESIA 1945," 1945. Accessed: Nov. 05, 2023. [Online]. Available: https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=&cad=rja&uact=8&ved=2ahUKEwjEgYWMpqyCAxXd3TgGHbWAC2AQFnoECAgQAQ&url=https%3A%2F%2Fwww.mkri.id%2Fpublic%2Fcontent%2Finfoumum%2Fregulation%2Fpdf%2FUUD45%2520ASLI.pdf&usg=AOvVaw1lGtYko293LLZY_XpVoYei&opi=89978449

[2] M. Farhan Syam, L. Nur Hayati, and L. Syafie, "Klasifikasi Pemenuhan Pilar Sanitasi Puskesmas Menggunakan Metode Naive Bayes Classification of Fulfillment of Health Center Sanitation Pillars Using the Naive Bayes Method," *Jurnal Sistem Komputer*, vol. 12, no. 2, p. 2020, 2023.

[3] D. Darwis, N. Siskawati, and Z. Abidin, "Penerapan Algoritma Naive Bayes untuk Analisis Sentimen Review Data Twitter BMKG Nasional," vol. 15, no. 1.

[4] T. Muhammad Firdausy and P. Pandu Adikara, "Deteksi Iklan pada Twit menggunakan Metode Naive Bayes," 2022. [Online]. Available: http://j-ptiik.ub.ac.id

[5] E. Fitri, Y. Yuliani, S. Rosyida, and W. Gata, "Analisis Sentimen Terhadap Aplikasi Ruangguru Menggunakan Algoritma Naive Bayes, Random Forest Dan Support Vector Machine," *TRANSFORMTIKA*, vol. 18, no. 1, pp. 71–80, 2020, [Online]. Available: www.nusamandiri.ac.id,

[6] A. Tri Wijaya and A. Hermawan, "Analisis Sentimen Terhadap Dampak Inflasi di Indonesia Menggunakan Metode Multinomial Naïve Bayes Sentiment Analysis of the Impact of Inflation in Indonesia Using the Naïve Bayes Multinomial Method."

[7] R. Rasenda, H. Lubis, and R. Ridwan, "Implementasi K-NN Dalam Analisa Sentimen Riba Pada Bunga Bank Berdasarkan Data Twitter," *JURNAL MEDIA INFORMATIKA BUDIDARMA*, vol. 4, no. 2, p. 369, Apr. 2020, doi: 10.30865/mib.v4i2.2051.

[8] A. M. Priyatno and L. Ningsih, "SISTEMASI: Jurnal Sistem Informasi Pembobotan TF-IDF untuk Mendeteksi Akun Spammer di Twitter berdasarkan Tweet dan Representasi Retweet dari Tweet TF-IDF Weighting to Detect Spammer Accounts on Twitter based on Tweets and Retweet Representation of Tweets." [Online]. Available: http://sistemasi.ftik.unisi.ac.id

[9] F. S. Pamungkas and I. Kharisudin, "Analisis Sentimen dengan SVM," vol. 4, pp. 628–634, 2021, [Online]. Available: https://journal.unnes.ac.id/sju/index.php/prisma/

[10] R. Adinugroho, "PERBANDINGAN RASIO SPLIT DATA TRAINING DAN DATA TESTING MENGGUNAKAN METODE LSTM DALAM MEMPREDIKSI HARGA INDEKS SAHAM ASIA."

[11] U. Negeri *et al.*, "Komparasi Algoritma Random Forest, Naïve Bayes, dan Bert Untuk Multi-Class Classification Pada Artikel Cable News Network (CNN) Nanang Husin," 2023.

[12] Erfina A and Lestari Ajeng R., "SISTEMASI: Jurnal Sistem Informasi Analisis Sentimen terhadap Kendaraan Listrik menggunakan Algoritma Naïve Bayes Sentiment Analysis of Electric Vehicles using the Naïve Bayes Algorithm." [Online]. Available: http://sistemasi.ftik.unisi.ac.id

[13] I. A. Mashudi, S. N. Arief, D. Sandhya, T. Fatmawati, M. Hani'ah, and I. T. Alfarid, "JURNAL MEDIA INFORMATIKA BUDIDARMA Klasterisasi Jawaban Uraian Mahasiswa Menggunakan TF-IDF dan K-Means untuk Membantu Koreksi Ujian," vol. 7, pp. 2159–2167, 2023, doi: 10.30865/mib.v7i4.6688.

[14] Affandy Fahrizain, "Bag of Words vs TF-IDF — Penjelasan dan Perbedaannya," Medium - Data Folks Indonesia.

[15] Merinda Lestandy, Abdurrahim Abdurrahim, and Lailis Syafa'ah, "Analisis Sentimen Tweet Vaksin COVID-19 Menggunakan Recurrent Neural Network dan Naïve Bayes," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 5, no. 4, pp. 802–808, Aug. 2021, doi: 10.29207/resti.v5i4.3308.