UNP JOURNAL OF STATISTICS AND DATA SCIENCE
Volume 2 Nomor 1 Tahun 2024 Halaman 99-104.
ISSN(*Print*) : 3025-5511
ISSN(*Online*): 2985-475X

Twitter Data Sentimen Analysis for 2024 Presidential Candidate using Algorithm Naïve Bayes Classifier by K-Fold Cross Validation Methods

Aldi Prajela, Syafriandi*, Dony Permana, Dina Fitria
Departemen Statistika, Universitas Negeri Padang, Padang, Indonesia
*Corresponding author: syafriandi_math@fmipa.unp.ac.id

Submitted : 25 Januari 2024
Revised : 15 Februari 2024
Accepted : 23 Februari 2024

ABSTRACT

Indonesia implements a democratic system by involving the public in general elections for specific political positions. The active community expresses opinions on social media, especially regarding the 2024 Presidential Election and respective presidential candidates, which have become trending topics on Twitter. The analysis used to absorb these tweets into information is sentimen analysis using the Naïve Bayes Classifier (NBC) algorithm with the K-fold cross-validation method. Through the stages of pre-processing, weighting, labeling, classification using NBC, and testing using a confusion matrix, the results of the classification from NBC showed that Anies got 80% positive tweets and 20% negative tweets from 1186 tweets, Prabowo Subianto got 78% positive tweets and 22% negative tweets from 1149 tweets, and Ganjar Pranowo got 77% positive tweets and 23% negative tweets from 1075 tweets. The testing process was carried out using the NBC algorithm with the K-Fold Cross Validation method using values k=1 to k=10. The function of K-fold cross validation is to maximize the confusion matrix result. It can be concluded that Anies Baswedan has the highest score in iteration 4, namely a precision value of 90%, a recall value of 99%, and an accuracy value of 91%. Further, Ganjar Pranowo had the highest score in iteration 9, namely a precision value of 95%, a recall value of 97%, and an accuracy value of 92%. Meanwhile, Prabowo Subianto had the highest score in iteration 9, namely a precision value of 97%, a recall value of 99%, and an accuracy value of 94%.

Keywords: *K-fold Cross Validation, NBC, Presidential Candidates, Sentiment Analysis, Twitter.*

I. PENDAHULUAN

Negara Indonesia menerapkan sistem demokrasi, yang membuat keterlibatan warga negara dalam proses pemilihan untuk menduduki jabatan-jabatan politik tertentu menjadi suatu hal yang penting. Proses tersebut dikenal sebagai Pemilihan Umum (Pemilu). Salah satu bentuk pemilu adalah pemilihan Presiden (Pilpres) dimana Pilpres 2024 adalah sebuah proses demokrasi untuk memilih Presiden Republik Indonesia periode 2024–2029 pada tanggal 14 Februari 2024. Pemilihan ini merupakan Pilpres yang ke-8 di Indonesia.

Memasuki era sebelum Pilpres, masyarakat turut aktif menyampaikan opininya mengenai berbagai hal yang berkaitan dengan Pilpres. Salah satu topik yang tidak luput dari pembahasan masyarakat adalah opini tentang setiap calon yang mendaftar pada Pilpres 2024. Masyarakat juga menjadi sasaran elit politik, dimana suara mereka merupakan penentu keberlangsungan arah politik untuk lima tahun kedepan. Opini-opini positif serta negatif yang disampaikan oleh masyarakat pada sosial media sangat beragam dan tergantung pada berbagai faktor, termasuk latar belakang politik, keyakinan, dan pengalaman pribadi. Pada media sosial twitter, topik mengenai Pilpres dan masing-masing calonnya sering menjadi *trending*, ini menunjukkan bahwa banyak masyarakat mengungkapkan pendapat mereka di sosial media Twitter. Data opini seperti *tweet* yang disampaikan oleh masyarakat tersebut dapat diserap menjadi sebuah informasi apabila dilakukan analisis untuk memilah apakah opini tersebut bersifat positif ataupun negatif. Analisis yang digunakan untuk menyerap *tweet* tersebut menjadi sebuah informsasi adalah analisis sentimen. Analisis sentimen terhadap calon presiden Republik Indonesia periode 2024-2029 penting karena membantu menentukan popularitas dan kemungkinan sukses kandidat dalam pemilihan umum. Informasi sentimen mengenai ketiga calon tersebut berguna untuk memahami persepsi dan keinginan masyarakat terhadap mereka. Dengan analisis sentimen, pihak-pihak yang berkaitan dapat memprediksi *trend* dan membuat strategi yang lebih efektif untuk mendukung calon yang paling cocok untuk posisi presiden.

DOI: https://doi.org/10.24036/ujsds/vol2-iss1/149

ISSN(*Print*) : 3025-5511
ISSN(*Online*): 2985-475X

Menurut Sabily (2019), analisis sentimen adalah sebuah metode yang mengkaji pendapat, opini, evaluasi, sikap, atau penilaian seseorang terhadap individu. Berdasarkan penelitian Mahbubah (2019), metode klasifikasi dokumen yang sangat populer adalah *Naïve Bayes Classifier*. Dalam *Naïve Bayes Classifier* masih terdapat kemungkinan terjadinya *Overfitting* sehinga bisa merusak kinerja model dan mengurangi hasil akurasi dari sebuah model. Untuk mengatasi kemungkinan *overfitting* pada klasifikasi Naïve Bayes dapat dilakukan dengan *cross validation.*

Menurut Raschka (2018), *cross validation* berfungsi membagi data menjadi dua bagian yaitu data *training* dan data *testing* secara berulang, sehingga setiap data memiliki kesempatan yang sama untuk dibagi kembali menjadi data *training* dan *testing*. Raschka (2018) menyatakan bahwa salah satu teknik *cross validation* adalah *K-fold cross validation. K-fold cross validation* membagi data secara acak menjadi k kelompok. Selanjutnya masing-masing kelompok dibagi menjadi data latih data dan uji. Proses ini diulang sebanyak k kali, dengan menjadikan satu kelompok sebagai data uji pada setiap iterasi.

Penelitian yang berkaitan dengan analisis sentimen ini telah dilakukan oleh Allif & Firman (2023), Pada penelitian tersebut menggunakan algoritma *Naive Bayes Classifier* (NBC) dalam mengklasifikasikan sentimen masyarakat mengenai Pilpres 2024 menggunakan Rapid Miner. Dimana ada empat capres yang diteliti yaitu, Anies memperoleh 74% tanggapan positif dan 26% tanggapan negatif, diikuti oleh Sandi dengan 57% tanggapan positif dan 43% tanggapan negatif. Ganjar mendapatkan 53% tanggapan positif dan 47% tanggapan negatif, sementara Prabowo hanya memperoleh 32% tanggapan positif dan 68% tanggapan negatif.

II. METODE PENELITIAN

Penelitian ini menggunakan data sekunder, yaitu sekelompok *tweet* yang diperoleh dari pengguna Twitter di Indonesia, data diambil dari tanggal 01 Oktober hingga 30 November 2023, *keyword* yang digunakan "Anies Baswedan", "Ganjar Pranowo" dan "Prabowo Subianto" terkumpul sebanyak 1600 *tweet* untuk masing-masing *keyword.*

Tahapan-tahapan yang dilakukan dalam menganalisis data adalah sebagai berikut:

1. Melakukan *pre-proccesing* data yaitu, mengolah data mentah untuk menghasilkan data yang lebih mudah dipahami. dan memiliki informasi untuk di analisis. Dalam *pre-proccesing* data memiliki beberapa tahapan yaitu, *cleansing data, case folding, tokenizing, stopword removal, normalizing, dan stemming* (Wahyuni, 2020).
2. Melakukan tahapan TF-IDF merupakan metode untuk melihat hubungan suatu kata dengan sebuah dokumen dengan cara dibobotkan. Rumus TF-IDF dapat dilihat pada persamaan 1 (Irham & Wisesty, 2019).

$$w_{ij} = t_{i,j} \times d \quad (1)$$

$$d = log\left(\frac{N}{f_{i,j}}\right) \quad (2)$$

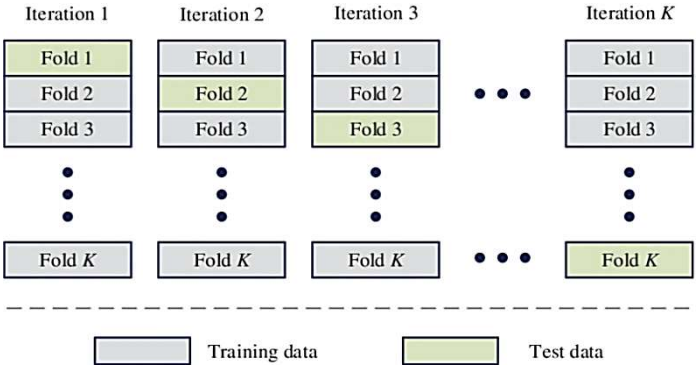
Keterangan :

w_{ij} : Bobot dari kata I pada dokumen ke-j
N : Total seluruh dokumen pada $t_{i,j}$
$t_{i,j}$: Total kemunculan kata ke-i pada dokumen ke-j
$f_{i,j}$: Total kemunculan kata ke-i pada dokumen ke-i
d_i : Total dokumen pada setiap kata ke-i
$n_{i,j}$: Banyak kata ke-I dalam dokumen ke-j

3. Mengunakan metode *Lexicon Based* dalam pelabelan data yang melibatkan penilaian bobot positif atau negatif pada setiap kata berdasarkan kamus *lexicon*. Pada penelitian ini kamus *lexicon* yang digunakan berasal dari *library VaderSentimen* yang terdapat dalam *Google Colab.*
4. Melakukan klasifikasi menggunakan algoritma *Naïve Bayes Classifier*
   Sebelum melakukan klasifikasi dengan menggunakan algoritma NBC, terlebih dahulu data dibagi menjadi dua kelompok yaitu data *training* dan data *testing.* Data *training* digunakan untuk membangun model, sedangkan data *training* untuk digunakan untuk menguji model. Untuk membagi data menjadi dua bagian digunakan metode *k-fold cross validation* dengan nilai *k=10.* Gambar 1 memberikan ilustrasi visual dari *metode k-fold cross-validation*, Refaeilzadeh (2016).

DOI: https://doi.org/10.24036/ujsds/vol2-iss1/149

ISSN(*Print*) : 3025-5511
ISSN(*Online*): 2985-475X



Gambar 1. Ilustrasi Metode *K-fold Cross Validation*

Setelah diperoleh data *training* dan data *testing*, dilakukan pengklasifikasian sentimen menggunakan metode NBC. NBC adalah salah satu algoritma klasifikasi yang menggunakan konsep probabilitas untuk mengklasifikasikan data ke dalam kategori yang paling tepat. NBC dikenal sebagai metode yang relatif sederhana, tetapi pada saat yang sama dapat memberikan hasil klasifikasi yang cukup baik. Pada saat klasifikasi menggunakan algoritma *Naive Bayes*, probabilitas tertinggi dari semua kategori/kelas dokumen yang diuji (dihitung dengan menggunakan rumus Teorema Bayes (Rodiansyah, 2013).

$$P(v_j) = \frac{doc_j}{training} \tag{3}$$

$$P(a_i|v_j) = \frac{n_i + 1}{n + kosakata} \tag{4}$$

Adapun persamaan V_{MAP} adalah sebagain berikut.

$$V_{MAP} = {}^{arg\,max}_{v_j \epsilon v} P(V_j) \times \prod_i P(a_i|v_j) \tag{5}$$

Keterangan :

V_{MAP} : Semua kategori/kelas untuk data *test*
$P(a_i|v_j)$: Probabilitas a_i pada kategori
$P(v_j)$: Probabilitas v_j
doc_j : Total dokumen pada kelas ke-I dalam *training*
$training$: Total dokumen pada data *training*
$kosakata$: Banyaknya kata dalam *training*

5. Melakukan evaluasi model menggunakan *confusion matrix. Confusion matrix* adalah salah satu teknik yang dapat dipakai untuk menilai efektivitas suatu metode klasifikasi. Pada prinsipnya, *confusion matrix* berisi data perbandingan antara *output* klasifikasi yang dihasilkan oleh sistem dengan hasil klasifikasi yang seharusnya. Menurut Han & Kember (2006), Dalam pengukuran kinerja menggunakan c*onfusion matrix*, terdapat empat istilah yang digunakan untuk merepresentasikan hasil proses klasifikasi, keempat istilah tersebut adalah *True Positive* (TP), *True Negative* (TN), *False Positive* (FP), *False Negative* (FN). Pada *confusion matrix* terdapat beberapa nilai yang dihitung, seperti akurasi, presisi, dan *recall.* Perhitungan dan tabel perhitungan akurasi menggunakan *confusion matrix* dapat dilihat pada tabel dibawah ini.

Tabel 1. *Confusuin matrix*

Predicted Class	Actual Class		
		Yes	No
	Yes	TP (True Positive)	FP (False Positive)
	No	FN (False Negative)	TN (True Negative

DOI: https://doi.org/10.24036/ujsds/vol2-iss1/149

ISSN(*Print*) : 3025-5511
ISSN(*Online*): 2985-475X

$$Akurasi = \frac{TN + TP}{TN + TP + FN + FP} \times 100\% \quad (6)$$

$$Presisi = \frac{TP}{TP + FP} \times 100\% \quad (7)$$

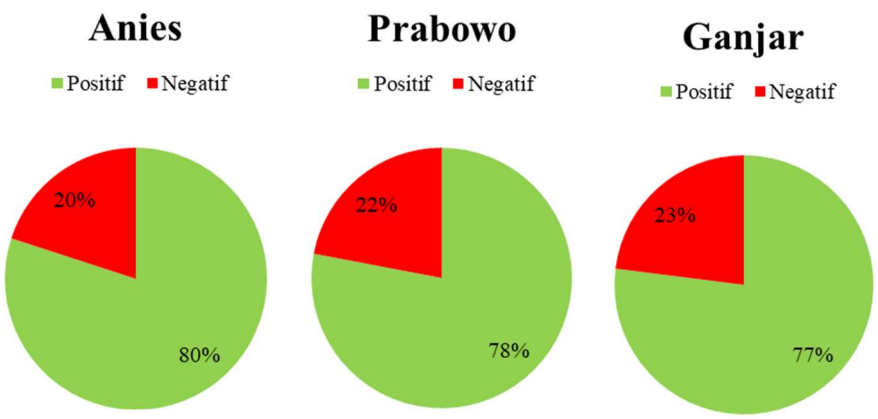
$$Recall_{positif} = \frac{TN}{TP + FN} \times 100\% \quad (8)$$

Akurasi adalah seberapa tepat suatu model dalam memprediksi nilai dengan membandingkan data yang diklasifikasikan secara benar terhadap keseluruhan dataset. Selanjutnya presisi merupakan rasio antara jumlah data teks yang diklasifikasikan dengan benar sebagai positif dan dibagi oleh keseluruhan data yang diklasifikasikan sebagai positif. Recall, atau sering disebut sebagai sensitivitas, merupakan proporsi dari data dalam kategori positif yang berhasil terdeteksi dengan tepat oleh system, fungsinya adalah untuk mengukur seberapa efektif sistem dalam mengambil kembali informasi yang relevan. Menurut Hilmiyah (2017), untuk memaksimalkan nilai *confusion matrix* dapat menggunakan *k-folds cross validation* dengan nilai k = 10, dan nilai yang maksimal dilihat dari nilai *akurasi*, *presisi* dan *recall* tertinggi yang dihasilkan oleh setiap iterasi *k-fold cross validation*

.

III. HASIL DAN PEMBAHASAN

Hasil pemodelan klasifikasi sentimen menggunakan algoritma *Naïve Bayes Classifier* pada masing-masing *keyword* disajikan dalam Gambar 2.



Gambar 2. Diagram Hasil Prediksi Model NBC

Setelah dilakukan klasifikasi menggunakan algoritma NBC didapatkan bahwa hasil klasifikasi sentimen masyarakat pada media sosial twitter terhadap Capres 2024 yaitu, Anies Baswedan mendapatkan perolehan sentimen positif tertinggi yaitu sebesar 80% dan 20% sentimen negatif dari 1186 *tweet*, disusul dengan Prabowo Subianto yang mendapatkan sentimen positif sebesar 78% dan 22% sentimen negatif dari 1149 *tweet*, sedangkan Ganjar Pranowo mendapatkan 77% sentimen positif dan 23% sentimen negatif dari 1075 *tweet*. Hal ini menandakan banyaknya pengguna media sosial twitter yang menyuarakan opini bersifat positif terhadap masing-masing Capres Indoensia 2024, dimana Anies Baswedan memperoleh sentimen positif tertinggi, sedangkan sentimen negatif tertinggi diperoleh oleh Ganjar Pranowo.

Performa dari model yang dibangun dapat dilihat dari nilai *precision*, *recall*, dan *accuracy* yang dihitung dari *confusion matrix.* Berikut hasil perhitungan pada pengujian data twitter capres 2024 menggunakan *confusion matrix* dapat dilihat pada Tabel 7.

DOI: https://doi.org/10.24036/ujsds/vol2-iss1/149

ISSN(*Print*) : 3025-5511
ISSN(*Online*): 2985-475X

Table 7. Hasil Evaluasi Model

Iteration	Anies			Ganjar Pranowo			Prabowo Subiato		
	precision	*recall*	*accuracy*	*precision*	*recall*	*accuracy*	*precision*	*recall*	*accuracy*
1	88%	95%	86%	92%	94%	87%	96%	93%	91%
2	83%	93%	78%	91%	95%	86%	89%	93%	85%
3	84%	99%	85%	93%	94%	92%	97%	93%	92%
4	90%	99%	91%	92%	95%	88%	94%	95%	91%
5	88%	98%	88%	92%	96%	88%	93%	91%	87%
6	84%	98%	84%	94%	97%	91%	95%	97%	93%
7	88%	96%	86%	92%	93%	86%	96%	95%	92%
8	89%	93%	85%	93%	96%	90%	95%	93%	90%
9	91%	94%	88%	95%	97%	92%	97%	99%	94%
10	86%	91%	82%	92%	98%	90%	93%	94%	89%

Tabel 7 menunjukkan hasil pengujian dengan menggunakan *confusion matrix*, perulangan yang dilakukan sebanyak 10. Jika diambil nilai tertinggi dari *confusion matrix* untuk setiap perulangan dari tabel 7, terlihat bahwa Anies memiliki nilai tertinggi pada perulangan 4 dengan ketepatan model dalam memprediksi kelas sebesar 90%, proporsi data yang tepat terprediksi dalam suatu kelas sebesar 99%, serta ketepatan model dalam memprediksi data secara keseluruhan sebesar 91%. Selanjutnya Ganjar Pranowo memiliki nilai tertinggi pada *iteration* 9 dengan ketepatan model dalam memprediksi kelas sebesar 95%, dan ketepatan model dalam memprediksi data secara keseluruhan sebesar 97%, serta ketepatan model dalam memprediksi data secara keseluruhan sebesar 92%. Sedangkan Prabowo Subianto memiliki nilai tertinggi pada *iteration* 9 dengan ketepatan model dalam memprediksi kelas sebesar 97%, dan ketepatan model dalam memprediksi data secara keseluruhan sebesar 99%, serta ketepatan model dalam memprediksi data secara keseluruhan sebesar 92% 94%.

IV. KESIMPULAN

Pada keseluruhan dataset yang melibatkan tiga *keyword* terlihat sentimen masyarakat Indonesia terhadap masing-masing calon presiden, Setelah dilakukan klasifikasi menggunakan NBC dapat disimpulkan bahwa Anies mendapatkan 80% *tweet* positif dan 20% *tweet* negatif dari 1186 *tweet*, Prabowo Subianto mendapatkan 78% *tweet* positif dan 22% *tweet* negatif dari 1149 *tweet*, dan Ganjar Pranowo mendapatkan 77% *tweet* positif dan 23% *tweet* negatif dari 1075 *tweet.* Terlihat bahwa kelompok data dengan *keywrod* Anies memporoleh sentimen positif tertinggi sebanyak 80%, sedangkan kelompok data dengan *keyword* negatif tertinggi diperoleh oleh Ganjar sebanyak 23%.

Pengujian menggunakan NBC memiliki kemampuan dalam mengklasifikasikan data Twitter calon presiden 2024 dengan menggunakan metode *k-fold cross validation confusion* untuk memaksimalkan hasil *confusion matrix,* dapat disimpulkan bahwa Anies memiliki nilai tertinggi pada *iteration 4* yaitu nilai *precision* sebesar 90%, nilai *recall* sebesar 99%, dan nilai *acuracy* sebesar 91%. Selanjutnya Prabowo Subianto memiliki memiliki nilai tertinggi pada *iteration* 9 yaitu nilai *precision* sebesar 97%, nilai *recall* sebesar 99%, dan nilai *acuracy* sebesar 94%. Sedangkan Ganjar Pranowo memiliki nilai tertinggi pada *iteration* 9 yaitu nilai *precision* sebesar 95%, nilai *recall* sebesar 97%, dan nilai *acuracy* sebesar 92%.

DAFTAR PUSTAKA

Abdillah, A. R., & Hasan, F. N. (2023). Analisis Sentimen Terhadap Kandidat Calon Presiden Berdasarkan Tweets Di Sosial Media Menggunakan *Naive Bayes Classifier*. SMATIKA JURNAL: STIKI Informatika Jurnal, 13(01), 117-130.

Hilmiyah, F. A. T. H. I. N. "Prediksi Kinerja Mahasiswa Menggunakan Support Vector Machine untuk Pengelola Program Studi di Perguruan Tinggi (Studi Kasus: Program Studi Magister Statistika ITS)." Departemen Manajemen Teknologi Bidang Keahlian Manajemen Teknologi Informasi Fakultas Bisnis Dan Manajemen Teknologi Institut Teknologi Sepuluh Nopember Surabaya (2017): 1-99.

DOI: https://doi.org/10.24036/ujsds/vol2-iss1/149

ISSN(*Print*) : 3025-5511
ISSN(*Online*): 2985-475X

Mahbubah, L. D., & Zuliarso, E. 2019. *Analisa Sentimen Twitter Pada Pilpres 2019 Menggunakan Algoritma Naive Bayes*. *Prosiding SINTAK 2019,ISBN*: 978-602-8557-20-7,(p) : 193- 199, 2019.

Raschka, S. 2018. MLxtend: Providing machine learning and data science utilities and extensions to Python's scientific computing stack. *Journal of open source software*, *3*(24), 638.

Refaeilzadeh, P., Tang, L. & Liu, H. 2016. *Cross Validation. Encyclopedia of Database Systems*. New York: Springer. DOI 10.1007/978-1-4899-7993- 3_565-2.

Rodiyansyah, S. F., & Winarko, E. (2012). Klasifikasi Posting Twitter Kemacetan Lalu Lintas Kota Bandung Menggunakan Naive Bayesian Classification. *IJCCS (Indonesian Journal of Computing and Cybernetics Systems)*.

Sabily, A. F., Adikara, P. P., & Fauzi, M. A. 2019. *q*. *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, *3*(5), 4204-4209.

Wahyuni, E. D., Arifiyanti, A. A. & Afandi, I. M., 2020. *Klasifikasi Teks Dengan Python*. Sidoarjo: Indomedika Pustaka.

DOI: https://doi.org/10.24036/ujsds/vol2-iss1/149