

ANALISIS SENTIMEN PARA KANDIDAT PILPRES 2024 DENGAN MODEL BAHASA BERT

Luis Ricardo Pandiangan^{1§}, IGN Lanang Wijayakusuma²

¹Program Studi Matematika, Universitas Udayana [Email: luisrichardo777@gmail.com]

²Program Studi Matematika, Universitas Udayana [Email: lanang_wijaya@unud.ac.id]

[§]Corresponding Author

ABSTRACT

Indonesia has conducted a general election for presidential and legislative positions for central and regional government on 14 February 2024. With the advancement of technology, each candidate has utilized social media to engage with the voters. One of the most common platforms for people to have political discussions is X. This study aims to perform sentiment analysis better to understand people's perceptions towards each 2024 presidential candidate. Using BERT as the primary model, this research classifies sentiment toward each candidate separately with only two labels: positive and negative. The metrics used in this study are accuracy and F1-score with astonishing results, 85% and 89% for Anies's dataset, 88% and 89% for Prabowo's dataset, and 89% for both metrics for Ganjar's dataset. This research also generated keywords for each sentiment towards candidates from BERT's validation data predictions. This study contributes insights into the narratives and sentiments surrounding the 2024 presidential election.

Keywords: BERT, pilpres, sentiment analysis

1. PENDAHULUAN

Indonesia telah melakukan pemilu presiden, DPR, DPD, dan DPRD pada 14 Februari 2024. Pesta demokrasi tersebut memiliki jumlah pemilih sebesar 204,8 juta menurut Komisi Pemilihan Umum. Pemerintah harus mengetahui kebutuhan masyarakat dan persepsinya pada musim politik kali ini agar proses transisi pemerintahan dapat berjalan dengan baik. Tentunya, para politisi yang terlibat harus mengetahui peta tempat mereka bermain agar bisa meraih suara yang diinginkan dengan cara yang relevan dengan keadaan.

Setiap hari manusia telah menggunakan media sosial untuk memperbarui perkembangan, berkomunikasi dengan orang lain, serta berinteraksi dalam berbagai diskusi. Kecepatan dan kemudahan akses media sosial telah membuat kita mampu mengetahui berbagai peristiwa dan informasi dari berbagai belahan dunia. Selain itu, berbagai platform media sosial telah membentuk pola pikir, kebiasaan, dan cara kita mengambil keputusan mulai dari ranah pribadi, ekonomi, politik, hingga budaya (Belcastro et al., 2022).

Persepsi masyarakat mampu dibentuk oleh masifnya informasi yang di media sosial tidak terbatas pada satu isu saja. Dinamika ini juga

telah mengubah proses berpolitik di era kontemporer. Pemilih mendapatkan informasi mengenai para kandidat melalui media sosial yang telah memiliki algoritmanya masing-masing. Ini membuat para kandidat harus beradaptasi agar bisa melakukan pendekatan pada media baru dengan segala karakteristiknya. Para politisi harus menyadari berbagai tantangan seperti pola interaksi agar bisa memposisikan diri dengan tepat (Noorikhshan et al., 2023).

X adalah salah media sosial yang sangat populer digunakan. Ada jutaan pesan atau *postingan* yang diunggah di X setiap detiknya dari berbagai pengguna. Ia memiliki fitur *tweet* yang berisi teks hingga 280 karakter serta dapat dilengkapi dengan gambar, video, tautan, atau polling. Pengguna lain juga bisa merespon, membagikan (*retweet*), atau menyukai *tweet* yang membuat interaksi di dalam aplikasi menjadi dinamis. Fitur tersebut membuat X dikenal sebagai media sosial *microblogging*. Data yang banyak dihasilkan oleh X menjadikan platform ini menjadi media sosial paling populer untuk diteliti dengan *machine learning* (Aljabri et al., 2023).

Analisis sentimen dapat berguna untuk mempelajari persepsi masyarakat (Phan et al., 2020). Terkhusus untuk ranah politik, analisis sentimen mampu memberikan pemahaman mengenai opini atau sikap terhadap politisi, kebijakan pemerintah, rancangan undang-undang, kampanye, hingga kejadian politik tertentu. Media sosial menjadi tempat yang cocok untuk studi ini dikarenakan banyaknya pengguna dari Indonesia.

Ada banyak metode yang telah digunakan untuk analisis sentimen salah satunya ialah Naïve Bayes. Ini adalah algoritma probabilistik yang sering dipakai pada masalah klasifikasi dengan hasil yang baik walaupun simpel dan intuitif menjadikannya cocok untuk dataset yang kecil. Model ini juga tidak terlalu rentan akan *overfitting* dan hanya memerlukan daya komputasi yang rendah. Hal-hal tersebut membuat Naïve Bayes sering digunakan dalam analisis sentimen dibandingkan model lainnya (Huang et al., 2023; Tan et al., 2023).

Yusrizal & Sasongko, (2024) melakukan penelitian pada presiden dan wakil presiden terpilih, Prabowo-Gibran, dengan menggunakan Naïve Bayes dari data X sebanyak 1520 tweet. Penelitian tersebut menyimpulkan bahwa masih banyak sentimen negatif terhadap pasangan presiden baru itu.

Seiring berkembangnya teknologi, algoritma *deep learning* semakin canggih dalam tugas pengenalan teks. Pada tahun 2017 Google menerbitkan artikel “*Attention is All You Need*” untuk mengenalkan arsitektur Transformers yang memberikan pendekatan berbeda dalam memprediksi kata dari model RNN atau LSTM yang bertumpu pada pemrosesan data deret. Model itu mampu mengenali konteks dengan *self-attention mechanism*, *positional encoding*, dan *parallelization*. Arsitektur ini memunculkan berbagai model pengenalan bahasa besar seperti BERT, GPT, dan T5 (Vaswani et al., 2017).

BERT (*Bidirectional Encoder Representation from Transformer*) pertama kali dikenalkan oleh (Devlin et al., 2019). Model ini mampu membaca teks dari kedua arah (depan dan belakang) untuk memahami makna dan konteks dalam kalimat penuh. Dengan demikian, BERT mampu digunakan untuk berbagai tugas bahasa termasuk analisis sentimen. Model ini menjadi memiliki keunggulan untuk menangkap hubungan konteks kata secara mendalam sehingga memberikan performa yang lebih akurat dibandingkan model bahasa terdahulu seperti

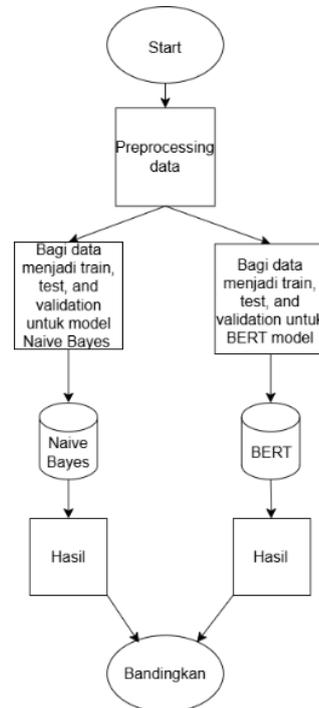
Naïve Bayes (Fernández-Martínez et al., 2022).

Zardak et al., (2023) melakukan studi analisis sentimen pada teks dokumen berbahasa Farsi dengan menggunakan ParsBERT. Ring et al., (2024) menggunakan BERT untuk menganalisis sentimen berita politik berbahasa Hungaria. Putri et al., (2023) melakukan penelitian analisis sentimen pilpres 2014 dan 2019 menggunakan BERT dan DistilBERT.

Artikel ini membahas mengenai analisis sentimen pilpres 2024 kepada masing-masing kandidat secara terpisah. Model yang digunakan adalah BERT dan memberikan *wordclouds* kata-kata kunci sentimen dari data hasil prediksi model BERT. Model Naïve Bayes digunakan sebagai *baseline* pembanding.

2. METODE PENELITIAN

Proses penelitian adalah sebagai berikut:



Gambar 1. Alur Penelitian

2.1 Dataset

Penelitian ini menggunakan data sekunder yang dapat diakses di Kaggle yaitu *Indonesia Presidential Candidate's Dataset, 2024*. Kontributor data ini adalah Firdaus, (2023). *Dataset* ini terdiri dari 3 folder yaitu *original data*, *cleaned data*, dan *labeled data*. Tiap folder berisi file CSV dari masing-masing kandidat. Total ada 30,000 *tweets* yang dikumpulkan dalam data tersebut. Data untuk Ganjar Pranowo diambil pada Oktober 2022 – April 2023,

Prabowo Subianto pada Desember 2022 – April 2023, dan Anies Baswedan pada Januari 2023 – April 2023. Kontributor data tidak menjelaskan alasan perbedaan waktu pengambilan data. Perbedaan waktu pengambilan data dapat memengaruhi sentimen yang tercatat dalam mencerminkan dinamika politik yang berbeda. Misalnya, muncul survei elektabilitas yang membuat kandidat Prabowo Subianto melesat sepanjang 2023 yang sebelum selalu didominasi oleh Ganjar Pranowo menurut berita oleh Suryana & Akbar, (2023). Meskipun terdapat perbedaan waktu pengambilan, penulis tetap menggunakan *dataset* ini karena masih memberikan gambaran dinamika politik yang berkembang sepanjang periode tersebut. Selain itu, perbedaan periode waktu dapat memberikan variabilitas perubahan sentimen seiring perkembangan masa kampanye. Label yang digunakan pada data hanya positif dan negatif. *Dataset* ini sudah diterjemahkan ke dalam bahasa Inggris oleh kontributor.

Tabel 1. Contoh *Dataset*

<i>Tweets</i>	Label
mr. anies continued, we will guard him until he becomes president	Positif
it must be continued by the next president, therefore, don't vote for anies. jakarta has become a victim of anies' stupidity	Negatif

2.2 Preprocessing

Tahap ini dilakukan untuk menyederhanakan data teks agar model dapat mengidentifikasi sentimen dengan lebih akurat. Proses ini dimulai dengan *case folding* semua teks menjadi huruf kecil, penghapusan karakter khusus, penghapusan tagar, penghapusan mention, penghapusan data duplikat, penghapusan *multiple spaces*, dan lematisasi.

Setelah itu, data kemudian dibagi menjadi 2 untuk dilakukan penelitian dengan model yang berbeda. Hal ini juga dilakukan karena proses tokenisasi yang berbeda pada kedua model tersebut. Tokenisasi Naïve Bayes mengandalkan *Term Frequency-Inverse Document Frequency* (TF-IDF) yang mengakibatkan setiap kata diperlakukan sebagai satu unit independen. Di sisi lain, BERT menggunakan tokenisasi

berbasis *WordPiece* yang membagi kata menjadi sub-kata jika kata tersebut tidak ditemukan secara keseluruhan dalam kosakata model. Proses ini memungkinkan memahami konteks kata lebih penuh (Duong & Nguyen-Thi, 2021; Pota et al., 2021; Tan et al., 2023)

Tabel 2. Contoh Hasil *Preprocessing*

<i>Tweets</i>	Label
mr anies continue we will guard him until he become president	Positif
it must be continue by the next president therefore dont vote for anies jakarta has become a victim of anies stupidity	Negatif

2.3 Naïve Bayes

Penggunaan model Naïve Bayes pada penelitian ini hanya berfungsi sebagai *baseline* dengan penggunaan model BERT. Model ini memprediksi kelas data berdasarkan probabilitas fiturnya yaitu frekuensi jumlah kata yang dimiliki tiap kelasnya (Suppala & Rao, 2020). Berikut adalah persamaan teorema Bayes yang menjadi dasar model ini:

$$P(X|Y) = \frac{P(X|Y) \cdot P(X)}{P(Y)} \quad (1)$$

$P(X)$ adalah probabilitas awal bahwa hipotesis X benar (probabilitas prior), misalnya suatu teks memiliki sentimen positif tanpa melihat kata-katanya. $P(Y)$ adalah probabilitas dari bukti Y, yaitu probabilitas munculnya kata-kata dalam teks tanpa dikaitkan dengan sentimen tertentu. $P(X|Y)$ adalah probabilitas sentimen pada tertentu pada teks setelah mengetahui kata-kata yang muncul (probabilitas posterior). $P(Y|X)$ adalah probabilitas kata-kata muncul jika tahu sentimennya (probabilitas *likelihood*).

Pada analisis sentimen, model menerapkan probabilitas awal untuk setiap kelas sentimen pada data (probabilitas prior). Setelah itu, model menghitung probabilitas kemunculan setiap kata dalam teks untuk semua kelas sentimen (*likelihood*) lalu model akan menghitung probabilitas total teks bersama sentimen tertentu dengan menggabungkan probabilitas prior dan probabilitas *likelihood* setiap kata dalam teks (Mustakim & Priyanta, 2022; Huang et al., 2023).

2.4 BERT

Pre-training pada model ini melibatkan dua tugas utama yaitu *masked language model* (MLM) dan *next sentence prediction* (NSP). Pada bagian MLM, model mengacak sebagian kata dalam teks dengan cara “disembunyikan” dengan memberikan token [MASK]. Model kemudian akan mengisi bagian tersebut berdasarkan konteks kata-kata lainnya. Pada bagian NSP, model memprediksi hubungan logis antar dua kalimat berturut-turut supaya memahami konteks antar kalimat. Model ini juga memberikan token-token khusus yaitu [CLS] pada awal *input* kalimat, [SEP] pada bagian yang memisahkan kalimat, dan [PAD] untuk menyamakan panjang *input* dalam batch. Posisi setiap kata tidak diperhatikan dalam BERT karena adanya *position embeddings* (Devlin et al., 2019).

BERT digunakan untuk mampu menangkap nuansa semantik dalam teks agar mampu memahami sentimen pada kalimat dengan cara melakukan *fine-tuning* menggunakan dataset dari Twitter. Setelah itu, gunakan arsitektur model BERT *uncased* klasifikasi agar bisa mengenali kelas sentimen dari data yang diberikan tanpa memedulikan kapitalisasi (Joloudari et al., 2023).

3. PEMBAHASAN

Kedua model dievaluasi menggunakan metrik akurasi (*accuracy*) dan nilai F1 (F1-score). Akurasi diketahui dengan persamaan berikut (Lai et al., 2024):

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (2)$$

di mana:

TP = *true positive*

TN = *true negative*

FP = *false positive*

FN = *false negative*.

Nilai F-1 menggunakan presisi dan *recall*. Menurut Lai et al., (2024), metrik ini mengukur keseimbangan antara presisi dan *recall* menggunakan rata-rata harmonik. Hal tersebut bertujuan untuk mengevaluasi performa model pada dataset dengan distribusi kelas yang tidak seimbang. Berikut adalah persamaan-persamaannya:

$$Precision = \frac{TP}{TP+FP} \quad (3)$$

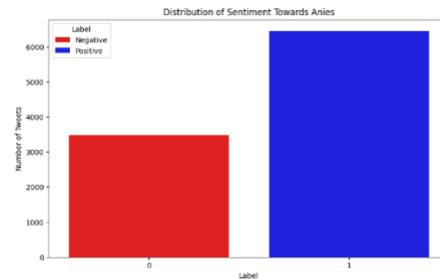
$$Recall = \frac{TP}{TP+FN} \quad (4)$$

$$F1 \text{ score} = 2 * \frac{Precision*Recall}{Precision+Recall} \quad (5)$$

Penelitian membagi dataset pada tiap kandidat sebanyak 60% *training data*, 20% *testing data*, dan 20% *validation data* pada model Naïve Bayes sedangkan model BERT mendapat pembagian data 75% *training data*, 15% *testing data*, dan 15% *validation data*.

Ketiga kandidat memiliki data yang tidak seimbang antar label. Menyeimbangkan jumlah data dilakukan agar mencapai akurasi yang terbaik. Penelitian ini menggunakan pendekatan *oversampling* pada model Naïve Bayes dan *undersampling* pada model BERT.

3.1 Hasil Pada Kandidat Anies Baswedan



Gambar 2. Label Sentimen Anies

Dataset Anies yang telah dibersihkan memiliki sebanyak 6454 data berlabel positif dan 3477 data berlabel negatif. Tabel 3 menunjukkan hasil dari analisis sentimen menggunakan Naïve Bayes dan BERT.

Tabel 3. Perbandingan Hasil pada Kandidat Anies Baswedan

Metrik	Naïve Bayes	BERT
<i>Precision</i>	0,76	0,89
<i>Recall</i>	0,88	0,82
F1-score	0,82	0,85
<i>Accuracy</i>	0,80	0,85

Dari tabel terlihat terjadi peningkatan akurasi dari 80% menggunakan Naïve Bayes menjadi 85% dengan BERT. Hasil ini selaras dengan nilai F1 yang lebih besar pada model BERT daripada hasil Naïve Bayes untuk kedua jenis label.

mendatang, maka penulis memberikan beberapa saran berikut:

1. Mengambil data secara langsung dari X menggunakan API dengan periode pengambilan data yang sama untuk semua kandidat di masa depan.
2. Menggunakan data *tweets* berbahasa Indonesia dan menggunakan model IndoBERT agar langsung memahami konteks dan keakuratan dengan lebih baik.
3. Membuat variasi analisis sentimen dari berbagai aplikasi media sosial yang lain seperti TikTok ataupun Instagram.

DAFTAR PUSTAKA

- Aljabri, M., Zagrouba, R., Shaahid, A., Alnasser, F., Saleh, A., & Alomari, D. M. (2023). Machine learning-based social media bot detection: a comprehensive literature review. *Social Network Analysis and Mining*, 13(1). <https://doi.org/10.1007/s13278-022-01020-5>
- Belcastro, L., Cantini, R., & Marozzo, F. (2022). Knowledge Discovery from Large Amounts of Social Media Data. *Applied Sciences (Switzerland)*, 12(3). <https://doi.org/10.3390/app12031209>
- Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. <http://arxiv.org/abs/1810.04805>
- Duong, H. T., & Nguyen-Thi, T. A. (2021). A review: preprocessing techniques and data augmentation for sentiment analysis. *Computational Social Networks*, 8(1). <https://doi.org/10.1186/s40649-020-00080-x>
- Fernández-Martínez, F., Luna-Jiménez, C., Kleinlein, R., Griol, D., Callejas, Z., & Montero, J. M. (2022). Fine-Tuning BERT Models for Intent Recognition Using a Frequency Cut-Off Strategy for Domain-Specific Vocabulary Extension. *Applied Sciences (Switzerland)*, 12(3). <https://doi.org/10.3390/app12031610>
- Firdaus, A. (2023). Indonesia Presidential Candidate's Dataset, 2024. In *Mendeley Data* (Issue 5). <https://doi.org/10.17632/7w5zvr8jgp.5>
- Huang, X., Jin, G., & Ruan, W. (2023). Naive Bayes. In *Machine Learning Safety. Artificial Intelligence: Foundations, Theory, and Algorithms* (pp. 95–102). Springer. https://doi.org/10.1007/978-981-19-6814-3_8
- Joloudari, J. H., Hussain, S., Nematollahi, M. A., Bagheri, R., Fazl, F., Alizadehsani, R., Lashgari, R., & Talukder, A. (2023). BERT-deep CNN: state of the art for sentiment analysis of COVID-19 tweets. *Social Network Analysis and Mining*, 13(1). <https://doi.org/10.1007/s13278-023-01102-y>
- Lai, J., Yang, X., Luo, W., Zhou, L., Li, L., Wang, Y., & Shi, X. (2024). RumorLLM: A Rumor Large Language Model-Based Fake-News-Detection Data-Augmentation Approach. *Applied Sciences (Switzerland)*, 14(8). <https://doi.org/10.3390/app14083532>
- Mustakim, H., & Priyanta, S. (2022). Aspect-Based Sentiment Analysis of KAI Access Reviews Using NBC and SVM. *IJCCS (Indonesian Journal of Computing and Cybernetics Systems)*, 16(2), 113. <https://doi.org/10.22146/ijccs.68903>
- Noorikhshan, F. F., Ramdhani, H., Sirait, B. C., & Khoerunisa, N. (2023). Dinamika Internet, Media Sosial, dan Politik di Era Kontemporer: Tinjauan Relasi Negara-Masyarakat. *Journal of Political Issues*, 5(1), 95–109. <https://doi.org/10.33019/jpi.v5i1.131>
- Phan, H. T., Tran, V. C., Nguyen, N. T., & Hwang, D. (2020). Improving the Performance of Sentiment Analysis of Tweets Containing Fuzzy Sentiment Using the Feature Ensemble Model. *IEEE Access*, 8, 14630–14641.
- Pota, M., Ventura, M., Fujita, H., & Esposito, M. (2021). Multilingual evaluation of pre-processing for BERT-based sentiment analysis of tweets. *Expert Systems with Applications*, 181. <https://doi.org/10.1016/j.eswa.2021.115119>
- Putri, M., Edy Sutanto, T., & Inna, S. (2023). Studi Empiris Model BERT dan DistilBERT: Analisis Sentimen pada Pemilihan Presiden Indonesia. *Indonesian*

- Journal of Computer Science Attribution*, 12(5), 2023–2972.
<https://doi.org/10.33022/ijcs.v12i5.3445>
- Ring, O., Szabó, M. K., Guba, C., Váradi, B., & Üveges, I. (2024). “Approaches to sentiment analysis of Hungarian political news at the sentence level.” *Language Resources and Evaluation*, 58, 1233–1261.
<https://doi.org/10.1007/s10579-023-09717-5>
- Suppala, K., & Rao, N. (2020). Sentiment analysis using naïve bayes classifier. *International Journal of Innovative Technology and Exploring Engineering*, 8(8), 264–269.
<https://doi.org/10.14445/22312803/ijctt-v68i4p141>
- Suryana, W., & Akbar, N. A. (2023, July 24). Elektabilitas Capres: Ganjar Unggul Sepanjang 2022, Prabowo Melesat pada 2023. *Republika*.
<https://news.republika.co.id/berita/ryalfh409/elektabilitas-capres-ganjar-unggul-sepanjang-2022-prabowo-melesat-pada-2023-part2>
- Tan, K. L., Lee, C. P., & Lim, K. M. (2023). A Survey of Sentiment Analysis: Approaches, Datasets, and Future Research. In *Applied Sciences (Switzerland)* (Vol. 13, Issue 7). MDPI.
<https://doi.org/10.3390/app13074550>
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). *Attention Is All You Need*.
<https://doi.org/10.48550/arXiv.1706.03762>
- Yusrizal, M., & Sasongko, T. B. (2024). Analisis Sentimen Masyarakat Terhadap Presiden dan Calon Presiden Terpilih 2024 Menggunakan Naïve Bayes. *JURNAL MEDIA INFORMATIKA BUDIDARMA*, 8(3), 1673.
<https://doi.org/10.30865/mib.v8i3.7882>
- Zardak, S. R., Rasekh, A. H., & Bashkari, M. S. (2023). Persian Text Sentiment Analysis Based on BERT and Neural Networks. *Iranian Journal of Science and Technology - Transactions of Electrical Engineering*, 47(4), 1623–1634.
<https://doi.org/10.1007/s40998-023-00626-5>