

Perbandingan Algoritma XGBoost dan SVM Dalam Analisis Opini Publik Pemilihan Presiden 2024**Dea Safitri¹, Susanti², Rahmaddeni³, Triyani Arita Fitri⁴**

safitridea77@gmail.com, susanti@usti.ac.id, rahmaddeni@usti.ac.id, triyani@usti.ac.id

Universitas Sains dan Teknologi Indonesia

Informasi Artikel	Abstrak
<p>Diterima : 18 Mei 2024 Direview : 14 Jun 2024 Disetujui : 30 Jun 2024</p>	<p>Pemilihan presiden dipengaruhi oleh berbagai faktor, termasuk latar belakang kandidat, isu politik, dan preferensi ideologi, sehingga pemilihan presiden menjadi subjek klasifikasi yang kompleks dan menarik. Menganalisis sentimen publik terhadap kandidat dan isu-isu politik memberikan wawasan penting mengenai dinamika politik selama pemilu. Penelitian ini berfokus pada pemilihan presiden dan membandingkan kinerja dua algoritma klasifikasi populer, XGBoost dan SVM, untuk menentukan metode mana yang lebih efektif. Setelah melalui beberapa tahapan text preprocessing terhadap 562 tweet, kami menemukan bahwa mayoritas pengguna Twitter cenderung memilih 347 tweet “Prabowo”. Model Extreme Gradient Boosting (XGBoost) menunjukkan performa terbaik dengan presisi 78%, presisi 76%, recall 78%, dan skor f1 76%. Hasil tersebut menunjukkan bahwa XGBoost merupakan model terbaik untuk mengklasifikasikan opini publik terkait pemilu presiden 2024 dan memberikan kontribusi penting dalam memahami efektivitas metode klasifikasi dalam konteks pemilu presiden.</p>
Kata Kunci	
<p>Extreme Gradient Boosting, Support Vector Machine, Pemilihan Presiden, SMOTE, AdaBoost</p>	
Keywords	Abstract
<p><i>Extreme Gradient Boosting, Support Vector Machine, Presidential Election, SMOTE, AdaBoost</i></p>	<p><i>Presidential elections are influenced by a variety of factors, including candidate backgrounds, political issues, and ideological preferences, making presidential elections a complex and interesting subject of classification. Analyzing public sentiment towards candidates and political issues provides important insights into political dynamics during elections. This research focuses on the presidential election and compares the performance of two popular classification algorithms, XGBoost and SVM, to determine which method is more effective. After some text preprocessing of 562 tweets, we found that the majority of Twitter users tend to vote for 347 “Prabowo” tweets. The Extreme Gradient Boosting (XGBoost) model showed the best performance with 78% precision, 76% precision, 78% recall, and 76% f1 score. These results show that XGBoost is the best model for classifying public opinion related to the 2024 presidential election and makes an important contribution to understanding the effectiveness of classification methods in the context of presidential elections.</i></p>

A. Pendahuluan

Pemilihan presiden merupakan bentuk kegiatan politik yang bersifat kedaulatan rakyat, dikarenakan pada saat pemilihan rakyat menjadi pihak penentu dalam pengerjaan politik pada suatu tempat lewat pemungutan suara yang dilaksanakan secara serentak [1]. Ada hal yang menarik saat pilpres pada tahun 2019 di mana para kandidat memanfaatkan media sosial seperti *Twitter* dalam melakukan kampanye [2]. Laporan *We Are Social* menunjukkan, jumlah pengguna aktif media sosial di Indonesia sebanyak 167 juta orang pada Januari 2023 [3]. Jumlah pengguna *Twitter* di Tanah Air mencapai 25,25 juta pengguna per Juli 2023, naik 71,2% secara kuartalan [4]. Semakin meluasnya penggunaan *Twitter* sebagai sarana untuk menyampaikan pendapat, membuat pentingnya dilakukan klasifikasi sentiment atau opini publik terhadap persentase pendukung tiap calon presiden guna untuk memprediksi calon presiden mana yang akan menang. Klasifikasi sentimen adalah teknik yang digunakan untuk mengidentifikasi dan mengklasifikasikan sentimen atau *opini* dalam teks [5]. Kategori klasifikasi pada penelitian ini adalah pro terhadap Prabowo, Ganjar dan Anies.

Dalam penelitian ini, dilakukan klasifikasi menggunakan dua model algoritma berbeda, yaitu *Extreme Gradient Boosting* (XGBoost) dan *Support Vector Machine* (SVM). Kedua algoritma ini diuji menggunakan teknik pemrosesan *data mining* dan *Natural Language Processing* (NLP). Selanjutnya digabungkan dengan fitur seleksi. Fitur seleksi merupakan bagian penting untuk mengoptimalkan kinerja dari *classifier* [6]. Ada beberapa metode yang dapat digunakan dalam seleksi fitur salah satunya yaitu *boosting*. Salah satu algoritma *boosting* yang populer adalah *adaboost* [7]. Algoritma *adaptive boosting* (*adaboost*) membangun pengklasifikasi kuat dengan cara mengombinasikannya. dengan sejumlah pengklasifikasi sederhana (lemah) [8]. Penelitian [9], membuktikan bahwa *adaboost* dapat menaikkan akurasi algoritma *Support Vector Machine* (SVM) dari 81,75% menjadi 82,48%.

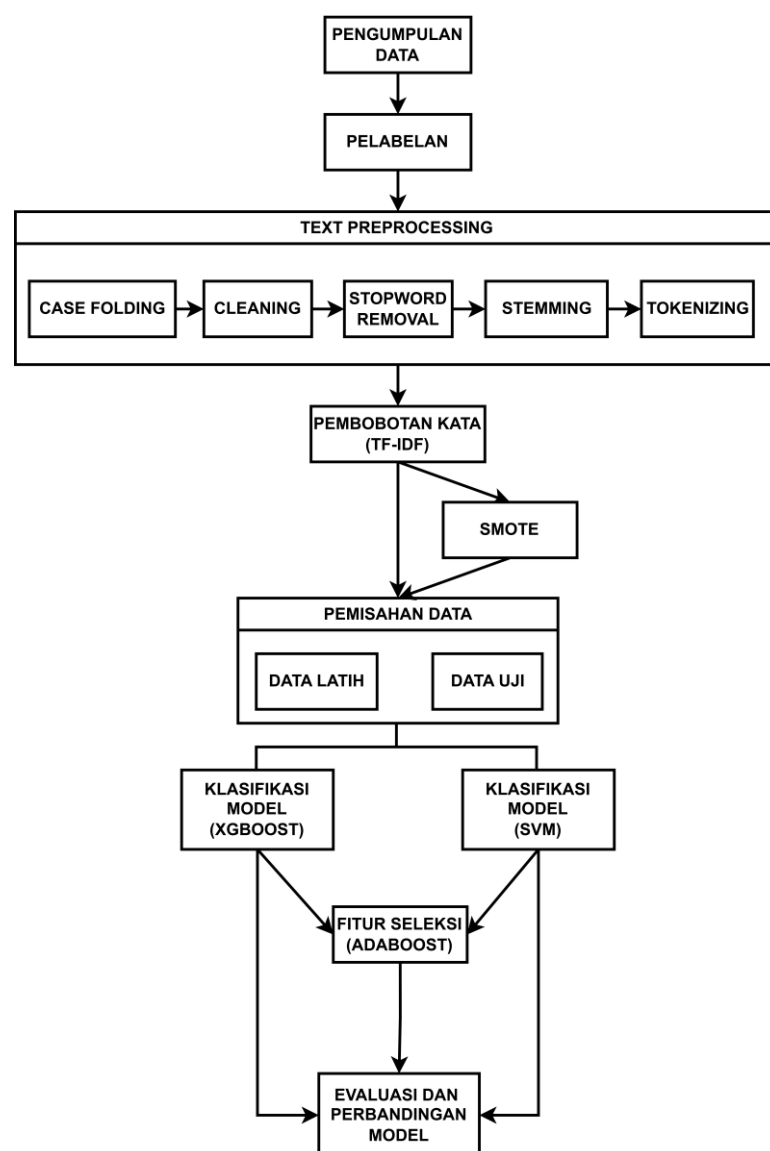
Dalam konteks pemilihan presiden, ketidakseimbangan data dapat menjadi masalah signifikan karena calon yang lebih populer mungkin memiliki jumlah dukungan yang jauh lebih besar daripada calon lainnya. Oleh karena itu, penelitian ini juga berusaha mengatasi ketidakseimbangan label dengan menerapkan teknik *oversampling* SMOTE (*Synthetic Minority Over-sampling Technique*). SMOTE menghasilkan sampel sintetik dari kelas minoritas dengan melakukan *oversampling* pada setiap titik data dengan mempertimbangkan kombinasi linier dari tetangga kelas minoritas yang ada [10]. Dalam studi yang dilakukan oleh [11], mengenai pemanfaatan *oversampling* dengan metode SMOTE menggunakan berbagai model prediktif, terbukti berhasil meningkatkan performa model, termasuk model *Support Vector Machine* (SVM) yang mencapai akurasi sebesar 98,9%. Sementara, penelitian yang dilakukan [12], yang meneliti analisis perbandingan klasifikasi dan penerapan strategi SMOTE dalam *data imbalanced credit card default* dengan algoritma *random forest* dan *extreme gradient boosting* (XGBoost), menemukan bahwa algoritma *extreme gradient boosting* (XGBoost) menggunakan SMOTE memiliki akurasi terbaik sebesar 72,78%. Penambahan SMOTE *oversampling* telah terbukti memahami dampak fitur terhadap pengembangan model [13].

Penelitian sebelumnya menunjukkan persamaan dalam penggunaan SMOTE pada algoritma *Support Vector Machine* (SVM) dan *Extreme Gradient Boosting*

(XGBoost) dapat meningkatkan performa model. Akan tetapi, belum terdapat penelitian yang membandingkan kedua algoritma tersebut menggunakan SMOTE untuk melakukan analisis opini publik terhadap pemilihan presiden.

Penelitian ini bertujuan untuk membandingkan kinerja algoritma *Extreme Gradient Boosting* (XGBoost) dan *Support Vector Machine* (SVM) dalam klasifikasi pemilihan presiden Indonesia 2024 dengan menerapkan teknik SMOTE. Dengan demikian, penelitian ini akan memberikan wawasan yang mendalam tentang efektivitas kedua algoritma tersebut dan dalam menangani masalah ketidakseimbangan data. Penelitian ini akan melakukan perbandingan dengan menggunakan hasil evaluasi seperti metric akurasi, presisi, *recall*, dan *F1-score*

B. Metode Penelitian



Gambar 1. Alur Penelitian

Pada Gambar 1, terdapat urutan langkah – langkah penelitian yaitu mengumpulkan data, kemudian tahap pra-pemrosesan data, selanjutnya tahap

pelabelan data, lalu tahap pemisahan data, dengan membagi data latih dan data uji, selanjutnya pada data latih, dilakukan perbandingan klasifikasi model dengan menerapkan teknik SMOTE *oversampling*, dan seleksi fitur Adaboost, setelah itu dilakukan tahap evaluasi model, dan diakhiri dengan kesimpulan dari hasil perbandingan metode *Extreme Gradient Boosting* (XGBoost) dan *Support Vector Machine* (SVM).

1. Pengumpulan Data

Data yang digunakan dalam penelitian ini yaitu data *public* yang bersumber dari *Github*. Data tersebut diperoleh pada bulan April – Mei 2023 yang berjumlah 4673 *tweet*. Data ini merupakan opini dari pengguna *Twitter* yang melakukan survey ataupun *voting* terhadap calon presiden 2024. Data ini berupa hasil dari *crawling data* melalui *Twitter*.

2. Text Preprocessing

Sebelum melakukan analisis teks, *preprocessing* seharusnya dilakukan terlebih dahulu untuk mengeliminasi kata-kata yang tidak diperlukan [14]. *Text preprocessing* adalah tahapan awal dari *text mining* yang bertujuan untuk mempersiapkan teks menjadi data yang akan mengalami pengolahan pada tahapan berikutnya [15]. Dalam penelitian ini, proses *preprocessing* melibatkan beberapa langkah, yaitu, *case folding*, *cleaning*, *stopword removal*, *stemming* dan *tokenization*.

1. Proses *case folding* bertujuan untuk merubah semua huruf yang ada pada sebuah dokumen teks menjadi huruf kecil, sebagai contoh kata “*Wifi*” menjadi “*wifi*” [16].
2. *Cleaning* adalah proses untuk menghilangkan kata yang tidak relevan dari dokumen [17]. Karakter yang dihilangkan seperti tanda atau simbol, *mention*, *emoticon*, *hashtag*, dan URL.
3. *Stopword removal* bertujuan untuk menghilangkan kata-kata yang tidak penting, proses ini dapat mengurangi dimensi ruang yang terlihat berat [16].
4. *Stemming* merupakan suatu proses yang terdapat dalam sistem IR yang mentransformasi kata – kata yang terdapat dalam suatu dokumen ke kata – kata akarnya (*root word*) dengan menggunakan aturan – aturan tertentu [18].
5. Proses tokenisasi memecah dokumen menjadi beberapa bagian kata yang disebut token [16].

3. Pelabelan Data

Setelah dilakukan *text preprocessing*, maka dataset yang telah dibersihkan akan diberikan label pada setiap data untuk mempermudah proses klasifikasi. Pada tahap ini dilakukan pelabelan secara manual menggunakan 3 label/kategori yaitu, pro terhadap “Prabowo”, “Ganjar”, dan “Anies”. Pelabelan dilakukan dengan membaca setiap *tweet* dan menganalisis *tweet* tersebut berpihak kepada calon presiden siapa.

4. Pembobotan Kata

Pembobotan kata adalah suatu proses yang mengubah kata menjadi representasi numerik (vektor kata) [19]. Salah satu pembobotan kata yang populer adalah algoritma *Term Frequency-Inverse Document Frequency* (TF-IDF). Algoritma ini menggabungkan dua konsep penghitungan bobot, yaitu frekuensi kata dalam dokumen tertentu dan frekuensi inverse dari dokumen yang mengandung kata

tersebut. Frekuensi kemunculan sebuah kata dalam dokumen tertentu menunjukkan pentingnya kata tersebut dalam dokumen [20]. Algoritma ini digunakan untuk menentukan bobot dari suatu *term* (kata) t , pada suatu dokumen d [21]. Berikut rumus pembobotan kata TF-IDF:

$$W_{ij} = tf_{ij} \times \log(ndf) \quad (1)$$

Dimana W_{ij} adalah bobot yang menentukan pentingnya kata atau *term*, tf_{ij} adalah frekuensi *term* yaitu berapa kali *term* muncul, dan ndf adalah frekuensi dokumen total dari *term* dalam seluruh koleksi dokumen.

5. Pemisahan Data

Setelah dataset melalui proses pembobotan kata, langkah selanjutnya adalah melakukan pemisahan data (*splitting data*). Pemisahan data ini bertujuan untuk membagi dataset menjadi dua subset yang berbeda, yaitu data latih (*training data*) dan data uji (*testing data*). Dalam penelitian ini, dilakukan empat eksperimen dengan rasio pemisahan data yang berbeda, yaitu 60:40, 70:30, 80:20, dan 90:10.

6. Synthetic Minority Oversampling Technique (SMOTE) oversampling

Setelah dilakukan tahapan pemisahan data, selanjutnya dilakukan *Synthetic Minority Oversampling Technique* (SMOTE) dapat meningkatkan akurasi dari metode klasifikasi [22]. Ketika melakukan klasifikasi data, salah satu masalah yang sering terjadi adalah ketidakseimbangan jumlah data antara kelas yang berbeda [23]. Oleh karena itu, digunakan teknik *Synthetic Minority Oversampling Technique* (SMOTE) adalah teknik canggih untuk memecahkan masalah *overfitting* pada kumpulan data yang tidak seimbang [24]. Teknik ini bekerja dengan cara menyeimbangkan jumlah distribusi data sampel pada kelas minoritas dengan menyeleksi data sampel tersebut sampai jumlah data sampel menjadi seimbang dengan jumlah sampel pada kelas mayoritas [25].

7. Extreme Gradient Boosting (XGBoost)

Extreme Gradient Boosting (XGBoost) merupakan salah satu teknik dalam *machine learning* untuk analisis regresi dan klasifikasi berdasarkan *Gradient Boosting Decision Tree* [12]. Algoritma ini memungkinkan melakukan optimasi 10 kali lebih cepat dibandingkan dengan *Gradient Boosting Machine* (GBM) lainnya. *Extreme Gradient Boosting* (XGBoost) mengambil ekspansi *Taylor* dari *loss function* hingga orde kedua dan menambahkan istilah regularisasi untuk menemukan solusi optimal, yang digunakan untuk menyeimbangkan penurunan fungsi tujuan dan kompleksitas model untuk menghindari *overfitting* [26]. Model *Extreme Gradient Boosting* (XGBoost) dapat direpresentasikan pada Persamaan (2) berikut.

$$\hat{y}_l = \sum_{k=1}^K f_k(x_i), f_k \in F \quad (2)$$

Keterangan:

K = Jumlah pohon keputusan

$f_k(x_i)$ = Fungsi *input* pada pohon keputusan ke- k

\hat{y}_l = Nilai prediksi

F = Himpunan dari semua CART yang mungkin

8. Support Vector Machine (SVM)

Support Vector Machine (SVM) adalah metode pembelajaran terawasi yang bekerja dengan cara mencari *hyperplane* dimana *hyperplane* ini berfungsi untuk memisahkan data dari dua atau lebih tipe kelas yang berbeda [27]. *Support Vector Machine* (SVM) mencoba untuk menemukan *hyperplane* yang paling baik memisahkan data *tweet* dengan sentimen negatif, netral, dan positif [13]. Prinsip *Support Vector Machine* (SVM) adalah membangun *hyperplane* yang memiliki ukuran *margin* yang sama dan tidak cenderung mendekati daerah dari salah satu kelas [28]. Rumus umum untuk perhitungan *Support Vector Machine* (SVM) adalah sebagai berikut:

$$f(x) = \text{sign}(w \cdot x + b) \quad (3)$$

Keterangan:

$F(x)$ = Fungsi prediksi

w = Vektor normal *hyperplane*

x = Vektor fitur *input*

b = bias atau *intercept*

9. *Adaptive Boosting* (Adaboost)

Algoritma *adaptive boosting* (adaboost) membangun pengklasifikasi kuat dengan cara mengombinasikannya dengan sejumlah pengklasifikasi sederhana (lemah) [8]. Cara kerja algoritma adaboost adalah memberikan bobot yang sama pada setiap sampel pada awalnya. Setelah setiap klasifikasi, bobot hasil yang benar berkurang, dan bobot hasil yang salah bertambah. Proses ini diulangi hingga mencapai ambang batas atau jumlah siklus maksimum [29].

10. Evaluasi Model

Selanjutnya, dilakukan proses evaluasi model menggunakan *confusion matrix* yang merepresentasikan kebenaran dari sebuah klasifikasi [30]. *Confusion matrix* digunakan untuk membandingkan kelas prediksi dengan kelas data yang sebenarnya [31]. Kinerja model pada penelitian ini diukur menggunakan metrik *accuracy*, *recall*, *precision*, dan *F1-Score*. Model *confusion matrix* dapat dilihat pada tabel berikut:

Tabel 1. *Confusion Matrix*

		Kelas Sebenarnya	
		<i>True</i>	<i>False</i>
Kelas Prediksi	<i>True</i>	TP	FP
	<i>False</i>	FN	TN

Keterangan:

True Positive (TP) = Jumlah dokumen dari kelas *true* yang benar diklasifikasikan sebagai kelas *true*.

True Negative (TN) = Jumlah dokumen yang berasal dari kelas *true* salah yang diklasifikasikan sebagai kelas *false*.

False Positive (FP) = Jumlah dokumen yang berasal dari kelas *false* yang salah yang diklasifikasikan sebagai kelas *false*.

False Negative (FN) = Jumlah dokumen yang berasal dari kelas *true* yang salah yang diklasifikasikan sebagai kelas *false*.

Fungsi matematika dari metrik akurasi, presisi, *recall*, dan *F1-Score* dapat dilihat di bawah ini:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$
$$Precision = \frac{TP}{TP + FN}$$
$$Recall = \frac{TP}{TP + FP}$$
$$F1 - score = 2 \times \frac{recall \times precision}{recall + precision}$$

C. Hasil dan Pembahasan

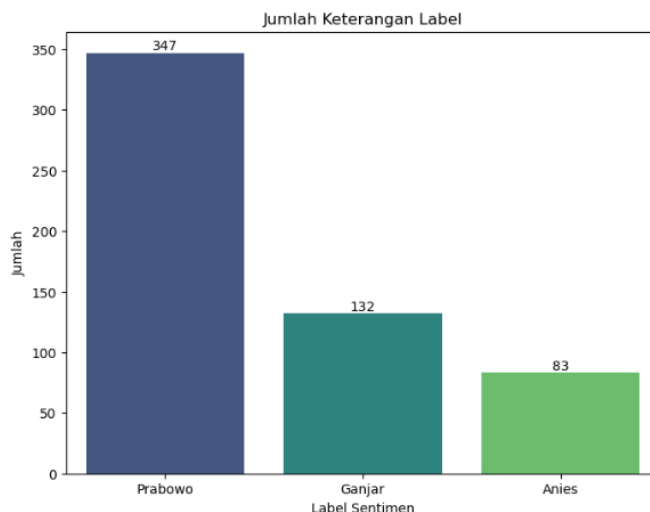
Terdapat sebanyak 4763 data mentah yang akan melalui *preprocessing text* agar data yang digunakan lebih akurat nantinya. Tahapan dalam *preprocessing text* yang dilakukan meliputi *case folding*, *cleaning* seperti menghapus *link* hingga karakter yang tidak berguna seperti titik, tanda tanya, dan lain – lain, kemudian dilakukan *stopword removal*, *stemming*, dan *tokenization*. Setelah dilakukan *preprocessing text*, maka didapatkan data bersih sebanyak 562 data dan dilakukan pelabelan data dalam 3 kategori, yaitu “Anies”, “Ganjar”, dan “Prabowo”.

	Sentimen	Stopword	Stemming	Token	Label
0	skor survei capres, 7 lembaga survei menyatakan...	skor survei capres , 7 lembaga survei ganjar p...	skor survei capres 7 lembaga survei ganjar pra...	['skor', 'survei', 'capres', '7', 'lembaga', '...', 'ganjar']	Ganjar
1	test ombak capres pilihan kalian anies baswed...	test ombak capres pilihan anies baswedan : bac...	test ombak capres pilih anies baswedan backhan...	['test', 'ombak', 'capres', 'pilih', 'anies', 'baswedan', 'backhan', '...']	Prabowo
2	hasil survei cncb indonesia tanggal 2 mei 2023...	hasil survei cncb indonesia tanggal 2 mei 2023...	hasil survei cncb indonesia tanggal 2 mei 2023...	['hasil', 'survei', 'cncb', 'indonesia', 'tanggal', '2', 'mei', '2023', '...']	Anies
3	capres cawapres terfavorit: 1. anies basweda...	capres cawapres terfavorit : 1. anies baswedan...	capres cawapres favorit 1 anies baswedan - ruh...	['capres', 'cawapres', 'favorit', '1', 'anies', 'baswedan', 'ruh', '...']	Anies
4	alhamdulillah, capres anies pernah jadi gub...	alhamdulillah , capres anies gub dki , capres ...	alhamdulillah capres anies gub dki capres prab...	['alhamdulillah', 'capres', 'anies', 'gub', 'dki', 'prabowo', '...']	Ganjar
...
557	soal ganjar pranowo dan jokowi, rocky gerung s...	ganjar pranowo jokowi , rocky gerung sarankan ...	ganjar pranowo jokowi rocky gerung saran prabo...	['ganjar', 'pranowo', 'jokowi', 'rocky', 'gerung', 'saran', 'prabowo', '...']	Ganjar
558	menurut anda siapakah calon presiden ri terbaik...	calon presiden ri terbaik terbaik periode 2024...	calon presiden ri baik baik periode 2024 - 202...	['calon', 'presiden', 'ri', 'baik', 'baik', 'periode', '2024', '2025', '...']	Anies
559	menuju ri 1 2024 :monaco: 1. letnan jenderal t...	ri 1 2024 : monaco : 1. letnan jenderal tni h...	ri 1 2024 monaco 1 letnan jenderal tni h prabo...	['ri', '1', '2024', 'monaco', '1', 'letnan', 'jenderal', 'tni', 'h', 'prabowo', '...']	Anies
560	buku yang dibaca pak principle for navigating...	buku dibaca principle for navigating big debt ...	buku baca principle for navigating big debt cr...	['buku', 'baca', 'principle', 'for', 'navigating', 'big', 'debt', 'crisis', '...']	Anies
561	semoga pak bisa bergabung dengan pak anies ba...	semoga bergabung anies baswedan .	moga gabung anies baswedan	['moga', 'gabung', 'anies', 'baswedan', '...']	Anies

562 rows × 5 columns

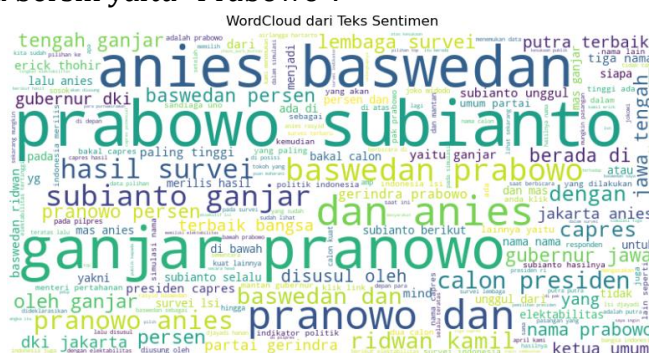
Gambar 2. Hasil *Text Preprocessing* dan Pelabelan

Setelah dilakukan *preprocessing text* dan pelabelan data, dilakukan penghitungan terhadap jumlah label untuk mengetahui seberapa banyak opini publik yang memilih tiap kandidat calon presiden 2024.



Gambar 3. Hasil Jumlah Per Label

Gambar diatas menunjukkan bahwa ternyata pengguna Twitter dominan memilih “Prabowo” pada pemilihan presiden 2024, disusul oleh “Ganjar” serta “Anies”. Proses selanjutnya yaitu mengimplementasikan data kedalam algoritma XGBoost, XGBoost dengan SMOTE, XGBoost dengan Adaboost, serta XGBoost menggunakan SMOTE dan Adaboost. Selanjutnya dilakukan perbandingan dengan algoritma SVM, SVM dengan SMOTE, SVM dengan Adaboost, serta SVM menggunakan SMOTE dan Adaboost ditiap pemisahan data 60:40, 70:30, 80:20, dan 90:10. Penelitian ini dilakukan menggunakan metode pembobotan kata yaitu TF-IDF. Visualisasi pada gambar 3 menunjukkan kata – kata yang sering muncul dalam keseluruhan data bersih yaitu “Prabowo”.



Gambar 4. Hasil *WordCloud*

Tahap akhir dalam penelitian ini adalah melakukan evaluasi model yang bertujuan untuk memastikan bahwa model yang telah dibangun dapat memenuhi kebutuhan dan tujuan yang telah ditentukan.

Tabel 2. *Report* Evaluasi Model 60:40

<i>Splitting</i>	Model	Akurasi	Presisi	Recall	F1-Score
60 : 40	SVM	71%	72%	71%	66%
	SVM – SMOTE	68%	66%	68%	66%
	SVM – AB	61%	38%	61%	47%
	SVM – SMOTE – AB	62%	66%	62%	60%

XGBoost	69%	67%	69%	67%
XGBoost – SMOTE	68%	66%	68%	66%
XGBoost – AB	64%	64%	64%	64%
XGBoost – SMOTE – AB	62%	65%	62%	63%

Tabel diatas menunjukkan hasil perbandingan tiap algoritma dari pemisahan data 60:40. Melalui proses klasifikasi dan evaluasi model pada pemisahan data 60:40, didapatkan bahwa model *Support Vector Machine* (SVM) berhasil memperoleh akurasi, presisi, recall, dan f1-score tertinggi dibanding model lain, yaitu sebesar 71%, 72%, 71%, dan 66%.

Tabel 3. Report Evaluasi Model 70:30

<i>Splitting</i>	Model	Akurasi	Presisi	Recall	F1-Score
70 : 30	SVM	73%	71%	73%	70%
	SVM – SMOTE	65%	68%	65%	66%
	SVM – AB	63%	40%	63%	49%
	SVM – SMOTE – AB	61%	59%	61%	59%
	XGBoost	72%	69%	72%	70%
	XGBoost – SMOTE	70%	68%	70%	69%
	XGBoost – AB	57%	67%	57%	60%
	XGBoost – SMOTE – AB	65%	71%	65%	67%

Selanjutnya pada tabel diatas, dilakukan klasifikasi dan evaluasi model terhadap tiap algoritma dengan pemisahan data 70:30. Hasil yang didapatkan bahwa model *Support Vector Machine* (SVM) berhasil memperoleh akurasi, presisi, recall, dan f1-score tertinggi dibanding model lain, yaitu sebesar 73%, 71%, 73%, dan 70%.

Tabel 4. Report Evaluasi Model 80:20

<i>Splitting</i>	Model	Akurasi	Presisi	Recall	F1-Score
80 : 20	SVM	74%	76%	74%	71%
	SVM – SMOTE	69%	68%	69%	68%
	SVM – AB	64%	41%	64%	50%
	SVM – SMOTE – AB	63%	70%	63%	62%
	XGBoost	78%	76%	78%	76%
	XGBoost – SMOTE	69%	65%	69%	67%
	XGBoost – AB	67%	70%	67%	67%
	XGBoost – SMOTE – AB	58%	62%	58%	69%

Selanjutnya pada tabel diatas, dilakukan klasifikasi dan evaluasi model terhadap tiap algoritma dengan pemisahan data 80:20. Hasil yang didapatkan bahwa model *Extreme Gradient Boosting* (XGBoost) berhasil memperoleh akurasi, presisi, recall, dan f1-score tertinggi dibanding model lain, yaitu sebesar 78%, 76%, 78%, dan 76%.

Tabel 5. Report Evaluasi Model 90:10

<i>Splitting</i>	Model	Akurasi	Presisi	Recall	F1-Score
90 : 10	SVM	75%	77%	75%	73%
	SVM – SMOTE	70%	69%	70%	69%
	SVM – AB	61%	38%	61%	47%
	SVM – SMOTE – AB	58%	68%	58%	58%
	XGBoost	74%	70%	74%	71%
	XGBoost – SMOTE	74%	70%	74%	71%
	XGBoost – AB	57%	67%	57%	60%
	XGBoost – SMOTE – AB	63%	64%	63%	63%

Terakhir, pada tabel diatas dilakukan klasifikasi dan evaluasi model terhadap tiap algoritma dengan pemisahan data 90:10. Hasil yang didapatkan bahwa model *Support Vector Machine* (SVM) berhasil memperoleh akurasi, presisi, recall, dan f1-score tertinggi dibanding model lain, yaitu sebesar 75%, 77%, 75%, dan 73%.

D. Simpulan

Penelitian ini menghasilkan data bersih sebanyak 562 *tweet* setelah melalui beberapa tahapan *text preprocessing*, yaitu *case folding*, *cleaning*, *stopword removal*, *stemming* dan *tokenizing*. Dataset yang telah dibersihkan menunjukkan bahwa pengguna *Twitter* lebih condong memilih “Prabowo” yaitu sebanyak 347 *tweet*. Setelah membandingkan beberapa model dengan keempat pemisahan data, didapatkan model *Extreme Gradient Boosting* (XGBoost) berhasil mencapai nilai evaluasi tertinggi dengan akurasi 78%, presisi 76%, *recall* 78%, dan *f1-score* 76%. Hal ini menunjukkan model *Extreme Gradient Boosting* (XGBoost) menjadi model terbaik dalam melakukan klasifikasi opini publik terhadap pemilihan presiden 2024.

E. Referensi

- [1] Delvika, B., Abror, N., & Gurning, U. R. (2023). Perbandingan Algoritma NBC dan C4. 5 Dalam Analisa Sentimen Pemilihan Presiden 2024 Pada Twitter: Comparison of the NBC and C4. 5 Algorithms in Sentiment *SENTIMAS: Seminar Nasional* ..., 41–48. <https://journal.irpi.or.id/index.php/sentimas/article/view/548%0Ahttps://journal.irpi.or.id/index.php/sentimas/article/download/548/336>
- [2] Akmaluddin, A., & Wempie, M. A. R. (2022). Budaya Politik Pemilih Milenial Dalam Pemilu Tahun 2024 Di Kabupaten Ogan Komering Ulu. *Jurnal Ilmu Pemerintahan Unbara*, 1(1), 49–57.
- [3] Widi, S. (2023). *Pengguna Media Sosial di Indonesia Sebanyak 167 Juta pada 2023*. <https://dataindonesia.id/internet/detail/pengguna-media-sosial-di-indonesia-sebanyak-167-juta-pada-2023>
- [4] Annur, C. M. (2023). *Jumlah Pengguna Twitter Indonesia Duduki Peringkat Ke-4 Dunia per Juli 2023*. <https://databoks.katadata.co.id/datapublish/2023/11/01/jumlah-pengguna-twitter-indonesia-duduki-peringkat-ke-4-dunia-per-juli-2023>
- [5] Parimala, M., Swarna Priya, R. M., Praveen Kumar Reddy, M., Lal Chowdhary, C., Kumar Poluru, R., & Khan, S. (2021). Spatiotemporal-based sentiment analysis

- on tweets for risk assessment of event using deep learning approach. *Software - Practice and Experience*, 51(3), 550–570. <https://doi.org/10.1002/spe.2851>
- [6] Harianto, H., Sunyoto, A., & Sudarmawan, S. (2020). Optimasi Algoritma Naïve Bayes Classifier untuk Mendeteksi Anomaly dengan Univariate Fitur Selection. *Edumatic: Jurnal Pendidikan Informatika*, 4(2), 40–49. <https://doi.org/10.29408/edumatic.v4i2.2433>
- [7] Pramadhana, D. (2021). Klasifikasi Penyakit Diabetes Menggunakan Metode CFS dan ROS dengan Algoritma J48 Berbasis Adaboost. *Edumatic: Jurnal Pendidikan Informatika*, 5(1), 89–98. <https://doi.org/10.29408/edumatic.v5i1.3336>
- [8] Mulyati, S., Informatika, T., Pamulang, U., & Pelanggan, C. (2017). 261256-Penerapan-Resampling-Dan-Adaboost-Untuk-a52Cf03a. 2(4), 190–199.
- [9] Andreyestha, Amir, R. F., Nawawi, I., Taufik, A., Eko Pramono, & Fajar Akbar. (2023). SELEKSI ATRIBUT PADA DATA TIDAK SEIMBANG NASABAH KOPERASI DENGAN OPTIMASI SMOTE DAN ADABOOST. *Jurnal Informatika Teknologi Dan Sains (Jinteks)*, 5(3), 322–327. <https://doi.org/10.51401/jinteks.v5i3.2757>
- [10] Turlapati, V. P. K., & Prusty, M. R. (2020). Outlier-SMOTE: A refined oversampling technique for improved detection of COVID-19. *Intelligence-based medicine*, 3, 100023.
- [11] Bujang, S. D. A., Selamat, A., Ibrahim, R., Krejcar, O., Herrera-Viedma, E., Fujita, H., & Ghani, N. A. M. (2021). Multiclass Prediction Model for Student Grade Prediction Using Machine Learning. *IEEE Access*, 9, 95608–95621. <https://doi.org/10.1109/ACCESS.2021.3093563>
- [12] Halim, S. F. N., & Azmi, U. (2023). Analisis Perbandingan Klasifikasi dan Penerapan Teknik SMOTE Dalam Imbalanced Data Pada Credit Card Default. *Jurnal Sains Dan Seni ITS*, 12(2). <https://doi.org/10.12962/j23373520.v12i2.111833>
- [13] Rabbani, S., Safitri, D., Rahmadhani, N., & Anam, M. K. (2023). Perbandingan Evaluasi Kernel SVM untuk Klasifikasi Sentimen dalam Analisis Kenaikan Harga BBM: Comparative Evaluation of SVM Kernels for Sentiment Classification in Fuel Price Increase Analysis. *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, 3(2), 153–160.
- [14] Rahutomo, F., Firdausi, A. T., & Rochmanshah, N. (2019). Pengembangan Sistem Analisa Keberpihakan Media Online Berdasarkan Trend Waktu Menggunakan Naive Bayes Classifier. *Jurnal Informatika Polinema*, 6(1), 33–40. <https://doi.org/10.33795/jip.v6i1.293>
- [15] Jayanti, L., Sentinuwo, S. R., Lantang, O. A., & Jacobus, A. (2016). Analisa Pola Penyalahgunaan Facebook Sebagai Alat Kejahatan Trafficking Menggunakan Data Mining. *Jurnal Teknik Informatika*, 8(1). <https://doi.org/10.35793/jti.8.1.2016.12231>
- [16] Rosid, M. A., Fitriani, A. S., Astutik, I. R. I., Mulloh, N. I., & Gozali, H. A. (2020, June). Improving text preprocessing for student complaint document classification using sastrawi. In *IOP Conference Series: Materials Science and Engineering* (Vol. 874, No. 1, p. 012017). IOP Publishing.
- [17] Palepa, M. J., Pratiwi, N., & Rohmansa, R. Q. (2024). Analisis Sentimen Masyarakat Tentang Pengaruh Politik Identitas Pada Pemilu 2024 Terhadap Toleransi Beragama Menggunakan Metode K-Nearest Neighbor. *JUPI (Jurnal*

- Ilmiah Penelitian dan Pembelajaran Informatika*), 9(1), 389-401.
- [18] Andika, R. (2023). Menentukan Dosen Pembimbing Secara Otomatis Dengan Algoritma Text Mining Dan TF-RF. *Bulletin of Information System Research*, 2(1), 165-173.
- [19] Rahmaddeni, R., & Akbar, F. (2023). Comparison of Naïve Bayes Algorithm, Support Vector Machine and Decision Tree in Analyzing Public Opinion on COVID-19 Vaccination in Indonesia. *Indonesian Journal of Artificial Intelligence and Data Mining*, 6(1), 8. <https://doi.org/10.24014/ijaidm.v6i1.19966>
- [20] Anam, M. K., Mahendra, M. I., Agustin, W., Rahmaddeni, R., & Nurjayadi, N. (2022). Framework for Analyzing Netizen Opinions on BPJS Using Sentiment Analysis and Social Network Analysis (SNA). *INTENSIF: Jurnal Ilmiah Penelitian Dan Penerapan Teknologi Sistem Informasi*, 6(1), 11-28. <https://doi.org/10.29407/intensif.v6i1.15870>
- [21] Harischandra, A. R., Pratama, M. F. A., Felix, F., & Laia, A. P. (2022). Aplikasi Pendukung Desain Interior dengan Sistem Rekomendasi Berdasarkan Nama Brand Perabot Menggunakan Algoritma Content-Based Filtering Berbasis Web. *Jurnal SIFO Mikroskil*, 23(1), 1-16. <https://doi.org/10.55601/jsm.v23i1.816>
- [22] Thanh-Long, N., Tran-Minh, & Hong-Chuong, L. (2022). A Back Propagation Neural Network Model with the Synthetic Minority Over-Sampling Technique for Construction Company Bankruptcy Prediction. *International Journal of Sustainable Construction Engineering and Technology*, 13(3), 68-79. <https://doi.org/10.30880/ijscet.2022.13.03.007>
- [23] Agustiningsih, A., Findawati, Y., & Kautsar, I. A. (2023). *Classification Of Vocational High School Graduates' Ability In Industry Using Extreme Gradient Boosting (Xgboost), Random Forest, And Logistic Regression*. 4(4), 977-985. <https://doi.org/10.52436/1.jutif.2023.4.4.945>
- [24] Rupapara, V., Rustam, F., Shahzad, H. F., Mehmood, A., Ashraf, I., & Choi, G. S. (2021). Impact of SMOTE on Imbalanced Text Features for Toxic Comments Classification Using RVVC Model. *IEEE Access*, 9, 78621-78634. <https://doi.org/10.1109/ACCESS.2021.3083638>
- [25] Kasanah, A. N., Muladi, M., & Pujiyanto, U. (2019). Penerapan Teknik SMOTE untuk Mengatasi Imbalance Class dalam Klasifikasi Objektivitas Berita Online Menggunakan Algoritma KNN. *Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi)*, 3(2), 196-201. <https://doi.org/10.29207/resti.v3i2.945>
- [26] Chen, Y., Wang, X., Jung, Y., Abedi, V., Zand, R., Bikak, M., & Adibuzzaman, M. (2018). Classification of short single-lead electrocardiograms (ECGs) for atrial fibrillation detection using piecewise linear spline and XGBoost. *Physiological Measurement*, 39(10), 104006. <https://doi.org/10.1088/1361-6579/aadf0f>
- [27] Mustamu, L. I., & Sibaroni, Y. (2023). Fuel Increase Sentiment Analysis Using Support Vector Machine With Particle Swarm Optimization and Genetic Algorithm As Feature Selection. *Jurnal Teknik Informatika (Jutif)*, 4(3), 521-528. <https://doi.org/10.52436/1.jutif.2023.4.3.881>
- [28] Fikri, M. I., Sabrila, T. S., & Azhar, Y. (2020). Perbandingan Metode Naïve Bayes dan Support Vector Machine pada Analisis Sentimen Twitter. *SMATIKA JURNAL*, 10(02), 71-76. <https://doi.org/10.32664/smatika.v10i02.455>
- [29] Tanti, T., Sirait, P., & Andri, A. (2021). Optimalisasi Kinerja Klasifikasi Melalui Seleksi Fitur dan AdaBoost dalam Penanganan Ketidakseimbangan Kelas.

-
- Jurnal Media Informatika Budidarma*, 5(4), 1377.
<https://doi.org/10.30865/mib.v5i4.3280>
- [30] Lutfiani, S., Saragih, T. H., Abadi, F., Faisal, M. R., & Kartini, D. (2023). Perbandingan Metode Extreme Gradient Boosting Dan Metode Decision Tree Untuk Klasifikasi Genre Musik. *Jurnal Informatika Polinema*, 9(4), 373–382. <https://doi.org/10.33795/jip.v9i4.1319>
- [31] Noveanto, M., Sastypratiwi, H., & Muhandi, H. (2022). Uji Akurasi Klasifikasi Emosi Pada Lirik Lagu Bahasa Indonesia. *Jurnal Sistem Dan Teknologi Informasi (JustIN)*, 10(3), 311. <https://doi.org/10.26418/justin.v10i3.56804>