Jurnal Ilmu Komputer dan Sistem Informasi (JIRSI) **e-ISSN 2830-3954**
Volume: 4, Nomor: 2, Mei 2025: 202-217 **p-ISSN 2830-6031**
https://jurnal.unity-academy.sch.id/index.php/jirsi/index

Optimalisasi Random Forest untuk Sentimen Bahasa Indonesia dengan GridSearch dan SMOTE

Random Forest Optimization for Indonesian Sentiment with GridSearch and SMOTE

Ahmad Fauzi[1*], Agus Heri Yunial[2], Dede Eko Saputro[3], Reza Saputra[4]
[1,2,3,4]Teknik Informatika, Universitas Pamulang Tangerang Selatan, Indonesia
E-mail: [1*]dosen02621@unpam.ac.id, [2]dosen02525@unpam.ac.id,
[3]dosen02627@unpam.ac.id, [4]dosen02620@unpam.ac.id

Abstrak

Penelitian ini berfokus pada optimasi algoritma Random Forest untuk analisis sentimen media sosial x berbahasa Indonesia dengan menggunakan TextBlob sebagai alat labeling, diikuti oleh teknik balancing data SMOTE dan optimasi hyperparameter dengan GridSearch. Data yang digunakan diambil dari 611 tweet dengan keyword ukt (uang kuliah tunggal). Labeling sentimen menggunakan TextBlob menghasilkan 438 sentimen negatif dan 173 sentimen positif. Metode SMOTE digunakan untuk menyeimbangkan data dengan terlebih dahulu membagi data menjadi 75% data latih dan 25% data uji. Vektorisasi data menggunakan tf-idf. Model algoritma Random Forest dievaluasi dengan akurasi awal menggunakan split data sebesar 73%, dan evaluasi cross validation dengan 10 k-fold menghasilkan nilai akurasi 75%. Optimasi yang dilakukan dengan hyperparameter GridSearch berhasil meningkatkan nilai akurasi menjadi 74%, sementara evaluasi cross validation menggunakan 10 k-fold akurasinya menjadi 89%. Dalam penelitian ini metode SMOTE efektif dalam menyeimbangkan data yang tidak seimbang, dan optimasi hyperparameter gridsearch berhasil meningkatkan nilai akurasi algoritma Random Forest dalam klasifikasi sentimen media sosial x berbahasa Indonesia dengan labeling otomatis texblob.

Kata kunci: Analisis sentimen; Hyperparameter Gridsearch; Random forest; TextBlob

Abstract

This research focuses on optimizing the Random Forest algorithm for sentiment analysis of social media x in Indonesian using TextBlob as a labeling tool, followed by the SMOTE data balancing technique and hyperparameter optimization with GridSearch. The data used was taken from 611 tweets with the keyword ukt (single tuition). Sentiment labeling using TextBlob produces 438 negative sentiments and 173 positive sentiments. The SMOTE method is used to balance the data by first dividing the data into 75% training data and 25% test data. Data vectorization using tf-idf. The Random Forest algorithm model was evaluated with an initial accuracy using split data of 73%, and cross validation evaluation with 10 k-folds produced an accuracy value of 75%. Optimization carried out with GridSearch hyperparameters succeeded in increasing the accuracy value to 74%, while cross validation evaluation using 10 k-fold accuracy was 89%. In this research, the SMOTE method was effective in balancing unbalanced data, and gridsearch hyperparameter optimization succeeded in increasing the accuracy value of the Random Forest algorithm in classifying social media sentiment x in Indonesian with automatic texblob labeling.

Keywords*: Sentiment analysis; Gridsearch Hyperparameters; Random forest; TextBlob.*

1. PENDAHULUAN

Analisis sentimen merupakan salah satu bidang penting dalam pengolahan bahasa alami (NLP) yang bertujuan untuk memahami opini publik terhadap berbagai isu, produk, atau layanan. Dalam konteks akademik dan kebijakan pendidikan, Uang Kuliah Tunggal (UKT) menjadi topik yang banyak dibicarakan di media sosial, terutama di Twitter. Oleh karena itu, analisis sentimen terhadap tweet dengan keyword "UKT" dapat memberikan wawasan berharga mengenai persepsi dan pendapat mahasiswa serta masyarakat umum terhadap kebijakan ini [1].

Salah satu tantangan utama dalam analisis sentimen adalah mengatasi masalah ketidakseimbangan data, di mana jumlah tweet dengan sentimen negatif dan positif seringkali tidak seimbang [2], [3]. Hal ini dapat mempengaruhi kinerja model klasifikasi, menyebabkan bias terhadap kelas yang dominan [4][5]. Oleh karena itu, diperlukan teknik khusus untuk mengatasi ketidakseimbangan ini dan meningkatkan akurasi model [5]. Penelitian sebelumnya memperkenalkan SMOTE sebagai solusi efektif untuk masalah ini, dan penelitian lain menunjukkan keberhasilan SMOTE dalam meningkatkan kinerja model klasifikasi pada dataset tidak seimbang [6]. Penelitian ini bertujuan untuk mengoptimalkan algoritma Random Forest dalam analisis sentimen tweet berbahasa Indonesia [7]. Dengan memanfaatkan teknik pra-pemrosesan data yang tepat, metode balancing data seperti SMOTE, dan optimasi hyperparameter menggunakan GridSearch, diharapkan dapat meningkatkan kinerja model dalam klasifikasi sentimen [2]. Penggunaan TextBlob sebagai alat labeling juga dieksplorasi untuk menentukan sentimen tweet setelah diterjemahkan ke dalam bahasa Inggris [8]. Penelitian membuktikan efektivitas TextBlob dalam analisis sentimen berbagai bahasa , sementara studi oleh Purnomo dan Sutopo menunjukkan bahwa GridSearch dapat meningkatkan akurasi model secara signifikan [2], [6].

Metodologi yang digunakan dalam penelitian ini mencakup beberapa tahap, mulai dari pengumpulan data tweet, pembersihan data, pra-pemrosesan teks, labeling sentimen, hingga pembangunan dan evaluasi model klasifikasi [9]. Data yang diambil dari Twitter dengan keyword "UKT" mengalami serangkaian proses pembersihan dan normalisasi untuk menghilangkan noise dan meningkatkan kualitas data [10]. Selanjutnya, data yang tidak seimbang diatasi dengan metode SMOTE sebelum dibagi menjadi data pelatihan dan pengujian [4]. Vektorisasi teks dilakukan menggunakan TF-IDF, dan model Random Forest dioptimalkan dengan GridSearch untuk mendapatkan kinerja terbaik [11]. Penelitian lain menunjukkan bahwa kombinasi TF-IDF dengan Random Forest dapat memberikan hasil yang kuat dalam klasifikasi teks [12]. Hasil dari penelitian ini diharapkan dapat memberikan kontribusi signifikan dalam bidang analisis sentimen, khususnya untuk data berbahasa Indonesia [13]. Dengan mengatasi masalah ketidakseimbangan data dan melakukan optimasi model, penelitian ini dapat menjadi acuan bagi penelitian serupa di masa depan [3]. Selain itu, temuan ini juga bermanfaat bagi para pengambil kebijakan di bidang pendidikan untuk memahami lebih baik persepsi dan opini mahasiswa terkait kebijakan UKT, sehingga dapat membuat keputusan yang lebih tepat dan responsif terhadap kebutuhan dan aspirasi mahasiswa [14]. Penelitian lain menggarisbawahi pentingnya analisis sentimen untuk pengambilan keputusan yang lebih baik dalam kebijakan publik [15].

2. METODE PENELITIAN

2.1 Analisis Data Media Sosial X (Twitter)

Proses *crawling* data dimulai dengan menggunakan *Tweet Harvest*, sebuah alat khusus untuk mengumpulkan *tweet* berdasarkan kata kunci. Untuk penelitian ini, kata kunci yang digunakan adalah "ukt" (Uang Kuliah Tunggal) karena relevansinya dengan topik penelitian. Menggunakan API *Twitter* melalui *Tweet Harvest*, melakukan autentikasi dan mengatur parameter pencarian untuk memastikan hanya *tweet* berbahasa Indonesia yang diambil. Parameter lain yang diatur termasuk batas waktu pengumpulan dan jumlah maksimal *tweet* yang diambil [16]. Setelah parameter ditetapkan, *Tweet Harvest* menjalankan pencarian dan mengumpulkan *tweet* beserta metadata yang terkait, seperti tanggal posting, jumlah *retweet*, dan jumlah *like*. Data yang terkumpul kemudian disimpan dalam format terstruktur, seperti CSV atau JSON. berikut gambar 1 *crawling* data *twitter* dengan *tweet harvest*:

```
# Crawling Data dengan keyword UKT

filename = 'UKT.csv'
search_keyword = 'ukt lang:id'
limit = 1000 #Gunakan limit secara bertahap

!npx --yes tweet-harvest@2.6.1 -o "{filename}" -s "{search_keyword}" -l {limit} --token ""
```

Gambar 1. Sourcode Crawling Tweet Harvest

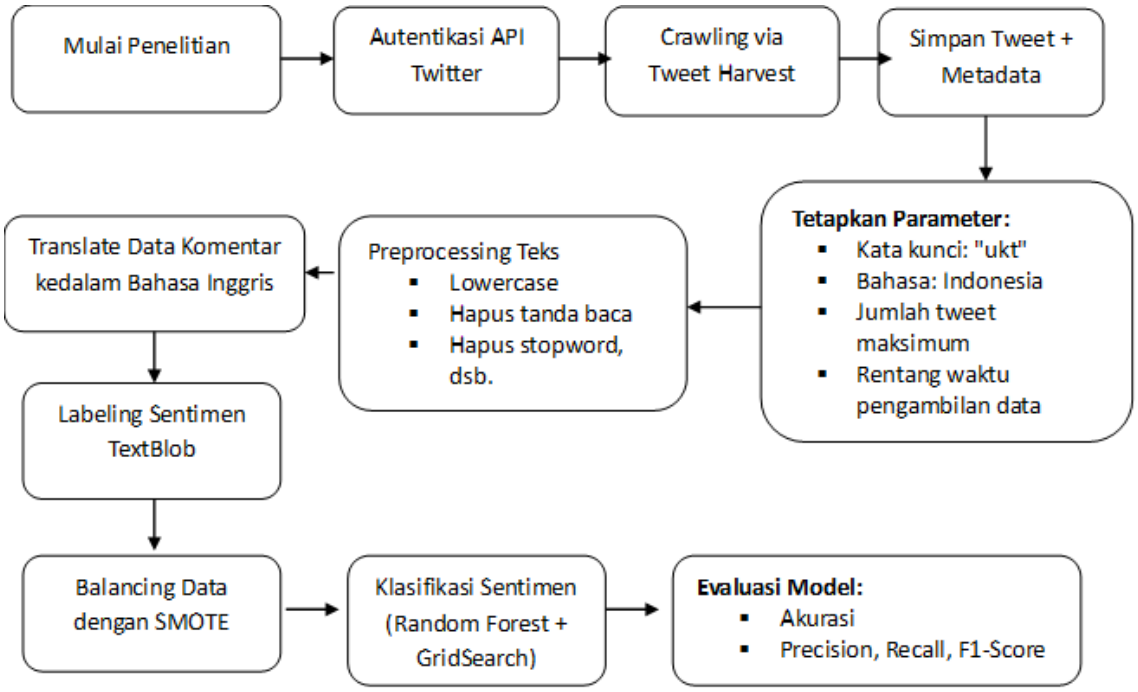
Dari proses ini, jumlah data yang diminta adalah 1.000 data namun hanya berhasil mengumpulkan 611 tweet yang relevan dengan kata kunci "ukt", hal ini karena tweet ukt pada media sosial x (twitter) hanya berjumlah 611 dari tahun 2019 sampai 28 Mei 2024. Data ini kemudian siap untuk tahap selanjutnya dalam penelitian, yaitu pembersihan dan pra-pemrosesan sebelum dilabeli dan dianalisis lebih lanjut.

Tahap selanjutnya adalah pembersihan dan pra-pemrosesan data. Pembersihan data adalah langkah awal yang sangat penting dalam analisis sentimen untuk menghilangkan noise dan memastikan data berkualitas tinggi [17], [18]. Dalam penelitian ini, pembersihan data mencakup penghapusan karakter khusus, URL, dan tanda baca yang tidak diperlukan. Selanjutnya, pra-pemrosesan data dilakukan dengan beberapa langkah penting: penghapusan stopword menggunakan NLTK stopword bahasa Indonesia, stemming menggunakan library Sastrawi, dan normalisasi menggunakan kamus slang Indonesia dari Kaggle. Penggunaan teknik pra-pemrosesan ini dapat secara signifikan meningkatkan kualitas data untuk analisis lebih lanjut [19], [20].

Setelah pra-pemrosesan, langkah berikutnya adalah labeling data. Dalam penelitian ini, TextBlob digunakan untuk menentukan sentimen positif atau negatif setelah mentranslate tweet ke bahasa Inggris. Hasil labeling menunjukkan adanya ketidakseimbangan data dengan 438 sentimen negatif dan 173 sentimen positif. Untuk menangani masalah ini, digunakan metode SMOTE (*Synthetic Minority Over-sampling Technique*). Metode ini terbukti efektif dalam menangani ketidakseimbangan data yang menunjukkan peningkatan kinerja model klasifikasi setelah menerapkan SMOTE.

Setelah penanganan data tidak seimbang, data dibagi menjadi 75% untuk data training dan 25% untuk data testing untuk memastikan bahwa model dapat diuji dengan data yang belum pernah dilihat sebelumnya. Vektorisasi data dilakukan menggunakan TF-IDF (*Term Frequency-Inverse Document Frequency*) untuk mengubah teks menjadi representasi numerik yang dapat diproses oleh algoritma machine learning.

Modeling dalam penelitian ini dilakukan dengan menggunakan algoritma Random Forest, yang dikenal karena kemampuannya dalam menangani dataset yang kompleks dan tidak seimbang. Setelah data divisualisasikan menggunakan TF-IDF, model Random Forest dibangun dan dioptimalkan menggunakan GridSearch untuk menemukan kombinasi hyperparameter terbaik. Evaluasi model dilakukan dengan membagi dataset menjadi 75% data training dan 25% data testing, kemudian mengukur kinerja model menggunakan metrik akurasi. Hasil awal menunjukkan akurasi sebesar 73%, yang kemudian meningkat menjadi 74% setelah optimasi.



Gambar 2. Proses Penelitian Sentimen Analisis Media Sosial X (Twitter)

2.2 Synthetic Minority Over-sampling Technique (SMOTE)

SMOTE (*Synthetic Minority Over-sampling Technique*) adalah teknik yang digunakan untuk menyeimbangkan kelas minoritas dalam dataset klasifikasi dengan menciptakan sampel sintetis berdasarkan tetangga terdekat dari kelas minoritas . Misalkan x_i adalah sampel minoritas, x_{zi} adalah salah satu tetangganya, dan x_{new} adalah sampel sintetis yang dibuat, maka rumus SMOTE untuk menciptakan sampel sintetis adalah sebagai berikut:

$$X_{new} = X_i + \alpha(X_{zi} - X_i)$$

Di mana α adalah angka acak antara 0 dan 1, x_i adalah sampel minoritas, dan x_{zi} adalah salah satu tetangganya dari x_i. SMOTE menghasilkan sampel baru yang serupa dengan sampel minoritas asli namun dengan variasi yang cukup untuk mewakili variasi yang mungkin ada dalam kelas tersebut. Hal ini membantu meningkatkan performa model klasifikasi terutama pada kelas minoritas dengan memperluas ruang sampel yang ada.

2.3 Random Forest

Random Forest merupakan algoritma klasifikasi dalam data mining yang efektif dan populer. Algoritma ini bekerja dengan cara menggabungkan prediksi dari beberapa pohon keputusan (decision trees) yang dibangun secara acak. Setiap pohon dalam Random Forest dibangun menggunakan subset acak dari fitur-fitur dalam dataset dan juga menggunakan teknik bootstraping untuk menghasilkan variasi yang lebih besar. Proses penggabungan prediksi dari pohon-pohon ini dilakukan melalui voting atau averaging, sehingga menghasilkan prediksi akhir yang lebih stabil dan akurat. Kelebihan utama dari Random Forest adalah kemampuannya dalam mengatasi overfitting, memproses dataset yang besar dengan cepat, dan memberikan estimasi kelas yang baik bahkan untuk data yang belum pernah dilihat sebelumnya.
langkah-langkah klasifikasi dalam menggunakan algoritma Random Forest dimulai dengan membagi dataset menjadi data latih dan data uji. Selanjutnya, Random Forest membangun beberapa pohon keputusan secara acak dengan memilih subset acak dari fitur-fitur dalam data latih dan menggunakan teknik bootstraping untuk sampel data. Setiap pohon dilatih secara independen untuk membuat prediksi berdasarkan fitur-fitur yang dipilih. Setelah semua pohon selesai dibangun, prediksi dari setiap pohon dijadikan sebagai pemilih untuk menentukan prediksi akhir dengan menggunakan metode voting atau averaging. Prediksi akhir inilah yang digunakan untuk mengevaluasi performa model Random Forest pada data uji, yang dapat memberikan informasi tentang seberapa baik model dapat mengklasifikasikan data yang belum pernah dilihat sebelumnya.

2.4 Hyperparameter GridSearch

Optimasi hyperparameter dengan GridSearch adalah teknik yang digunakan untuk mencari kombinasi hyperparameter terbaik untuk sebuah model machine learning. Dalam GridSearch, kita mendefinisikan kumpulan nilai yang mungkin untuk setiap hyperparameter yang ingin dioptimalkan, kemudian sistem secara sistematis mencoba semua kombinasi nilai tersebut dan memilih kombinasi yang memberikan kinerja terbaik berdasarkan metrik evaluasi yang ditentukan. Dalam analisis sentimen menunjukkan bahwa GridSearch mampu meningkatkan kinerja model dengan menemukan kombinasi hyperparameter yang optimal. Rumus umum GridSearch adalah sebagai berikut:

$$Gridsearch = \arg max_{\theta}\ score(\theta)$$

Di mana θ adalah kombinasi hyperparameter yang ingin dioptimalkan (misalnya, jumlah pohon dalam Random Forest, kedalaman maksimum pohon, dll.),

dan Score (θ) adalah metrik evaluasi yang digunakan (misalnya, akurasi, presisi, recall, dll.).

Selain itu, dalam penelitian menggambarkan bahwa GridSearch dapat digunakan secara efektif dalam mengoptimalkan model machine learning pada dataset dengan kompleksitas yang beragam, memastikan bahwa model yang dihasilkan memiliki kinerja yang maksimal [21].

2.5 Confusion Matrix

Confusion matrix adalah alat evaluasi yang penting dalam pemodelan klasifikasi. Ini memberikan gambaran yang jelas tentang seberapa baik model klasifikasi dapat memprediksi kelas-kelas yang berbeda dalam dataset. Confusion matrix terdiri dari empat sel, yaitu True Positive (TP) yang menunjukkan jumlah data yang benar diprediksi sebagai positif, True Negative (TN) yang menunjukkan jumlah data yang benar diprediksi sebagai negatif, False Positive (FP) yang menunjukkan jumlah data yang salah diprediksi sebagai positif, dan False Negative (FN) yang menunjukkan jumlah data yang salah diprediksi sebagai negatif. Rumus umum confusion matrix dan tabel sebagai berikut:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$$

Accuracy merupakan pengukuran seberapa sering model klasifikasi benar-benar memprediksi kelas dengan benar. di mana TP adalah True Positive, TN adalah True Negative, FP adalah False Positive, dan FN adalah False Negative.

$$Precision = \frac{TP}{TP+FP}$$

Precision digunakan untuk mengukur seberapa banyak dari semua prediksi positif yang sebenarnya benar. Di mana TP adalah True Positive dan FP adalah False Positive.

$$Recall = \frac{TP}{TP + FN}$$

Recall (Sensitivity) digunakan untuk mengukur seberapa banyak dari semua kelas positif yang berhasil diprediksi oleh model. Di mana TP adalah True Positive dan FN adalah False Negative.

$$F1 - Score = 2\ x \frac{Precision\ x\ Recall}{Precision + Recall}$$

F1-Score merupakan harmonic mean dari precision dan recall, memberikan gambaran yang seimbang antara keduanya. F1-Score berguna ketika kelas-kelas yang diprediksi positif dan negatif tidak seimbang.

Tabel 1. Confusion Matrix

	Prediksi Positif	Prediksi Negatif
Aktual Positif	TP	FN

Aktual Negatif	FP	TN

3. HASIL DAN PEMBAHASAN

3.1 Pembesihan dan Pra-Pemrosesan Data

Hasil pembersihan data merupakan tahap yang krusial dalam proses analisis data yang bertujuan untuk meningkatkan kualitas dan kebersihan dataset. Setelah melalui langkah-langkah seperti penghapusan karakter khusus, URL, dan tanda baca yang tidak diperlukan, serta penggunaan teknik seperti penghapusan stopword, stemming, dan normalisasi, dataset yang dihasilkan menjadi lebih bersih, terstruktur, dan siap untuk digunakan dalam analisis lebih lanjut. Proses ini membantu menghilangkan noise dan informasi yang tidak relevan, sehingga memungkinkan model machine learning untuk mempelajari pola yang lebih bermakna dan menghasilkan hasil yang lebih akurat dan dapat diandalkan. Berikut gambar merupakan sourcode pembersihan pada data twitter:

```
#menghapus karakter pada kolom full_text
def clean_twitter_text(text):
    text = re.sub(r'https?://\S+|www\.\S+','',text)
    text = re.sub(r'\$\w*','',text)
    text = re.sub(r'^RT[\s]+','',text)
    text = re.sub(r'https?://[^\s\n\r]+','',text)
    text = re.sub(r'#','',text)
    text = re.sub(r'http\S+','',text)
    text = re.sub(r'[^a-zA-Z\s]','', text)
    return text
data['full_text'] = data['full_text'].apply(clean_twitter_text)
data['full_text'].head(10)
```

Gambar 3. Sourcode Pembersihan Data Twitter

Dari Gambar 3 merupakan sourcode pembersihan data untuk menghilangkan karakter khusus, URL, dan tanda baca yang tidak diperlukan, kemudian data yang sudah dibersihkan akan disimpan kembali dalam atribut atau kolom full_text yang berisi kumpulan komentar pada media sosial x atau twitter dalam persoalan kenaikan ukt**.** selanjutnya dilakukan pembersihan dengan merubah teks komentar dalam kolom full_text menjadi huruf kecil dengan library lower pada python.

selanjutnya setelah dilakukan pembersihan data twitter yaitu pra-pemprosesan data, dalam pra-pemprosesan data akan dilakukan penghapusan kata yang tidak baku menjadi kata baku dalam bahasa Indonesia seperti kata yg akan diganti dengan yang, adapun caranya dengan melakukan normalisasi kata menggunakan kamus slang indonesia yang diambil dari situs kaggle respository. Dalam pra-pemrosesan data adalah melakukan penghapusan kata yang tidak memiliki arti dalam bahasa Indonesia seperti kata dan, atau, menggunakan library nltk yaitu stopword berbahasa indonesia.

Langkah selanjutnya dalam pra-pemrosesan adalah melakukan stemming kata dalam bahasa Indonesia menggunakan library sastrawi, stemming merupakan proses penghapusan kata menjadi kata dasar seperti pengumuman menjadi umum. Setelah proses pembersihan dan tahapan pra-pemrosesan selesai, maka akan di dapati data yang lebih bersih dari data sebelumnya. Berikut gambar 3 hasil pembersihan dan pra-pemrosesan data.

	full_text	stemmed_text
0	pengumuman terkait ukt	umum kait ukt
1	breaking kemendikbudristek memutuskan membatal...	breaking kemendikbudristek putus batal naik ukt
2	bpjs disesuaikan kelas dihapus ppn ukt publik ...	bpjs sesuai kelas hapus ppn ukt publik protes ...
3	ukt terbitlah tapera indonesia negara bu giman...	ukt terbit tapera indonesia negara bu gimana s...
4	ptn alerta alerta memanggil rekan rekan mahasi...	ptn alerta alerta panggil rekan rekan mahasisw...
...	...	...
606	ukt melambung ugm ui ptn unsoed kenaikan perse...	ukt lambung ugm ui ptn unsoed naik persen temp...
607	ketua pimpinan pusat pp muhammadiyah anwar abb...	ketua pimpin pusat pp muhammadiyah anwar abbas...
608	gilanghamidy lpdp urgent urgent ningkatin kual...	gilanghamidy lpdp urgent urgent ningkatin kual...
609	dibaca ya guys puluhan camaba unri mundur ukt ...	baca ya guys puluh camaba unri mundur ukt maha...
610	dips mama ayah penjual bakpau keliling kuliahi...	dips mama ayah jual bakpau keliling kuliahin u...

611 rows × 2 columns

Gambar 4. Hasil stemming dengan satrawi

3.2 Labeling data dengan TextBlob

Dalam melakukan labeling data twitter bahasa Indonesia terdapat banyak cara yaitu dengan melakukan labeling secara manual, otomatis dan semi otomatis. Pada penelitian ini labeling dilakukan secara otomatis dengan library textblob, akan tetapi karena pada textblob belum dapat melabeli langsung kata berbahasa Indonesia maka akan dilakukan penerjamahan kata twitter bahasa Indonesia kedalam bahasa inggris. Setelah proses penerjemahan kata selesai selanjutnya dilakukan labeling data twitter menggunakan textblob, untuk labeling sentimen pada penelitian ini hanya sentimen positif dan negatif.

Berikut merupakan tabel 2 hasil terjemahan teks kedalam bahasa inggris, dan hasil klasifikasi sentimen positif dan negatif menggunakan texblob.

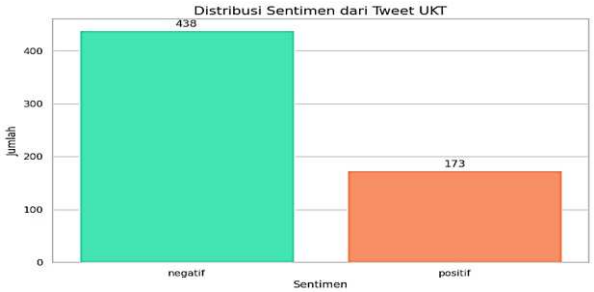
Tabel 2. Hasil Terjemahan Teks

No	Text sebelum Translated	Text setelah Translated	Sentiment
1	umum kait ukt	['announcement', 'related', 'to', 'ukt']	negatif
2	breaking kemendikbudristek putus batal naik ukt	['breaking', 'kemendikbudristek', 'decided', 'to', 'cancel', 'the', 'increase', 'in', 'ukt']	negatif
3	bpjs sesuai kelas hapus ppn ukt publik protes keras tunda tahun pertalite hapus tdl lpg	['bpjs', 'is', 'adjusted', 'for', 'the', 'class', 'to', 'be', 'deleted', 'by', 'the', 'ppn', 'ukt', 'public', 'protest', 'protest', 'postponed', 'a',	negatif

	sembako rakyat bayar tapera gaji utang nalar	'year', 'of', 'pertalite', 'deleted', 'tdl', 'lpg', 'sembako', 'people', 'pay', 'tapera', 'salary', 'payable', 'payable']	
4	ukt terbit tapera indonesia negara bu gimana sih biaya sejahtera	['ukt', 'is', 'published', 'by', 'tapera', 'indonesia', ',', 'how', 'about', 'the', 'cost', 'of', 'prosperity']	negatif
.....	....	.....	.....
611	dips mama ayah jual bakpau keliling kuliahin ukt atas gol kos aman jajan aman kip kipk bantu apa sekolah detik sma biayain swasta yaallah cont	['what', 'is', 'the', 'name', 'of', 'the', 'father', 'of', 'the', 'seller', 'of', 'buns', 'around', 'college', 'ukt', 'on', 'a', 'safe', 'boarding', 'house', 'for', 'safe', 'snacks', 'kip', 'kipk', 'aid', 'whatever', 'school', 'seconds', 'high', 'school', 'costs', 'private']	positif

3.3 Hasil Klasifikasi Sentimen Media Sosial X

Setelah data selesai dilakukan pembersihan, pra-prosesing dan sudah memilki labael sentiment langkah selanjutnya adalah melakukan klasifikasi data untuk diketahui tingkat akurasinya. Kalsifikasi dilakukan dengan menggunakan metode random forest, tahap awal dalam klasifikasi terlebih dahulu melihat jumlah Sentimen positif dan negatif pada data. Berikut merupakan gambar 5 jumlah sentimen positif dan negatif pada data media sosial x mengenai ukt.



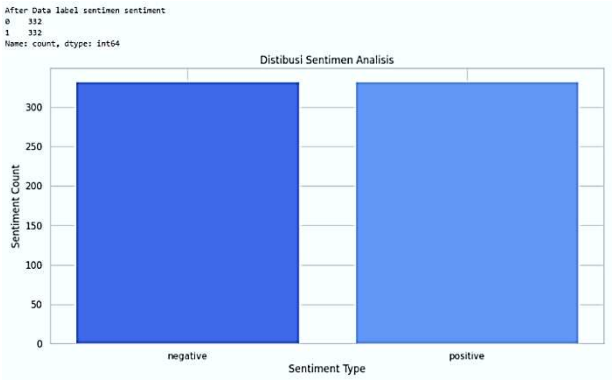
Gambar 5. Jumlah data sentimen

Dari gambar 5 dapat dilihat jumlah data sentimen positif sebanyak 173 dan jumlah sentiment negatif 438 dari total data 611. Jika dilihat persebaran data sentimen tidak seimbang, untuk melakukan ketidak seimbangan data dapat menggunakan salah satu metode imbalance data yaitu smote. Namun dalam melakukan imbalance data terlebih dahulu dilakukan pemisahan data dengan membagi data latih dan data uji serta merubah label data menjadi numerik dengan *encoder*. Pembagian data latih sebesar 75% dan data uji 25%. Sementara label sentiment positif dan negatif akan dirubah mejadi 0 dan 1, dimana 0 untuk sentiment negatif dan 1 untuk sentiment positif.

Selanjutnya data yang sudah dibagi menjadi data latih dan data uji, kemudian dilakukan vektorisasi data untuk memecah kata berdasarkan jumlah kemunculannya. Vektorisasi yang digunakan pada penelitian ini adalah *tf-idf*.

3.4 Hasil Imbalance Metode SMOTE

Setelah dilakukan pembagian data dan vektorisasi pada data teks, tahap selanjutnya mengatasi ketidak seimbangan data dengan metode smote, jumlah sentimen positif dan negatif akan di seimbangkan untuk memperoleh data yang lebih baik. Adapun data yang di seimbangkan adalah data latih dari hasil pembagian data. Berikut gambar 6 hasil imbalance data metode smote.



Gambar 6. Hasil Imbalance Metode Smote

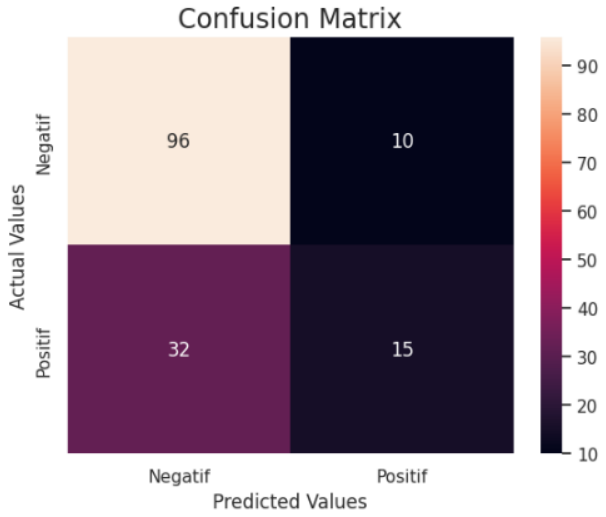
3.5 Hasil Evaluasi Algoritma Random Forest.

Data sentimen pada media sosial x mengenai ukt, selanjutnya akan di klasifikasin dengan algoritma random forest. Pada penghitungan algoritma random forest hasil akurasi yang diperoleh sebesar 73% dari pembagian data latih sebesar 75% dan data uji 25% dengan imbalance smote. Hasil akurasi algoritma random forest dapat dilihat dengan confusion matrix. Gambar 7 merupakan confusion matrix algoritma random forest.

	precision	recall	f1-score	support
0	0.75	0.91	0.82	106
1	0.60	0.32	0.42	47
accuracy			0.73	153
macro avg	0.68	0.61	0.62	153
weighted avg	0.70	0.73	0.70	153

Gambar 7. Confusion Matrix Algoritma Random Forest

Dari gambar 7 dapat dilihat akurasi algoritma yang dihasil oleh algoritma random forest sebesar 73%. Untuk mengetahui hasil prediksi yang dihasilkan dari algoritma random forest dapat dilakukan dengan visualisasi confusion matrix. Gambar 8 merupakan visualisasi confusion matrix algoritma random forest.



Gambar 8. Visualisasi Confusion Matrix Algoritma Random Forest

Dari gambar 8 dapat diketahui untuk sentimen negatif yang sesuai dengan prediksi sebanyak 96 adapun sebanyak 10 data yang bermula sentimen negatif diprediksi menjadi positif. Untuk sentimen positif terdapat 32 data sentimen yang diprediksi negatif dan 15 data yang diprediksi sesuai dari data aktualnya.
Untuk melihat konsistensi akurasi dari algoritma random forest dapat dilakukan penghitungan dengan cross validation dengan 10 k-fold. Dapat diketahui hasil penghitungan cross validation dengan 10 k-fold akurasi yang diperoleh sebesar 75%, cross validation sendiri melakukan pengujian dengan membagi 10 k-fold dari data latih.

3.6 Hasil Hyperparameter Gridsearch

Setelah diketahui hasil akurasi algoritma random forest untuk analisis sentimen pada media sosial x, dimana hasilnya adalah 73% hasil akurasi yang diperoleh akan ditingkatkan menggunakan hyperparameter gridsearch dengan melakukan penambahan parameter n_estimators, max_depth, min_samples_split, min_samples_leaf, dan bootstrap. Penambahan parameter ini bertujuan untuk mencari nilai parameter terbaik dari algoritma random forest. Untuk hasil parameter terbaik dapat dilihat pada tabel 3 berikut.

Tabel 3. Parameter gridsearch

Parameter gridsearch	Nilai Parameter	Hasil Best Parameter
n_estimators	[50, 100, 200]	200
max_depth	[None, 10, 20, 30]	None
min_samples_split	[2, 5, 10]	2
min_samples_leaf	[1, 2, 4]	1
bootstrap	[True, False]	True

Setelah mendapat parameter terbaik dari hasil hyperparameter gridsearch, selanjutnya melakukan penghitungan algoritma random forest dengan menambahkan hyperparameter gridsearch untuk meningkatkan hasil akurasi yang diperoleh sebelumnya. Algoritma random forest di optimasi dengan hasil hyperparameter gridsearch terbaik untuk meningkatkan nilai akurasi dalam analisis sentimen media sosial x mengenai ukt, adapun nilai optimasi yang berhasil diperoleh sebesar 74%, nilai tersebut lebih besar dari hasil akurasi sebelum dilakukan hyperparameter gridsearch.
Nilai akurasi yang dihasilkan dapat dilihat dengan confusion matrix, sehingga prediksi dari analisis sentimen data media sosial x dapat diketahui. Berikut gambar 9 hasil evaluasi dan confusion matrix algoritma random forest dengan hyperparameter gridsearch.

```
Akurasi Random Forest: 0.7450980392156863
Classification Report:
              precision    recall  f1-score   support

           0       0.73      0.99      0.84       106
           1       0.90      0.19      0.32        47

    accuracy                           0.75       153
   macro avg       0.82      0.59      0.58       153
weighted avg       0.79      0.75      0.68       153

Confusion Matrix:
 [[105   1]
 [ 38   9]]
```

Gambar 9. Hasil evaluasi dan confusion matrix algoritma random forest dengan hyperparameter gridsearch

Berdasarkan gambar 9 diketahui jumlah sentimen negatif yang berhasil diprediksi negatif sebanyak 105 dan 1 sentimen negatif yang diprediksi menjadi sentimen positif. Sementara untuk sentimen positif terdapat 38 yang diprediksi menjadi sentimen negatif, dan terdapat 9 data yang sesuai dengan prediksi sentimen positif. Dari hasil akurasi dengan hyperparameter gridsearch, dapat dilakukan penghitungan cross validation dengan 10 k-fold untuk melihat konsistensi dari nilai akurasinya, adapun hasil akurasi yang diperoleh adalah sebesar 89%. Berikut gambar 10 hasil akurasi cross validation optimasi algoritma random forest dengan hyperparameter gridsearch.

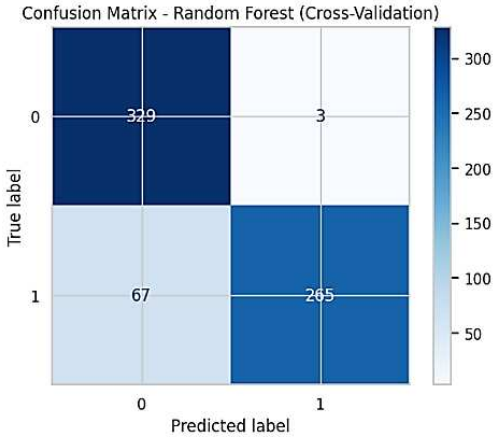
```
Cross-Validation Accuracy: 0.89
              precision    recall  f1-score   support

           0       0.83      0.99      0.90       332
           1       0.99      0.80      0.88       332

    accuracy                           0.89       664
   macro avg       0.91      0.89      0.89       664
weighted avg       0.91      0.89      0.89       664
```

Gambar 10. Hasil akurasi cross validation optimasi algoritma random forest dengan hyperparameter gridsearch

Dari gambar 10 dapat dilakukan visualisasi confusion matrix hasil akurasi cross validation algoritma random forest dengan hyperparameter gridsearch, visualisasi dilakukan untuk mengetahui jumlah prediksi data sentimen positif dan sentimen negatif. Berikut gambar 10 Visualisasi confusion matrix cross validation algoritma random forest dengan hyperparameter gridsearch.



Gambar 11. Visualisasi confusion matrix cross validation algoritma random forest dengan hyperparameter gridsearch

Hasil visualisasi gambar 11 menunjukkan bahwa dari hasil akurasi 89% terdapat 329 data sentimen negatif pada data latih yang diprediksi sesuai dengan data aktualnya, sedangkan 3 data sentimen negatif yang diprediksi menjadi sentimen positif. Sementara untuk sentimen positif terdapat 67 data yang diprediksi menjadi negatif, dan sebanyak 265 data diprediksi sesuai dengan sentimen positif. Hasil nilai dari gambar 10 dapat juga dilakukan penghitungan secara manual untuk mengetahui nilai akurasi, Precision, recall, f1-score, berikut penghitungannya.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$$

$$Accuracy = \frac{265 + 329}{265 + 329 + 67 + 3} = \frac{594}{664} = 0.89$$

$$Precision = \frac{TP}{TP + FP} = \frac{265}{265 + 67} = 0.79$$

$$Recall = \frac{TP}{TP + FN} = \frac{265}{265 + 3} = 0.98$$

$$F1 - Score = 2\ x \frac{Precision\ x\ Recall}{Precision + Recall}$$

$$F1 - Score = 2\ x \frac{0.79x0.98}{0.79 + 0.98} = 2\ x\ \frac{0.7742}{1.77} = 2\ x\ 0.437 = 0.87$$

Dari hasil penghitungan dapat diketahui nilai akurasi sebesar 89%, precision 79%, recall 98%, dan F1-scoresebesar 87%. Hasil tersebut memberikan pengetahuan bahwa data analisa sentimen media sosial x dengan bahasa Indonesia mengenai ukt memilki hasil prediksi yang cukup baik.

3.7 Perbandingan Hasil Evaluasi Algortima Randon Forest

Setelah dilakukan penghitungan untuk melakukan evaluasi terhadap data sentimen media sosial x bahasa Indonesia mengenai ukt, dapat dilihat hasil masing-masing nilai akurasi yang diperoleh dari algoritma random forest, baik yang sebelum dioptimasi ataupun yang sudah dioptimasi. Berikut tabel 4 perbandingan hasil evaluasi nilai akurasi data sentimen media sosial x bahasa Indonesia mengenai ukt menggunakan algoritma random forest.

Algoritma Data Mining	Split Data (75% : 25%)	Cross Validation K-fold = 10
Random Forest	73%	75%
Random Forest + Hyperparameter Gridsearch	74%	89%

Tabel 4. Parameter gridsearch

Dari hasil perbandingan tabel 4 diketahui untuk akurasi data analisis sentimen menggunakan algoritma random forest dengan pengujian split data sebesar 73%, dan hasil akurasi menggunakan cross validation dengan 10 k-fold sebesar 75%. Adapaun optimasi algoritma random forest dengan hyperparameter gridsearch berhasil meningkatkan nilai akurasi yaitu untuk evaluasi dengan split data meningkat menjadi 74%, dan evaluasi menggunakan cross validation dengan 10 k-fold meningkat cukup besar menjadi 89%.

KESIMPULAN

Analisis sentimen media sosial x dengan bahasa Indonesia dalam mengklasifikasikan sentimen positif dan negatif dari hasil lebeling otomatis mengggunakan textblob, memiliki nilai akurasi 73% dari hasil penghitungan algoritma random forest dengan pembagian data latih 75% dan data uji 25%, semtara hasil cross validation menggunakan 10 k-fold adalah sebesar 75%, vektorisasi data yang digunakan yaitu tf-idf dangan melakukan imbalance data menggunakan metode smote.

Adapun hasil akurasi algoritma random forest menggunakan hyperparameter gridsearch berhasil meningkatkan nilai akurasi sebesar 1% menjadi 74%, sementara evaluasi dengan cross validation menggunakan 10 k-fold mengalami peningkatan sebesar 14% sehingga nilai akurasi datanya menjadi 89%.

Selanjutnya penelitian ini dapat dilakukan dengan vektorisasi lain seperti BOW dengan algoritma klasifikasi seperti SVM, Decision Tree dan lain sebagainya, serta dilakukan labeling otomatis dengan menggunkan Lexicon base bahasa Indonesia untuk mendapatkan hasil akurasi yang lebih baik lagi.

DAFTAR PUSTAKA

[1] S. H. Fikri, W. R. W. R. Panji, and E. L. Fitriyah, "Urgensi Pelaksanaan Pendidikan Karakter Yang Terintegrasi: Analisis Kebijakan Penguatan Pendidikan Karakter," *Indonesian Journal of Educational Management and Leadership*, vol. 1, no. 1, pp. 45–56, 2023, doi: 10.51214/ijemal.v1i1.485.

[2] C. Suhaeni and H.-S. Yong, "Mitigating Class Imbalance in Sentiment Analysis Through GPT-3-Generated Synthetic Sentences," *Applied Sciences*, vol. 13, no. 17, p. 9766, 2023, doi: 10.3390/app13179766.

[3] R. Obiedat *et al.*, "Sentiment Analysis of Customers' Reviews Using a Hybrid Evolutionary SVM-Based Approach in an Imbalanced Data Distribution," *Ieee Access*, vol. 10, pp. 22260–22273, 2022, doi: 10.1109/access.2022.3149482.

[4] H. Wen and J. Zhao, "Sentiment Analysis Model of Imbalanced Comment Texts Based on BiLSTM," 2023, doi: 10.21203/rs.3.rs-2434519/v1.

[5] L. Chen, S. Shang, and Y. Wang, "Cross-Lingual Sentiment Analysis With MultiEmo: Exploring Language-Agnostic Models for Emotion Recognition," 2024, doi: 10.20944/preprints202408.1639.v1.

[6] T. W. Purnomo and J. Sutopo, "Comparison of Pre-Trained Bert-Based Transformer Models for Regional Language Text Sentiment Analysis in Indonesia," *International Journal Science and Technology*, vol. 3, no. 3, pp. 11–21, 2024, doi: 10.56127/ijst.v3i3.1739.

[7] R. Kusumaningrum, I. Z. Nisa, R. Jayanto, R. P. Nawangsari, and A. Wibowo, "Deep Learning-Based Application for Multilevel Sentiment Analysis of Indonesian Hotel Reviews," *Heliyon*, vol. 9, no. 6, p. e17147, 2023, doi: 10.1016/j.heliyon.2023.e17147.

[8] A. H. Nasution and A. Onan, "ChatGPT Label: Comparing the Quality of Human-Generated and LLM-Generated Annotations in Low-Resource Language NLP Tasks," *Ieee Access*, vol. 12, pp. 71876–71900, 2024, doi: 10.1109/access.2024.3402809.

[9] F. Fathoni, E. Erwin, and A. Abdiansah, "Multilabel Sentiment Analysis for Classification of the Spread of COVID-19 in Indonesia Using Machine Learning," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 31, no. 2, p. 968, 2023, doi: 10.11591/ijeecs.v31.i2.pp968-978.

[10] L. Damayanti and K. M. Lhaksmana, "Sentiment Analysis of the 2024 Indonesia Presidential Election on Twitter," *Sinkron*, vol. 8, no. 2, pp. 938–946, 2024, doi: 10.33395/sinkron.v8i2.13379.

[11] M. B. Ressan and R. F. Hassan, "Naïve-Bayes Family for Sentiment Analysis During COVID-19 Pandemic and Classification Tweets," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 28, no. 1, p. 375, 2022, doi: 10.11591/ijeecs.v28.i1.pp375-383.

[12] A. Romadhony, S. A. Faraby, R. Rismala, U. N. Wisesty, and A. Arifianto, "Sentiment Analysis on a Large Indonesian Product Review Dataset," *Journal of Information Systems Engineering and Business Intelligence*, vol. 10, no. 1, pp. 167–178, 2024, doi: 10.20473/jisebi.10.1.167-178.

[13] V. Vinardo and I. Wasito, "Two-Stage Sentiment Analysis on Indonesian Online News Using Lexicon-Based," *Sinkron*, vol. 8, no. 4, pp. 2109–2119, 2023, doi: 10.33395/sinkron.v8i4.12769.

[14] M. A. W. Sinaga, N. F. Nuzula, and C. R. Damayanti, "The Psychology of Risk Influence and Investor Sentiment on Investment Decision Making in the Indonesian Stock Market," *Jurnal Ilmiah Akuntansi Dan Bisnis*, vol. 18, no. 2, p. 197, 2023, doi: 10.24843/jiab.2023.v18.i02.p01.

[15] A. Ardisurya and M. Rizkinia, "Implementation of Diffusion Variational Autoencoder for Stock Price Prediction With the Integration of Historical and Market Sentiment Data," *Ijecbe*, vol. 2, no. 2, 2024, doi: 10.62146/ijecbe.v2i2.55.

[16] H. Sujadi, "Analisis Sentimen Pengguna Media Sosial Twitter Terhadap Wabah Covid-19 Dengan Metode Naive Bayes Classifier Dan Support Vector Machine," *Infotech Journal*, vol. 8, no. 1, pp. 22–27, 2022, doi: 10.31949/infotech.v8i1.1883.

[17] E. Hasibuan and E. A. Heriyanto, "Analisis Sentimen Pada Ulasan Aplikasi Amazon Shopping Di Google Play Store Menggunakan Naive Bayes Classifier," *Jurnal Teknik Dan Science*, vol. 1, no. 3, pp. 13–24, 2022, doi: 10.56127/jts.v1i3.434.

[18] D. Atmajaya, A. Febrianti, and H. Darwis, "Metode SVM Dan Naive Bayes Untuk Analisis Sentimen ChatGPT Di Twitter," *Indonesian Journal of Computer Science*, vol. 12, no. 4, 2023, doi: 10.33022/ijcs.v12i4.3341.

[19] E. Eviyanti, B. Irawan, and A. Bahtiar, "Penggunaan Algoritma Naïve Bayes Dalam Menganalisis Sentimen Ulasan Aplikasi Adakami Di Google Play Store," *Jati (Jurnal Mahasiswa Teknik Informatika)*, vol. 7, no. 6, pp. 3879–3885, 2024, doi: 10.36040/jati.v7i6.8272.

[20] A. I. Tanggraeni and M. N. N. Sitokdana, "Analisis Sentimen Aplikasi E-Government Pada Google Play Menggunakan Algoritma Naïve Bayes," *Jatisi (Jurnal Teknik Informatika Dan Sistem Informasi)*, vol. 9, no. 2, pp. 785–795, 2022, doi: 10.35957/jatisi.v9i2.1835.

[21] J.-H. Wang, C. Liu, Y.-R. Min, Z.-H. Wu, and P.-L. Hou, "Cancer Diagnosis by Gene-Environment Interactions via Combination of SMOTE-Tomek and Overlapped Group Screening Approaches With Application to Imbalanced TCGA Clinical and Genomic Data," *Mathematics*, vol. 12, no. 14, p. 2209, 2024, doi: 10.3390/math12142209.