

Klasifikasi Ulasan Konsumen Menggunakan *Random Forest* dan SMOTE

¹Nurul Istiqamah, ²Muhammad Rijal

^{1,2}Institut Teknologi dan Bisnis NOBEL Indonesia

Email: ¹nrlistiqamah127@gmail.com, ²rijal2303@gmail.com

Abstrak

Ulasan berupa deskripsi teks direpresentasikan dalam bentuk skala peringkat tidaklah cukup sebagai acuan dalam menentukan sentimen yang seringkali bias. Penelitian ini dilakukan bertujuan untuk mengklasifikasi ulasan konsumen pada *e-WOM* dengan menerapkan teknik klasifikasi sentimen pada tingkat dokumen. Penelitian ini berfokus pada *imbalanced class* klasifikasi ulasan konsumen. Metode yang digunakan dalam mengatasi permasalahan *imbalanced class* adalah kombinasi dari dua metode, yaitu *Synthetic Minority Oversampling Technique* (SMOTE) dan *Random Under-sampling* (RUS) pada tahap *pre-processing* dan pada tahap klasifikasi menggunakan *Random Forest*. Fitur data yang digunakan adalah fitur teks ulasan (proses analisis dan klasifikasi sentimen) dan fitur peringkat (proses pelabelan). Hasil penelitian setelah menggunakan kombinasi metode SMOTE dan RUS menunjukkan akurasi sebesar 75% dan menunjukkan peningkatan kembali dengan 8000 *max_features* menjadi 77%. Penerapan teknik *class imbalanced* (dalam hal ini SMOTE dan RUS) pada tahapan *pre-processing* menunjukkan bahwa kombinasi metode ini mampu memberikan perubahan peningkatan akurasi dan mengenali data yang awalnya dianggap minor.

Kata Kunci: *imbalanced class*, klasifikasi sentimen, *Random Forest*, RUS, SMOTE

Abstract

Reviews in the form of text descriptions represented in the form of a rating scale are not sufficient as a reference in determining sentiment which is often biased. This research was conducted with the aim of classifying consumer reviews on e-WOM by applying sentiment classification techniques at the document level. This research focuses on the imbalanced class classification of consumer reviews. The method used to overcome the imbalanced class problem is a combination of two methods, namely Synthetic Minority Oversampling Technique (SMOTE) and Random Under-sampling (RUS) at the pre-processing stage and for classification stage using Random Forest. The data features used are review text features (sentiment analysis and classification process) and ranking features (labeling process). The research results after using a combination of the SMOTE and RUS methods showed an accuracy of 75% and showed an increase again with 8000 max_features to 77%. The application of the class imbalanced technique (in this case SMOTE and RUS) at the pre-processing stage shows that this combination of methods is able to provide changes in increasing accuracy and recognizing data that was initially considered minor.

Keywords: imbalanced class, sentiment classification, Random Forest, RUS, SMOTE

1. Pendahuluan

e-WOM merupakan suatu media komunikasi yang berisi kumpulan opini (berbentuk ulasan, forum, maupun mikro blog) terkait suatu entitas tertentu seperti produk, jasa, isu sosial bahkan tokoh (Li dkk., 2018). Bagi pelaku bisnis, *e-WOM* berupa ulasan dapat membantu mereka dalam melakukan riset pasar baik terkait produk maupun terkait karakteristik kebiasaan konsumen. Ulasan yang berbentuk deskripsi teks juga direpresentasikan dalam bentuk skala peringkat. Skala peringkat ini dijadikan sebagai acuan dalam menentukan sentimen. Hal itu tidaklah cukup dan seringkali terjadi kebiasaan di antaranya (Tama dkk., 2019). Hal ini bisa teratasi dengan menerapkan salah satu teknik analisis sentimen yaitu klasifikasi sentimen. Klasifikasi sentimen sendiri terdiri dari tiga tingkatan yaitu tingkat dokumen, tingkat kalimat, hingga tingkat aspek (Liu, 2015). Pada penelitian ini, analisis sentimen yang digunakan adalah analisis tingkat dokumen yang digunakan pada tipe dataset *review* produk. Klasifikasi sentimen pada tingkat dokumen membahas mengenai bagaimana menampilkan sebuah sentimen pada seluruh dokumen menggunakan model yang tepat. Pada penelitian ini metode yang digunakan adalah algoritma *Random Forest classifier* dan *Synthetic minority oversampling technique* (SMOTE) dengan kolaborasi *Random Undersampling* (RUS) yang diterapkan untuk meningkatkan performa klasifikasi. Penelitian pada *imbalanced clas* menggunakan SMOTE telah dilakukan oleh Utari dkk. (2020), Mustaqim dkk. (2019), dan Umer dkk. (2021) menghasilkan peningkatan pada akurasi menjadi lebih dari 90% (Utari dkk., 2020; Mustaqim dkk., 2019; Umer dkk., 2021). Penelitian pada *imbalanced class* menggunakan RUS juga telah dilakukan oleh Rafrastara dkk. (2023) dan penelitian perbandingan kedua metode SMOTE dan RUS oleh Vinodhini dan Chandrasekaran (2017) (Rafrastara dkk., 2023; Vinodhini dan Chandrasekaran, 2017). Pada penelitian oleh Vinodhini dan Chandrasekaran (2017) menghasilkan bahwa kinerja SMOTE lebih baik pada rasio ketidakseimbangan yang lebih tinggi sedangkan kinerja RUS memiliki lebih baik pada rasio ketidakseimbangan yang lebih rendah (Vinodhini dan Chandrasekaran, 2017).

Pada penelitian ini dilakukan kolaborasi kedua metode tersebut (SMOTE dan RUS) untuk permasalahan *imbalanced class* dan meningkatkan kinerja pada tahap *pre-processing*. Hasil kinerja kolaborasi metode tersebut kemudian digunakan untuk tahap klasifikasi sentimen yang menggunakan algoritma *Random Forest*. Penelitian telah dilakukan oleh Bahwari tahun 2019 menggunakan *Random Forest* pada klasifikasi analisis sentimen terhadap ulasan di media sosial Twitter mengenai 6 perusahaan penerbangan menghasilkan akurasi sebesar 75% (Bahwari, 2019). Penelitian lain juga telah dilakukan oleh Neogi dkk. pada tahun 2021 yang membandingkan *Random Forest* dengan 3 algoritma lainnya menghasilkan bahwa *Random Forest* memiliki akurasi tertinggi (Neogi dkk., 2021). Hasil dari penelitian ini adalah tingkat akurasi dari metode yang telah diusulkan.

Dalam penelitian ini dibahas mengenai metode yang digunakan dalam penelitian pada bagian 2. Pembahasan mengenai hasil dan pembahasan penelitian ada pada bagian 3. Pembahasan kesimpulan dari penelitian ada pada bagian 4.

2. Metode Penelitian

2.1. Data Penelitian

Pada penelitian ini dataset yang digunakan adalah dataset ulasan pelanggan yang berasal dari *situsreviews.femaledaily.com* yang terdiri dari 11 fitur. Kesebelas fitur

tersebut antara lain, *product_brand*, *product_category*, *product_id*, *product_name*, *product_type*, *review_date*, *review_id*, *review_packaging*, *review_rating*, *review_rebuy*, *review_text*. Fitur yang digunakan sebagai data untuk proses analisis dan klasifikasi sentiment yaitu fitur teks ulasan (*review_text*) dan proses pelabelan menggunakan fitur peringkat (*review_rating*). Penggunaan dataset ulasan produk dari situs *reviews.femaledaily.com* ini berjumlah 14.999 data yang disimpan dalam format tabel (.xlsx) dan digunakan sebagai data inputan.

2.2. *Sentiment analysis (SA)*

Sentiment analysis (SA) merupakan salah satu dari beberapa sub-bidang studi komputasi yakni *Natural Language Processing (NLP)*. SA umumnya sebuah teknik dalam penambangan informasi berupa ekstraksi emosi berdasarkan opini atau sentimen (ulasan) (Liu, 2012). Analisis sentimen atau penambangan opini memiliki manfaat pada setiap tingkatan salah satunya pada tingkat dokumen. Tingkat dokumen ini membahas tentang bagaimana menampilkan sebuah sentimen pada keseluruhan dokumen menggunakan model yang tepat. Serangkaian proses SA ada beberapa variasi tugas dan salah satunya adalah *sentiment classification*. *Sentiment classification* ini melakukan pengelompokan terhadap dua kelas yakni positif dan negatif (Liu, 2012; Saberi dan Saad, 2017).

2.3. *Text Pre-processing*

Text Preprocessing adalah tahapan proses membersihkan dataset sentimen menjadi teks terstruktur yang siap untuk diolah. Proses ini dapat meningkatkan kinerja dan akurasi pada model klasifikasi (Shaheen, 2019). Penjelasan tahapan pada *preprocessing* adalah sebagai berikut:

1. *Case Folding*

Tahap ini adalah mengubah semua huruf dalam dokumen menjadi bentuk setara (*lowercase*) sedangkan karakter yang tidak termasuk dalam huruf 'a' sampai 'z', angka, dan tanda baca akan dianggap *delimiter* yang harus dihilangkan (Nugroho, 2019).

2. *Tokenization*

Tahap ini adalah memisahkan setiap kata yang menyusun sebuah dokumen sebagai sebuah token. Proses ini dilakukan pada paragraf ataupun kalimat sesuai dengan kebutuhan (Nugroho, 2019).

3. *Stopword Removal*

Tahap ini adalah proses filterisasi. *Stopword Removal* adalah tahap menghapus kata yang tidak penting atau tidak bermakna dari proses *Tokenization* (Nugroho, 2019).

4. *Text Normalization*

Tahap ini adalah proses normalisasi kata sesuai standar penulisan sehingga kata-kata tersebut bisa diolah.

2.4. *Term Frequency-Inverse Document Frequency (TF-IDF)*

TF-IDF adalah metode dalam proses pembobotan fitur kata berdasarkan frekuensi kemunculan kata pada suatu dokumen. TF (*Term Frequency*) merepresentasikan frekuensi kemunculan kata pada suatu dokumen sedangkan IDF (*Inverse Document*

Frequency) membantu dalam mengetahui apakah *term* yang dicari telah sesuai dengan kata kunci yang diharapkan. Berikut untuk persamaannya (Fitriyah, 2020).

$$tfidf(t, d, D) = tf(t, d) * idf(t, D) \quad (1)$$

Di mana :

$tf(t, d)$ = berapa kali *term* 't' muncul dalam dokumen 'd'

$idf(t, D)$ = *term* 't' dalam kumpulan dokumen keseluruhan 'D'

Untuk memperoleh hasil dari persamaan ' $idf(t, D)$ ', maka digunakan formula sebagai berikut :

$$idf(t, D) = \log \frac{N}{|d \in D: t \in d|} \quad (2)$$

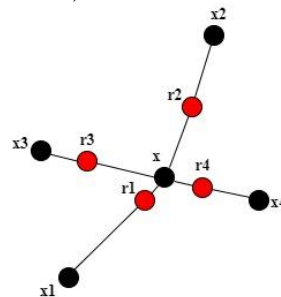
Di mana :

N = total dokumen yang ada di dalam corpus D

$|d \in D: t \in d|$ = jumlah dokumen dimana *term* 't' muncul.

2.5. Synthetic Minority Oversampling Technique (SMOTE)

SMOTE merupakan salah satu teknik yang digunakan dalam mengatasi permasalahan *class imbalanced* terkait sampel minoritas. SMOTE ini mampu menyeimbangkan antara data minor dan data mayor dan mampu mengurangi masalah *overfitting* pada proses *learning* (Mustaqim dkk., 2019). Proses pembuatan data sintetik pada teknik ini yakni dengan melakukan interpolasi antara sampel minor dengan k-tetangga terdekat (*k-Nearest Neighbor*) hingga seimbang dengan sampel mayor (Soltanzadeh and Hashemzadeh, 2021).

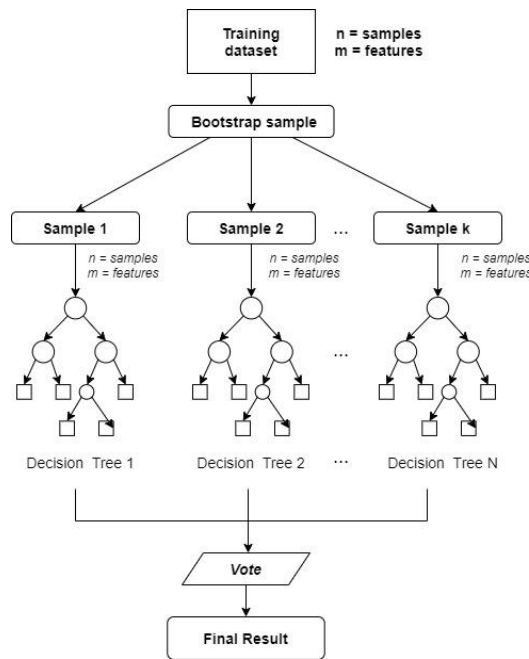


Gambar 1. Sampel Data Sintetik Hasil SMOTE (Qu dkk., 2020)

Salah satu kunci keberhasilan SMOTE dalam menghadapi *class imbalanced* adalah tergantung pada kemampuan algoritma ini mengikuti aturan distribusi sampel yang dekat dengan distribusi sebenarnya.

2.6. Random Forest Classifier

Algoritma ini merupakan hasil dari pengembangan klasifikasi dasar *Decision Tree* yang mampu beroperasi secara acak dengan memilih subset sampel dan subset fitur dalam menjamin independensi pada pohon keputusan serta generalisasi yang lebih baik (Parmar dkk., 2019). *Random Forest* memilih pilihan terbanyak sebagai hasil prediksi dan memiliki performa yang baik dalam mengatasi *overfitting* (Bahwari, 2019).



Gambar 2. Ilustrasi *Random Forest Classifier* (Park dkk)

Pada RF, setiap pemilihan *tree* sangat menentukan hasil prediksi yang dihasilkan pada klasifikasi karena semakin banyak *decision tree* dengan voting terbanyak maka semakin mudah mesin dalam menentukan kelas target (Breiman, 2001).

2.7. Confusion Matrix

Confusion Matrix (CM) merupakan suatu alat dalam mengevaluasi model pada *supervised learning* yaitu algoritma klasifikasi (Bekkar dkk., 2013).

		Actual Values	
		1 (Positive)	0 (Negative)
Predictive Values	1 (Positive)	TP True Positive	FP False Positive Type I Error
	0 (Negative)	FN False Negative Type II Error	TN True Negative

Gambar 3. *Confusion Matrix* Untuk Klasifikasi Dua Kelas

Pada Gambar 3, terdapat akronim yang mewakili setiap hasil kelas pada proses klasifikasi antara lain TP, TN, FP dan FN. Untuk menentukan persamaan nilai akurasi, presisi, dan recall berikut rumus dasar nilai akurasi pada metode evaluasi matriks :

$$Akurasi = \frac{TP + TN}{TP + TN + FP + FN} \quad (3)$$

$$Sensitivitas = recall = TP_{rate} = \frac{TP}{TP + FN} \quad (4)$$

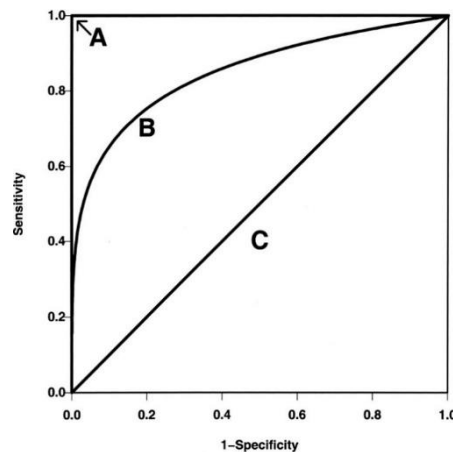
$$Presisi = \frac{TP}{TP+FP} \quad (5)$$

Berikut maksud dari beberapa akronim terkait :

- TP = *True Positive*, dimana data kasus positif yang diprediksi benar.
- TN = *True Negative*, dimana data kasus negatif yang diprediksi benar.
- FP = *False Positive*, dimana data kasus negatif namun diprediksi sebagai data positif.
- FN = *False Negatif*, dimana data kasus positif namun diprediksi sebagai data negatif.

2.8. AUC ROC

AUC (*Area Under the ROC Curve*) dinilai tepat untuk kasus masalah tidak seimbang sebab AUC mampu mengevaluasi prediktor secara komprehensif (Zhang and Wang, 2011) serta mampu menilai model mana yang lebih baik secara rata-rata dalam bentuk grafik kurva. Evaluasi AUC (*Area Under the ROC Curve*) juga merupakan salah satu metode evaluasi yang populer digunakan pada klasifikasi dengan kasus data tidak seimbang (Saifudin dan Wahono, 2015).



Gambar 4. Bentuk Penyajian Analisis Kurva ROC (Zou dkk., 2007)

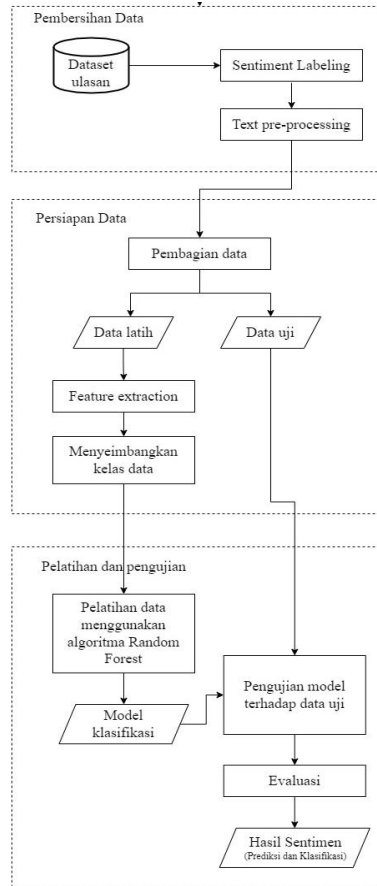
Terdapat tiga hipotesis kurva yang masing-masing merepresentasikan analisis akurasi. Simbol huruf 'A' mewakili area dengan hasil akurasi terbaik (di atas standar) yakni AUC = 1. Simbol huruf 'B' merepresentasikan bentuk kurva ROC di mana AUC = 0.85, dan garis diagonal atau simbol huruf 'C' yang mewakili *random classifier* yang sesuai. Salah satu tanda adanya peningkatan pada hasil analisis adalah ketika AUC mendekati angka 1 atau kurva ROC bergerak ke arah simbol A. (Zou et al., 2007).

3. Hasil dan Pembahasan

3.1. Kerangka Penelitian

Penelitian ini terdiri dari 3 tahapan yang terdiri dari Pembersihan Data, Persiapan Data, dan Pelatihan dan Pengujian. Pada tahap pertama yaitu Pembersihan Data, dilakukan proses *Sentiment Labeling* dan *Text Pre-processing*. Tahap kedua yaitu Persiapan data dengan dilakukan proses Pembagian Data. Pada tahap kedua ini, data akan dibagi menjadi 2, yaitu Data Latih dan Data Uji. Data latih diproses terlebih dahulu untuk mengekstrak fitur (*feature extraction*) menggunakan TF-IDF kemudian dilakukan proses penyeimbangan kelas data (*class imbalanced*) menggunakan SMOTE dan RUS. Tahap

ketiga yaitu Pelatihan dan Pengujian. Pada tahap ketiga ini hasil output data latih dari tahap kedua menjadi input tahap ketiga menggunakan *Random Forest* yang kemudian dilakukan pengujian dengan data uji. Hasil dari tahap ketiga ini adalah nilai akurasi dari klasifikasi. Untuk lebih jelasnya ada pada gambar berikut ini.



Gambar 5. Kerangka Penelitian

3.1.1. Pembersihan Dataset

Tahap ini dilakukan pembersihan dataset dari data *noise* dan mengurangi dimensi pada dataset berbasis teks ke dalam bentuk token. Pada data informasi mentah memiliki 11 fitur dan hanya 9 fitur yang digunakan. Fitur-fitur tersebut adalah *product brand*, *product category*, *product id*, *product name*, *product type*, *review date*, *review id*, *review rating*, dan *review text*. Fitur yang digunakan dalam proses analisis sentimen yakni fitur data teks ulasan sedangkan fitur peringkat digunakan dalam proses pelabelan. Data informasi yang dimiliki selanjutnya diubah ke dalam format tabel (.xlsx) dengan jumlah data sebanyak kurang lebih 14.999 baris data.

1. *Sentiment Labeling*

Proses pelabelan dilakukan karena dataset belum memiliki kelas target yang berupa polarisasi positif maupun negatif. Proses pelabelan sentimen dilakukan dengan menggunakan fitur peringkat sebagai acuan untuk menentukan pola sentimen positif dan negatif.

Tabel 1. Indikator Pelabelan Kelas (*Threshold*)

Peringkat	Sentiment
-----------	-----------

4-5	Positif
1-2	Negatif

Tabel 1. merupakan indikator dalam menentukan pelabelan awal sebelum dilakukan *pre-processing* dan pembuatan model. Hasil dari proses pelabelan terbentuk satu kolom fitur baru yakni fitur sentimen. Fitur sentimen tersebut selanjutnya digunakan sebagai data kelas (target).

2. Text Pre-processing

Dataset berupa teks memiliki bentuk yang sangat tidak terstruktur sehingga dibutuhkan usaha lebih dalam proses *text pre-processing*. Dataset ulasan pada penelitian ini menggunakan campuran bahasan Indonesia dan bahasa Inggris. Penelitian ini fokus pada penggunaan bahasa Indonesia yang dinormalisasi dari bahasa gaul ke bahasa baku. Berikut beberapa tahapan prosesnya:

1. Case Folding

Proses ini mengubah huruf menjadi kecil (*lowercase*) dan karakter angka, tanda baca, karakter kosong (*spacing*) dihilangkan. Berikut ilustrasinya:

2. Text Normalization

Text Normalization merupakan tahapan yang dilakukan untuk mengubah kata-kata informal menjadi kata yang sesuai dengan standar penulisan. Berikut ilustrasinya:

Tabel 2. Ilustrasi Hasil *Text Normalization*

Masukan
“suka banget sama lippenya wnw ini warnanya ok banget dan murah finishnya matte tapi ga bikin bibir kering dan ga bikin bibir terlihat pecah aplikasi pun gampang banget cuman ga suka sama bentuk pekejingnya karena harus super duper hati kalau buka dan tutup karena lipsticknya ga masuk semua ke tube dan punyaku ada yang bpmpal karena ga hati waktu nutup”
Keluaran
“suka sangat sama lippenya wnw ini warnanya ok sangat dan murah akhir matte tetapi tidak buat bibi kering dan tidak buat bibir terlihat pecah aplikasi pun mudah sangat hanya tidak suka sama bentuk kemasan karena harus super sangat hati kalau buka dan tutup karena lipsticknya tidak masuk semua ke tube dan punyaku ada yang karena tidak hati waktu tutup”

3. Stopwods Removal

Stopword Removal merupakan proses pembuangan kata-kata yang termasuk ke dalam kategori *stopword* seperti ini, itu, yang, ke, di, dalam, dan lain sebagainya. Berikut ilustrasinya:

Tabel 3. Ilustrasi Hasil *Stopword Removal*

Masukan
“suka sangat sama lippenya wnw ini warnanya ok sangat dan murah akhir matte tetapi tidak buat bibi kering dan tidak buat bibir terlihat pecah aplikasi pun mudah sangat hanya tidak suka sama bentuk kemasan karena harus super sangat hati kalau buka dan tutup karena lipsticknya tidak masuk semua ke tube dan punyaku ada yang karena tidak hati waktu tutup”
Keluaran
“suka sangat lippenya wnw warnanya sangat murah finishnya matte bibir kering bibir terlihat pecah aplikasi mudah sangat suka bentuk kemasan super sangat hati buka tutup lipsticknya masuk semua tube punyaku hati tutup”

4. Tokenization

Proses tokenisasi atau pemisahan kata terjadi ketika suatu kata terpisahkan dengan kata yang lain maupun dengan adanya karakter spasi diantara kata. Berikut ilustrasinya:

Tabel 4. Ilustrasi Hasil *Tokenization*

Masukan
“suka sangat sama lippennya wnw warnanya ok sangat murah finishnya matte bibir kering bibir terlihat pecah aplikasi mudah sangat suka bentuk kemasan super sangat hati buka tutup lipsticknya masuk semua tube punyaku bpmpal hati tutup”
Keluaran
['suka', 'sangat', 'sama', 'lippennya', 'wnw', 'warnanya', 'sangat', 'murah', 'finishnya', 'matte', 'bibir', 'kering', 'bibir', 'terlihat', 'pecah', 'aplikasi', 'gampang', 'sangat', 'suka', 'bentuk', 'kemasan', 'super', 'sangat', 'hati', 'buka', 'tutup', 'lipsticknya', 'masuk', 'semua', 'tube', 'punyaku', 'bpmpal', 'hati', 'tutup']

3.1.2. Persiapan Data

Pada tahap persiapan data, dilakukan pembagian data menjadi 2 bagian, yaitu data latih (*training data*) dan data uji (*testing data*). Pada penelitian ini, dari 14.999 data, data dibagi dengan perbandingan 70:30 persen. Data sebanyak 70 persen atau berjumlah 10.500 data digunakan dalam proses pelatihan dan 30 persen atau berjumlah 4.500 data digunakan guna proses pengujian berupa data yang tidak pernah digunakan sebelumnya pada proses pelatihan.

Data latih kemudian melalui proses *Feature Extraction* (ekstraksi fitur) atau vektorisasi menggunakan metode *Term Frequency-Inverse Document Frequency* (TF-IDF). Ekstraksi fitur adalah suatu proses yang digunakan untuk men ekstrak suatu dokumen ke dalam kumpulan vector yang akan memudahkan mesin untuk membacanya. Berikut sampelnya:

Tabel 5. Sampel Data Untuk Representasi Fitur Menggunakan TF-IDF

Dokumen yang telah melalui proses preprocessing	
D1	['suka', 'lippen', 'warna', 'murah', 'akhir', 'matte', 'bibir', 'kering', 'bibir', 'pecah', 'aplikasi', 'mudah', 'sangat', 'suka', 'bentuk', 'kemasan', 'super', 'lipstick']
D2	['suka', 'lipstik', 'lembab', 'kilap', 'warna', 'bronze', 'orange', 'cocok', 'kulit', 'cenderung', 'coklat', 'pakai', 'lip', 'balm', 'lembab']

Hasil ekstraksi fitur yang digunakan dalam proses pelatihan hanya sejumlah 8.000 fitur. Hal tersebut untuk meminimalisir peningkatan dimensi disebabkan representasi fitur dan dapat mengakibatkan proses eksekusi menjadi lambat.

Proses selanjutnya adalah proses menyeimbangkan kelas (*class imbalanced*) data. Proses ini bertujuan untuk lebih meningkatkan performa model yang dibuat. Proses ini menggunakan SMOTE dan RUS.

3.1.3. Pelatihan dan Pengujian

Data latih yang sudah melalui proses penyeimbangan kelas data kemudian masuk ke proses berikutnya yaitu proses pelatihan menggunakan *Random Forest* untuk diklasifikasikan berdasarkan hasil pemilihan pohon keputusan. Tahap pengujian menggunakan dua metode yaitu *Confusion Matrix* (CF) untuk menilai seberapa baik tingkat akurasi yang dihasilkan oleh model dalam mengklasifikasi berdasarkan kelas dan metode evaluasi AUC-ROC dimaksudkan untuk menilai kinerja prediktor secara keseluruhan berdasarkan hasil rata dari perhitungan *False Positive Rate* (FPR) dengan *True Positive Rate*.

3.2. Hasil dan Pembahasan

Penelitian ini dilakukan dua skenario pengujian. Skenario pengujian pertama adalah menggunakan RF tanpa teknik *class imbalanced*, RF dengan SMOTE, dan RF dengan kombinasi dua metode, yaitu SMOTE dan RUS. Skenario pengujian pertama ini bertujuan untuk memaksimalkan model dalam mengenali atau memprediksi sentimen negatif. Adapun hasil yang diperoleh adalah sebagai berikut:

Tabel 6. Hasil Skenario Pengujian 1 Menggunakan *Confusion Matrix*

Model	Akurasi	Recall	Presisi
RF	73%	73%	98%
RF+SMOTE	72%	75%	89%
RF+SMOTE+RUS	75%	78%	89%

Berdasarkan tabel di atas, terjadi peningkatan performa hampir di seluruh aspek penilaian menggunakan metode *confusion matrix*. Jika dilihat berdasarkan nilai akurasi, model RF dengan implementasi teknik *class imbalanced* kombinasi dua metode SMOTE dan RUS memberikan peningkatan sebesar 2% lebih baik dari pada performa tradisional RF dengan nilai akurasi menjadi 75%.

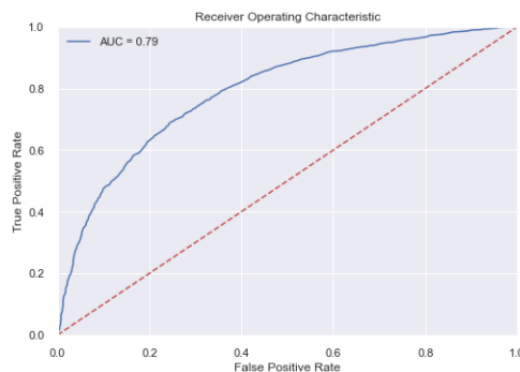
Skenario pengujian kedua untuk melihat hasil penerapan fitur vektor terhadap satu model terbaik yang terpilih yang berasal dari skenario pertama. Perbedaan dari skenario yang pertama adalah adanya penambahan parameter vektor, yaitu *max_features* yang nilainya berasal dari hasil pemilihan fitur *brute force*. Jumlah *max_features* yang digunakan adalah 8.000 dengan *n_estimator* sebanyak 500 *tree* pada model klasifikasi RF. Adapun hasil yang diperoleh dan dibandingkan oleh skenario pengujian pertama adalah sebagai berikut:

Tabel 7. Komparasi Hasil Pengujian 1 dan 2 dengan *Confusion Matrix*

Model	Akurasi
RF + SMOTE + RUS	75%
RF + SMOTE + RUS Dengan 8000 <i>max_features</i>	77%

Berdasarkan Tabel 8 jika dibandingkan dengan model pada skenario pertama, model dengan penambahan teknik kelas *imbalanced* mampu memberikan peningkatan dibandingkan dengan model tanpa teknik kelas *imbalanced*. Pada Tabel 8, menunjukkan bahwa peningkatan hasil akurasi dengan penambahan parameter vektor *max_features* dengan nilai hasil pemilihan fitur *bruto force* dengan 8000 *max_feature* yang semula sebesar 75% menjadi 77% dan hal itu menunjukkan peningkatan sebesar 2% nilai akurasi.

Pada evaluasi ROC AUC, hasil data yang digunakan bukan data prediksi kelas tetapi menggunakan data hasil probabilitas yang diprediksi. Hasil dari evaluasi ROC AUC sebagai berikut:



Gambar 6. Kurva Hasil Evaluasi Model Menggunakan ROC

Berdasarkan pada hasil yang disajikan pada Gambar 6, nilai AUC yang diperoleh adalah 0,79 atau jika dibulatkan menjadi 0,80 dimana hal tersebut dapat disimpulkan bahwa hasil tersebut termasuk dalam kategori baik sebab nilai mendekati angka 1.

4. Kesimpulan

Berdasarkan hasil dari penelitian klasifikasi sentimen ulasan tingkat dokumen dari situs *reviews.femaledaily.com* menggunakan *Random Forest* dan teknik kelas data tidak seimbang (*imbalanced data*) dapat disimpulkan bahwa, penerapan teknik *class imbalanced* (dalam hal ini SMOTE dan RUS) pada tahapan *pre-processing* mampu memberikan perubahan peningkatan akurasi pada performa model kurang lebih sekitar 3% dan peningkatan skor AUC kurang lebih sekitar 4%. Berdasarkan hasil yang diperoleh. Meskipun hasil akurasi tidak tinggi namun model ini mampu mengenali data yang awalnya dianggap minor. Adapun saran untuk pengembangan selanjutnya diharapkan adanya peningkatan pada segi klasifikasi multi bahasa dan teknik *class imbalanced* lainnya.

Daftar Pustaka

- Bahwari, 2019, Sentiment Analysis Using Random Forest Algorithm-Online Social Media Based, *Journal Of Information Technology and Its Utilization* 2 (2), 29–33.
- Bekkar, M., Djemaa, H dan Alitouch, T., 2013, Evaluation Measures for Models Assessment over Imbalanced Data Sets, *Journal of Information Engineering and Applications* 3 (10), 27–38.
- Breiman, L., 2001, Random Forests, *Machine Learning* 45 (1), 5–32.
- Fitriyah, N., Warsito, B., Maruddani, D. A. I., 2020, Analisis Sentimen Gojek Pada Media Sosial Twitter Dengan Klasifikasi Support Vector Machine (SVM), *Jurnal Gaussian* 9 (3), 376–390.
- Li, Y., Guo, H., Zhang, Q., Gu, M dan Yang., 2018, Imbalanced Text Sentiment Classification Using Universal and Domain-Specific Knowledge, *Knowledge-Based Systems* 160 (June), 1–15.
- Liu, B., 2015, *Sentiment Analysis*, Cambridge University Press, Cambridge.
- Liu, B., 2012, Sentiment Analysis and Opinion Mining, *Synthesis Lectures on Human Language Technologies* 5 (1), 1–167.
- Mustaqim, M., Warsito, B dan Surarso, B., 2019, Kombinasi Synthetic Minority Oversampling Technique (SMOTE) Dan Neural Network Backpropagation Untuk Menangani Data Tidak Seimbang Pada Prediksi Pemakaian Alat Kontrasepsi Implan, *Jurnal Ilmiah Teknologi Sistem Informasi* 5 (34), 116–127.
- Neogi, A.S., Garg, K. A., Mishra, R. K., Dwivedi, Y. K., 2021, Sentiment Analysis and Classification of Indian Farmers' Protest Using Twitter Data, *International Journal of Information Management Data Insights* 1 (2), 100019.
- Nugroho, K.S., 2019, *Dasar Text Preprocessing Dengan Python*, *medium.com*.
- Park, M., Jung, D., Lee, S dan Park, S., 2020, Heatwave Damage Prediction Using Random Forest Model in Korea, *Applied sciences* 10 (22), 1–12.
- Parmar, A., Kataruya, R dan Petal, V., 2019, A Review on Random Forest: An Ensemble Classifier, *Lecture Notes on Data Engineering and Communications Technologies*

- 26, 758–763.
- Qu, Z., Li, H., Wang, Y., Zhang, J., Abu-Siada, A., & Yao, Y., 2020, Detection of Electricity Theft Behavior Based on Technique and Random Forest Classifier, *Energies* 13 (8), 2039.
- Rafrastara, F. A., Supriyanto, C., Paramita, C., Astuti, Y. P., & Ahmed, F. (2023). Performance Improvement of Random Forest Algorithm for Malware Detection on Imbalanced Dataset using Random Under-Sampling Method. *Jurnal Informatika: Jurnal Pengembangan IT*, 8(2), 113-118.
- Saberi, B. dan Saad, S., 2017, Sentiment Analysis or Opinion Mining: A Review, *International Journal on Advanced Science, Engineering and Information Technology* 7 (5), 1660–1666.
- Saifudin, A. dan Wahono, S., 2015, Pendekatan Level Data Untuk Menangani Ketidakseimbangan Kelas Pada Prediksi Cacat Software, *Journal of Software Engineering* 1 (2), 76–85.
- Shaheen, M., 2019, Sentiment Analysis on Mobile Phone Reviews Using Supervised Learning Techniques, *International Journal of Modern Education and Computer Science* 11 (7), 32–43.
- Soltanzadeh, P. dan Hashemzadeh, M., 2021, RCSMOTE: Range-Controlled Synthetic Minority over-Sampling Technique for Handling the Class Imbalance Problem, *Information Sciences* 542, 92–111.
- Tama, V.O., Sibaroni, Y dan Adiwijaya., 2019, Labeling Analysis in the Classification of Product Review Sentiments by Using Multinomial Naive Bayes Algorithm, *Journal of Physics: Conference Series* 1192 (1).
- Umer, M., Sadiq, S., Missen, M., Hameed, Z., Siddique, M dan Nappi, M., 2021, Scientific Papers Citation Analysis Using Textual Features and SMOTE Resampling Techniques, *Pattern Recognition Letters* 150, 250–257.
- Utari, M., Warsito, B., Kusumaningrum, R., 2020, Implementation of Data Mining for Drop-Out Prediction Using Random Forest Method. In *2020 8th International Conference on Information and Communication Technology (ICoICT)*, IEEE, 1–5.
- Vinodhini, G. dan Chandrasekaran, R.M., 2017, A Sampling Based Sentiment Mining Approach for E-Commerce Applications, *Information Processing and Management* 53 (1), 223–236.
- Zhang, H. dan Wang, Z., 2011, A Normal Distribution-Based Over-Sampling Approach to Imbalanced Data Classification. In *ADMA*,
- Zou, K.H. O'Malley, J, A, dan Mauri, L., 2007, Receiver-Operating Characteristic Analysis for Evaluating Diagnostic Tests and Predictive Models, *Circulation* 115 (5), 654–657.