Sentiment Analysis on Twitter About Domestic Violence Using Random Forest and Extreme Gradient Boosting Methods
[Analisa Sentimen Pada Twitter Tentang Kekerasan Dalam Rumah Tangga Menggunakan Metode Random Forest dan Extreme Gradient Boosting]

Robi'atul Asyaroh[1)], Arif Senja Fitrani [*,2)]

[1)]Program Studi Informatika, Universitas Muhammadiyah Sidoarjo, Indonesia
[2)] Program Studi Informatika, Universitas Muhammadiyah Sidoarjo, Indonesia
*Email Penulis Korespondensi:asfjim@umsida.ac.id

Abstract. *Social media Twitter is one of the communication media that is in great demand by the public. Currently, the topic of domestic violence is being discussed by Twitter social media users. Users will provide comments and opinions about cases that were trending at the time, namely domestic violence via Twitter. In this study, researchers want to implement machine learning algorithms to perform sentiment analysis on Twitter users on the topic of domestic violence. After going through the text preprocessing stages, the researcher then used SMOTE for handling imbalanced data and TFIDF for word weighting. After that, classification was carried out using two machine learning algorithms, namely random forest and XGBoost. The results of this study obtained a train score of 94% and a test score of 76% for the Random Forest algorithm and a train score of 86% and a test score of 73% for XGBoost*.

Keywords - *Analysis Sentiment; KDRT; Random Forest; Twitter; XGBoost*

Abstrak. *Media social twitter menjadi salah satu media komunikasi yang tengah diminati oleh masyarakat. Saat ini topik KDRT tengah menjadi perbincangan oleh para pengguna media social twitter. Pengguna akan memberikan komentar dan opini tentang kasus yang sedang trending pada saat itu yaitu KDRT melalui twitter. Pada penelitian ini, peneliti ingin mengimplementasikan algoritma machine learning untuk melakukan Analisa sentiment pada pengguna twitter terhadap topik KDRT. Setelah melalui tahapan text preprocessing, selanjutnya peneliti menggunakan SMOTE untuk handling imbalanced data dan TFIDF untuk pembobotan kata. Setelah itu dilakukan klasifikasi dengan dua algoritma machine learning yaitu random forest dan XGBoost. Hasil dari penelitian ini mendapatkan skor train sebesar 97% dan skor tes 76% untuk algoritma Random Forest dan skor test 73% untuk XSBoost.*

Kata Kunci - *Analisa Sentimen; KDRT; Random Forest; Twitter; XGBoost.*

I. PENDAHULUAN

Media sosial Twitter menjadi salah satu pilihan utama masyarakat global dalam berkomunikasi, terbukti dari lonjakan jumlah pengguna Twitter secara global. Pada tahun 2016, Twitter berhasil mencatatkan sekitar 313 juta pengguna yang aktif setiap bulannya [1]. Individu akan berbagi informasi terkini atau pendapat mengenai isu-isu hangat global melalui media sosial Twitter. Isu-isu yang sedang tren dan mendapatkan banyak tanggapan dari pengguna dapat menghasilkan topik yang sedang populer di platform ini, dikenal dengan sebutan "trending topic".

Dengan meningkatnya jumlah pengguna Twitter, terjadi lonjakan dalam jumlah tweet yang diposting. Tweet-tweet tersebut berisi pandangan dan komentar publik yang beragam, meliputi bidang ekonomi, perilaku sosial, fenomena alam, perdagangan, pendidikan, hiburan, dan berbagai aspek lainnya. Salah satu isu yang menonjol dalam konteks perilaku sosial adalah kasus KDRT yang baru-baru ini mencuat. Pengguna Twitter aktif dalam memberikan komentar dan pandangan mengenai isu tersebut, menghasilkan diskusi yang ramai di platform ini.

Tweet-tweet yang dihasilkan oleh pengguna menjadi referensi bagi pengguna lainnya, memungkinkan berbagai sudut pandang dalam opini tentang kasus KDRT ini. Namun, tumpang tindihnya informasi dalam tweet-tweet yang tersebar secara acak kadang membuat pengguna kesulitan untuk memahami apakah opini yang disampaikan bersifat positif, negatif, atau netral.

Contoh penelitian analisis sentimen yang dilakukan oleh Alim Ikegami pada tahun 2022 dalam konteks Pemodelan Topik Ulasan Aplikasi di Google Play Store memberikan panduan. Dalam penelitian ini, analisis sentimen dan pemodelan topik digunakan untuk mengidentifikasi polaritas sentimen dan topik yang dibahas dalam setiap polaritas sentimen. Metode XGBoost dan Latent Dirichlet Allocation (LDA) digunakan dalam analisis ulasan yang terdapat di Google Play Store. Hasil analisis sentimen dari penelitian tersebut mencapai tingkat akurasi, presisi, recall, dan skor F1 sebesar 86,8%, 83,9%, 77,9%, dan 80,2%. Selanjutnya, pemodelan topik berhasil mengungkap masing-masing 3 dan 6 topik dalam ulasan yang memiliki polaritas sentimen positif dan negatif.f [1].

Pada tahun 2021, Rimbun Siringoringo melakukan penelitian dalam analisis sentimen yang berkaitan dengan segmentasi dan peramalan pasar. Dalam penelitian ini, metode XGBoost digunakan sebagai alat analisis utama. Untuk mengoptimalkan parameter XGBoost pada masalah segmentasi pasar, Genetic Algorithm (GA) digunakan sebagai pendekatan dalam mencari nilai optimal hyperparameter. Evaluasi performa model dilakukan dengan memanfaatkan kurva Receiver Operating Characteristic (ROC). Hasil pengujian penelitian ini menunjukkan hasil pengujian ROC untuk beberapa model, yaitu SVM, Logistic Regression, dan Genetic-XGBoost masing-masing memiliki nilai sebesar 0,89; 0,98; 0,99. Dari hasil ini dapat disimpulkan bahwa model Genetic-XGBoost memiliki kinerja yang sangat baik dalam konteks segmentasi dan peramalan pasar. Hal ini menunjukkan bahwa pendekatan ini dapat berhasil diterapkan dalam mengatasi tantangan dalam analisis sentimen pada bidang tersebut [2].

Pada tahun 2021, Rita Rismala melaksanakan sebuah penelitian yang terfokus pada masalah Adiksi Internet dan Media Sosial di Indonesia. Dengan memandang dari perspektif filsafat sains, penelitian ini menganggap adiksi terhadap internet dan media sosial sebagai fenomena yang dapat dianggap saintifik karena memenuhi kriteria falsifikasi dan dapat diujikan secara empiris. Melalui survei yang melibatkan 1980 responden, hasilnya menunjukkan bahwa sekitar 25,56% dari responden mengalami adiksi terhadap internet dan sekitar 20,2% mengalami adiksi terhadap media sosial. Penelitian ini berhasil membangun sebuah model menggunakan metode XGBoost untuk mendeteksi adiksi terhadap internet dan media sosial. Model ini mencapai tingkat akurasi yang signifikan, diukur dengan F-Measure sebesar 69,23% untuk adiksi terhadap internet dan 67,66% untuk adiksi terhadap media sosial. Hasil penelitian ini memberikan pemahaman yang lebih mendalam terkait adiksi terhadap internet dan media sosial di Indonesia serta memberikan kontribusi dalam pengembangan model pendeteksian menggunakan XGBoost untuk fenomena tersebut [3].

Seperti dalam penelitian analisis sentimen PBB yang dilaksanakan oleh M.R Adrian pada tahun 2021, tujuannya adalah untuk mengkaji sentimen publik terkait penerapan PSBB (Pembatasan Sosial Berskala Besar) menggunakan data tweet yang ditemukan di platform media sosial Twitter. Penelitian ini melibatkan sebanyak 466 data tweet yang digunakan sebagai data latih dan data tes, dengan rasio perbandingan 7 banding 3 antara keduanya. Selanjutnya, data tersebut dianalisis menggunakan dua algoritma yang berbeda untuk perbandingan, yaitu algoritma Support Vector Machine (SVM) dan Random Forest. Tujuan dari analisis ini adalah untuk mendapatkan hasil analisis sentimen yang paling akurat. Hasil penelitian ini memberikan pandangan mendalam tentang bagaimana opini publik terhadap penerapan PSBB dapat diidentifikasi dan dianalisis melalui tweet di platform Twitter. Penggunaan dua algoritma yang berbeda, yaitu SVM dan Random Forest, memberikan pembandingan yang kuat terhadap hasil analisis sentimen yang dicapai. Penelitian ini menjadi sumbangan berharga dalam memahami sentimen masyarakat terkait isu sosial dan kebijakan yang relevan dengan platform media sosial [4].

Pada tahun 2020, Debby Alita melakukan penelitian dalam analisis sentimen yang difokuskan pada Deteksi Sarkasme dalam konteks masyarakat. Penelitian ini melibatkan berbagai tahap, dimulai dari preprocessing dan ekstraksi fitur dalam analisis sentimen. Selanjutnya, metode Support Vector Machine (SVM) digunakan untuk klasifikasi data. Proses pendeteksian sarkasme kemudian dilakukan dengan pendekatan yang melibatkan ekstraksi fitur dari empat set yang berbeda: sentiment related, punctuation-related, lexical and syntactic, serta pattern-related. Klasifikasi untuk pendeteksian sarkasme menggunakan metode Random Forest Classifier. Hasil dari penelitian ini menunjukkan adanya peningkatan yang signifikan dalam sejumlah metrik performa. Nilai rata-rata akurasi meningkat sebesar 16,61%, presisi naik sebesar 5,45%, recall meningkat sebesar 9,64%, dan kenaikan nilai F1-score sebesar 11,27%. Data yang digunakan dalam penelitian ini terdiri dari 2.027 data dengan rincian: 1.023 data dengan label positif, 587 data dengan label negatif, dan 462 data dengan label netral. Data sarkasme diidentifikasi dari tweet yang sebelumnya telah diberi label positif dan kemudian diberi label khusus untuk menandai apakah sarkasme atau tidak. Hasilnya, dari jumlah total 1.023 data dengan label positif, sebanyak 354 diantaranya diklasifikasikan sebagai sarkasme, sementara 669 data dianggap tidak mengandung sarkasme [5].

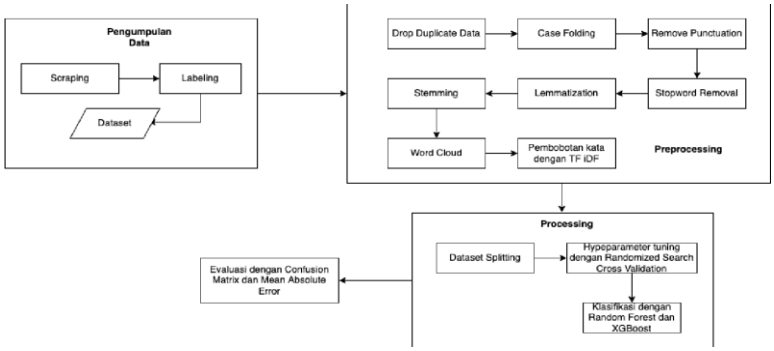
Berdasarkan penjelasan yang telah diuraikan, penelitian ini akan melakukan sebuah studi untuk menganalisis sentimen di Twitter dengan tujuan mengategorikan tweet-tweet yang berkaitan dengan kasus KDRT. Informasi yang dihimpun akan diolah menggunakan teknik text mining, lalu langkah selanjutnya adalah membagi tweet-tweet tersebut ke dalam tiga kategori, yakni positif, negatif, dan netral. Pemisahan ini akan dilakukan dengan menggunakan algoritma random forest dan XGBoost. Proses klasifikasi ini diharapkan dapat mempermudah pengguna dalam mengetahui pandangan yang bersifat positif, negatif, dan netral terkait topik tersebut. Tingkat akurasi dari kedua algoritma ini akan sangat mempengaruhi hasil akhir dari proses klasifikasi tersebut.

II. METODE

Metode penelitian mempunyai beberapa tahapan yang harus dilakukan untuk . Berikut adalah tahapan metode penilitiannya.

A. Tahap Penilitian

Rangkaian langkah penelitian menggambarkan secara menyeluruh urutan prosedur yang akan dilaksanakan dalam pelaksanaan penelitian ini, dimulai dari awal hingga tahap akhir. Serangkaian langkah penelitian yang akan ditempuh dapat dijelaskan lebih rinci melalui sebuah diagram alir, sebagaimana tampak pada gambar 1.



Gambar 1. Tahap Penelitian

.

B. Pengumpulan Data

Data yang digunakan dalam studi ini diperoleh dari platform media sosial Twitter, dengan melakukan pencarian berdasarkan kata kunci KDRT. Dalam proses perolehan data, para peneliti mengimplementasikan Teknik scraping melalui Bahasa pemrograman Python serta memanfaatkan pustaka (library) Tweepy. Tweepy merupakan suatu pustaka Python yang dirancang untuk mengakses antarmuka pemrograman aplikasi (API) Twitter, guna mengambil informasi dari platform tersebut melalui kode sumber dalam bahasa pemrograman Python [6]. Setelah berhasil mengumpulkan data, langkah berikutnya adalah melakukan proses labeling pada data tersebut dengan membaginya ke dalam tiga kategori, yakni positif, negatif, dan netral. Proses penentuan kategori ini dilakukan berdasarkan sentimen yang diungkapkan oleh pengguna Twitter terhadap kasus KDRT yang terjadi di Indonesia..

C. Preprocessing

Langkah selanjutnya melibatkan proses pra-pemrosesan. Tujuan dari langkah ini adalah untuk mengubah informasi teks yang tidak teratur menjadi informasi yang terstruktur. Proses pra-pemrosesan dilakukan dengan niat untuk menangani dataset yang mungkin tidak utuh atau kurang lengkap [7]. Berikut adalah langkah-langkah dalam proses pra-pemrosesan.

1. Case Folding

Langkah berikutnya adalah case folding. Dalam sebuah dokumen, tidak semua kata ditulis dengan huruf besar atau huruf kecil. Oleh karena itu, diperlukan langkah untuk mengubah semua huruf menjadi huruf kecil agar memiliki format yang seragam. Proses ini disebut case folding, yang berfungsi untuk mengubah huruf-huruf abjad menjadi huruf kecil atau lower case [8].

2. Remove Punctuation

Tahap "Remove Punctuation" adalah langkah yang diambil untuk menghilangkan tanda baca atau simbol yang terdapat dalam teks [9]. Hal ini dilakukan karena tanda baca dan simbol tidak memiliki pengaruh yang signifikan terhadap hasil analisis sentimen. Oleh karena itu, penghapusan tanda baca dan simbol menjadi perlu dalam tahap preprocessing.

D. Tokenization

Tokenisasi adalah proses memecah teks yang awalnya dalam bentuk kalimat menjadi unit-unit yang lebih kecil, yaitu kata-kata. Dalam tokenisasi, setiap kata biasanya dipisahkan oleh karakter spasi. Karakter spasi ini menjadi kunci dalam proses tokenisasi karena digunakan untuk memisahkan kata-kata dalam teks dokumen [10]. Hasil dari tahapan ini adalah kumpulan kata

1. Stopword Removal

Tahapan ini merupakan langkah penyaringan (filtering) yang bertujuan untuk mengambil kata-kata yang memiliki makna penting dari hasil tokenisasi [11]. Dalam penelitian ini, peneliti memanfaatkan pustaka (library) Sastrawi untuk melakukan tahap penghapusan kata-kunci (stopword removal).

2. Steaming

Proses ini dikenal sebagai stemming, yang bertujuan untuk mengubah kata-kata menjadi bentuk dasarnya. Fungsi utama dari stemming adalah mengubah kata-kata menjadi bentuk dasar tanpa memperhatikan konteks dari kata tersebut [12].

3. Pembobotan Kata

Tahap ini dilakukan dengan tujuan untuk meningkatkan kemampuan analisis sentimen dalam proses penambangan teks (text mining). Dalam tahap ini, peneliti menerapkan metode Term Frequency – Inverse Document Frequency (TF-IDF). Proses pembobotan dilakukan dengan menghitung nilai Term Frequency (TF) dan Inverse Document Frequency (IDF) secara terpisah. TF mengindikasikan seberapa sering sebuah term muncul dalam satu dokumen tertentu, sementara IDF digunakan untuk mengurangi bobot suatu term apabila kemunculannya menyebar di sejumlah dokumen dalam koleksi tersebut [13]

4. Handling Imbalanced Data

Dalam penelitian ini, terdapat ketidakseimbangan jumlah data pada setiap kelas. Kelas negatif memiliki 671 rekaman, kelas netral memiliki 303 rekaman, dan kelas positif memiliki 599 rekaman. Mengingat distribusi kelas yang tidak seimbang tersebut, penting untuk mengatasi masalah ini melalui pendekatan penanganan data tidak seimbang. Peneliti memutuskan untuk menggunakan Teknik SMOTE (Synthetic Minority Over-sampling Technique) untuk melakukan oversampling. Dalam pendekatan ini, ukuran sampel pada kelas netral dan positif akan ditingkatkan dengan tujuan menyamakan jumlah sampel pada kelas negative [14].

E. Procesing

melewati tahap preprocessing, langkah berikutnya adalah tahap processing, yang mencakup sub-tahapan hyperparameter tuning dan klasifikasi menggunakan dua algoritma machine learning, yaitu random forest dan Extreme Gradient Boosting.

1. Hyperpameter tuning dengan Randomized Cross Validation

Dalam metode machine learning, terdapat sejumlah parameter yang dapat disesuaikan untuk meningkatkan kinerja model, yang disebut sebagai hyperparameter. Pada penelitian ini, metode yang akan digunakan untuk mengoptimalkan hyperparameter adalah Randomized Search Cross Validation. Meskipun memiliki fungsi yang serupa dengan Grid Search CV, Randomized Search CV memiliki keunggulan dalam efisiensi waktu. Metode ini lebih cepat dalam mencari parameter terbaik, terutama ketika terdapat banyak parameter dan volume data yang besar [15]. Adapun parameter yang akan dilakukan tuning adalah sebagai berikut.

Tabel 1. Hyperparameter tuning untuk algoritma Random Forest

Hyperparameter	Kegunaan
n_estimator	Jumlah pohon yang digunakan dalam proses klarifikasi
max_depth	Tingkat kedalaman pohon
Max_features	Jumlah fitur yang perlu dipertimbangkan
Min_sampels_leaf	Jumlah minimum sample yang diperlukan

Tabel 2. Hyperparameter tuning untuk algoritma XGBoost

Hyperparameter	Kegunaan
max_depth	Kedalaman maksimum yang diberikan pada setiap pohon dalam proses klasifikasi.
learning_rate	Digunakan untuk mencegah model agar tidak mengalami overfitting.
n_estimators	Jumlah pohon yang diaplikasikan dalam langkah klasifikasi.
subsample	Rasio jumlah contoh (instance) dalam data latih.
gamma	Menentukan proses pemangkasan pada simpul-simpul pohon yang terbentuk. Semakin besar nilai gamma, semakin konservatif model yang dihasilkan.
colsample_bytree	Parameter untuk memilih jumlah sampel kolom yang akan digunakan.
reg_alpha	Regularisasi L1 pada bobot adalah istilah yang merujuk pada penalti yang diterapkan pada bobot model dengan menggunakan norma L1. Jika nilai regularisasi L1 ditingkatkan, maka model akan menjadi lebih konservatif karena beberapa bobot akan didorong mendekati nol, mengakibatkan fitur yang kurang penting memiliki pengaruh yang lebih kecil pada prediksi model.

reg_lambda	Regularisasi L2 pada bobot adalah istilah yang mengacu pada penalti yang diterapkan pada bobot model dengan menggunakan norma L2. Jika nilai regularisasi L2 ditingkatkan, maka model akan menjadi lebih konservatif karena bobot-bobot yang lebih besar akan diberikan penalti yang lebih besar, sehingga mengurangi kompleksitas model dan potensi overfitting.

2. Klasifikasi dengan Algoritma Random Forest dan XGBost

Setelah dataset melewati tahap pra-pemrosesan, langkah berikutnya adalah memisahkan dataset menjadi data pelatihan (data latih) dan data pengujian (data uji) dengan perbandingan 90% untuk data latih dan 10% untuk data uji. Setelah pemisahan data dan proses penalaan parameter hiper melalui validasi silang (cross-validation) sebanyak 5 kali, langkah selanjutnya adalah mengadaptasi (fitting) model pada data latih. Data latih digunakan untuk membangun model klasifikasi, sementara data uji akan digunakan untuk menguji model tersebut.

Proses pemodelan dilakukan menggunakan dua metode, yaitu Random Forest dan Extreme Gradient Boosting (XGBoost). XGBoost merupakan pengembangan dari metode Random Forest yang menggabungkan konsep dari Random Forest dan Gradient Boosting. Metode XGBoost ini mengoptimalkan kinerja model dengan memadukan keunggulan keduanya [12].

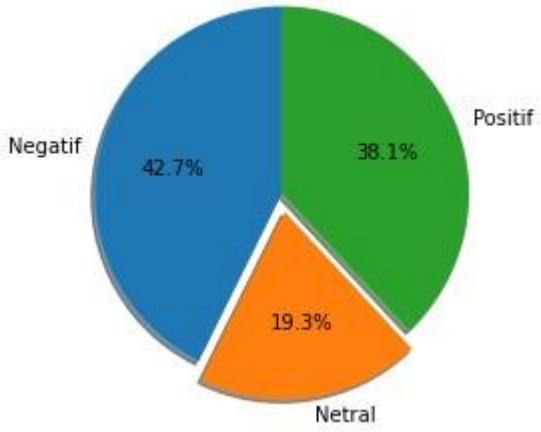
F. Evaluasi

Tahap ini memiliki tujuan untuk mengevaluasi kinerja dari model machine learning yang telah dibangun. Dalam tahap evaluasi, peneliti menggunakan confusion matrix untuk menganalisis kinerja model yang telah dihasilkan. Matriks kebingungan adalah alat pengukuran performa untuk tugas klasifikasi pada machine learning, di mana hasilnya melibatkan dua kelas atau lebih. Matriks kebingungan ini menggambarkan empat kombinasi berbeda antara nilai prediksi dan nilai aktual dalam bentuk tabel [16]. Setelah mendapatkan matriks kebingungan, langkah selanjutnya adalah menghitung nilai akurasi, presisi (precision), recall, dan F1 Score dari model. Akurasi mengukur sejauh mana model memberikan prediksi yang benar secara keseluruhan. Presisi mengukur proporsi hasil positif yang benar dari semua hasil yang dinyatakan positif oleh model. Recall mengukur proporsi hasil positif yang benar dari semua hasil aktual yang seharusnya positif. F1 Score adalah harmonic mean dari presisi dan recall, memberikan gambaran holistik tentang kinerja model. Selanjutnya, peneliti juga memanfaatkan Mean Absolute Error (MAE) untuk menilai sejauh mana kesalahan rata-rata mutlak model dalam memprediksi nilai. MAE mengukur deviasi rata-rata antara nilai prediksi dan nilai aktual [17]. Mean Absolute Error dan Mean Squared Error adalah dua dari berbagai metode yang digunakan untuk mengukur akurasi suatu model peramalan. MAE menghitung rata-rata kesalahan mutlak antara nilai prediksi dan nilai sebenarnya, sementara MSE mengukur rata-rata kesalahan kuadrat antara prediksi dan data aktual. Kedua metode ini memberikan gambaran tentang sejauh mana perbedaan antara prediksi dan nilai aktual, dengan MAE memberikan informasi tentang kesalahan mutlak dan MSE lebih fokus pada kesalahan yang lebih besar.

III. HASIL DAN PEMBAHASAN

A. Dataset

Masukan untuk penelitian ini berupa data penelitian analisis sentimen Twitter terkait kejadian KDRT dengan rincian 599 data berlabel positif, 671 data berlabel negatif dan 303 data apakah ada label netral.



a. Gambar 2 Sebaran sentiment

Dari visualisasi data pada gambar 2 dapat diketahui bahwa persebaran data tidak seimbang dimana kelas negatif terdapat 42.7%, kelas positif 38.1% dan Netral 19.3%.

B. Preprocessing Data

Dalam rangkaian langkah-langkah text mining, teks yang terdapat dalam dokumen harus melalui tahap persiapan sebelum dapat dimanfaatkan dalam proses utama. Proses persiapan data mentah ini juga dikenal sebagai pra-pemrosesan teks. Text preprocessor berperan penting dalam mengubah teks yang awalnya tidak terstruktur atau acak menjadi format yang terstruktur. Pra-pemrosesan dilakukan guna mengatasi potensi dataset yang tidak sempurna, adanya gangguan dalam dataset, ketidaksesuaian data, serta untuk meningkatkan efisiensi proses pemrosesan dokumen [7]. Berikut merupakan hasil dari text preprocessing yang telah melewati beberapa tahapan yang sudah dijelaskan pada tahapan penelitian.

Table 3. Hasil Text Preprocessing

Sebelum Preprocessing	Setelah Preprocessing
Kerja nyatanya adalah mengatasi masalah yang berkaitan dengan perlindungan perempuan dan anak. Saat ini KDRT	kerja nyata adalah atas masalah yang kait dengan lind ung perempuan dan anak

Hasil dari tahap text preprocessing kemudian diberi bobot kata dan pengolahan data yang tidak seimbang sebelum masuk ke tahap pengolahan. Berikut adalah source code untuk menghitung bobot kata menggunakan IDF TF.

Table 4. Kode Sumber TF IDF

```
tfidf_vectorizer = TfidfVectorizer() tfidf_vector = tfidf_vectorizer.fit_transform(X)
tfidf_vector.shape
```

Hasil dari pembobotan kata tersebut selanjutnya dilakukan handling imbalanced data karena jumlah dari masingmasing kelas tidak sama. Berikut merupakan kode sumber dari oversampling dengan SMOTE.

Table 5. Kode Sumber Over Sampling

```
from imblearn.over_sampling import SMOTE from
collections import Counter
oversample = SMOTE()
tfidf_vector, label = oversample.fit_resample(tfidf_vector, label)
```

hasil dari kode sumber di atas menghasilkan jumlah kelas yang sama yaitu positif sebanyak 671 record, negatif 671 recod dan netral 671 record.

C. Processing

Setelah melalui tahapan preprocessing, selanjutnya dilakukan pemisahan data dimana data dibagi menjadi data latih dan data uji dengan persentase 10% untuk data uji dan 90% untuk data uji apakah pelatihan. Selanjutnya, kami menemukan penyetelan hyperparameter terbaik untuk metode Random Forest dan XGBoost. Ini adalah hasil penyetelan hyperparameter untuk metode Random Forest dan XGBoost.

Tabel 6. Hasil Hyperparameter tuning untuk algoritma Random Forest

Hyperparameter	Hasil
n_estimators	188
max_depth	48
Max_features	0.82
Min_samples_leaf	1

Tabel 7. Hasil Hyperparameter tuning untuk algoritma XGBoost

Hyperparameter	Hasil
max_depth	3
learning_rate	0.12
n_estimators	185
subsample	0.78

gamma	1
colsample_bytree	0.69
reg_alpha	0.01
reg_lambda	0.09

Dari Tabel 4 dan 5 diperoleh hasil belajar dan nilai tes metode random forest masing-masing sebesar 97% dan 76%. Sedangkan skor pelatihan dan tes metode XGBoost masing-masing adalah 86% dan 73%. Hasil dari kedua metode tersebut menghasilkan model yang sama-sama overfitt. Namun, hutan acak menghasilkan peralatan berlebih dalam jumlah yang lebih besar daripada metode XGBoost.

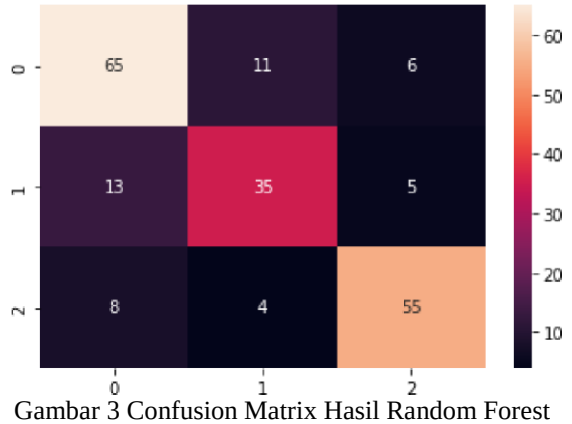
D. Evaluasi

Setelah melalui langkah-langkah pemrosesan, tahap berikutnya melibatkan evaluasi model untuk mengukur kinerjanya. Para peneliti menggunakan matriks kebingungan, yang juga dikenal sebagai matriks kesalahan. Pada dasarnya, Confusion Matrix memberikan gambaran perbandingan hasil klasifikasi yang dilakukan oleh model dengan hasil klasifikasi yang sebenarnya. Confusion matrix adalah sebuah tabel matriks yang menggambarkan performa model klasifikasi pada sejumlah data uji yang memiliki nilai sebenarnya diketahui. Berikut adalah tabel matriks konfusi dengan 4 kombinasi nilai prediksi dan nilai aktual yang berbeda.

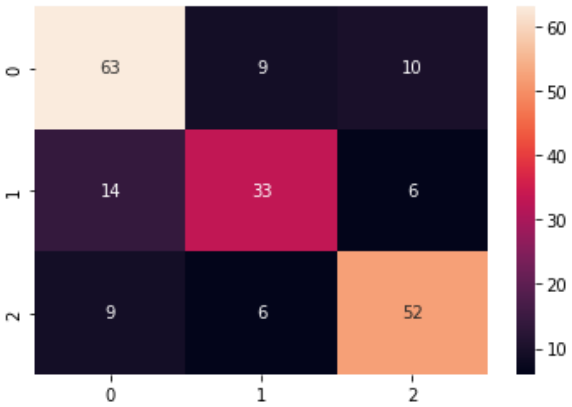
Tabel 8. Confusion Matrix

	Nilai Aktual Positif	**Nilai Aktual Negatif**
Prediksi Positif	True Positif (TP)	False Positif (FP)
Prediksi Negatif	False Negatif (FN)	True Negatif (TN)

Dalam tabel di atas, nilai-nilai seperti True Positive (TP), False Positive (FP), False Negative (FN), dan True Negative (TN) digunakan untuk menghitung berbagai metrik evaluasi seperti akurasi, presisi, recall, F1-score, dan lain-lain. Matriks konfusi membantu dalam mengevaluasi sejauh mana model dapat mengklasifikasikan data dengan benar dan menganalisis jenis-jenis kesalahan yang dilakukan oleh model. Berikut merupakan confusion matrix yang dihasilkan oleh metode random forest dan XGBoost.



Gambar 3 Confusion Matrix Hasil Random Forest



Gambar 4 Confusion Matrix Hasil XGBoost

Dari confusion matrix didapatkan akurasi 77%, akurasi 77%, recall 77n. Skor F1 adalah 77 n Rata-rata kesalahan absolut adalah 0,30. Hasil ini dihasilkan oleh metode hutan acak. Sedangkan metode XGBoost menghasilkan Accuracy 73%, Accuracy 73%, 73n F1 Score Recovery sebesar 73% dan Mean Absolute Error sebesar 0,36.

IV. SIMPULAN

Penelitian ini berhasil menyelesaikan semua tahapan mulai dari pretreatment hingga treatment dan evaluasi. Hasil yang diperoleh pada penelitian ini diperoleh nilai akurasi sebesar 86%. Namun model yang dibuat pada penelitian ini masih mengalami overfitting karena nilai tes yang diperoleh memiliki perbedaan yang besar dengan nilai pelatihan dengan selisih 12%. Dari hasil tersebut, pada penelitian selanjutnya diharapkan metode dari tahap pretreatment hingga tahap pengolahan dapat mengatasi kekurangan yang masih tersisa pada penelitian ini. Selain dapat menerapkan metode machine learning lainnya untuk menjadi benchmark antara algoritma regresi logistik dengan algoritma machine learning lainnya.

UCAPAN TERIMA KASIH

Ucapan terima kasih kepada orang tua saya yang telah mendoakan dan menyemangati saya untuk menyelesaikan penelitiaan ini dengan cepat. Terima kasih kepada Bapak dan Ibu Dosen Informatika Universitas Muhammadiyah Sidoarjo yang telah membimbing dan memberikan ilmu selama kuliah. Tidak lupa saya ucapkan terima kasih kepada teman-teman A1, terutama seseorang yang telah membantu dan memberikan dukungan selama penilitian saya berlangsung.

REFERENSI

[1] A. Ikegami, I. Dewa, M. Bayu, and A. Darmawan, "Analisis Sentimen dan Pemodelan Topik Ulasan Aplikasi Noice Menggunakan XGBoost dan LDA," *Jnatia*, vol. 1, no. 1, 2022.

[2] R. Siringoringo, R. Perangin-angin, and J. Jamaluddin, "MODEL HIBRID GENETIC-XGBOOST DAN PRINCIPAL COMPONENT ANALYSIS PADA SEGMENTASI DAN PERAMALAN PASAR," *METHOMIKA Jurnal Manajemen Informatika dan Komputerisasi Akuntansi*, vol. 5, no. 2, pp. 97–103, Oct. 2021, doi: 10.46880/jmika.Vol5No2.pp97-103.

[3] R. Rismala, L. Novamizanti, K. N. Ramadhani, Y. S. Rohmah, S. Parjuangan, and D. Mahayana, "Kajian Ilmiah dan Deteksi Adiksi Internet dan Media Sosial di Indonesia Menggunakan XGBoost," *Jurnal Edukasi dan Penelitian Informatika (JEPIN)*, vol. 7, no. 1, 2021, doi: 10.26418/jp.v7i1.43606.

[4] M. R. Adrian, M. P. Putra, M. H. Rafialdy, and N. A. Rakhmawati, "Perbandingan Metode Klasifikasi Random Forest dan SVM Pada Analisis Sentimen PSBB," *Jurnal Informatika Upgris*, vol. 7, no. 1, 2021, doi: 10.26877/jiu.v7i1.7099.

[5] D. Alita and A. R. Isnain, "Pendeteksian Sarkasme pada Proses Analisis Sentimen Menggunakan Random Forest Classifier," *jurnal komputasi*, vol. 8, no. 2, 2020, doi: 10.23960/komputasi.v8i2.2615.

[6] N. N. Pandika Pinata, I. M. Sukarsa, and N. K. Dwi Rusjayanthi, "Prediksi Kecelakaan Lalu Lintas di Bali dengan XGBoost pada Python," *Jurnal Ilmiah Merpati (Menara Penelitian Akademika Teknologi Informasi)*, 2020, doi: 10.24843/jim.2020.v08.i03.p04.

[7] Y. S. Nugroho and N. Emiliyawati, "Sistem Klasifikasi Variabel Tingkat Penerimaan Konsumen Terhadap Mobil Menggunakan Metode Random Forest," *Jurnal Teknik Elektro*, vol. 9, no. 1, 2017.

[8] U. Erdiansyah, A. Irmansyah Lubis, and K. Erwansyah, "Komparasi Metode K-Nearest Neighbor dan Random Forest Dalam Prediksi Akurasi Klasifikasi Pengobatan Penyakit Kutil," *JURNAL MEDIA INFORMATIKA BUDIDARMA*, vol. 6, no. 1, 2022, doi: 10.30865/mib.v6i1.3373.

[9] R. Siringoringo, R. Perangin-angin, and M. J. Purba, "SEGMENTASI DAN PERAMALAN PASAR RETAIL MENGGUNAKAN XGBOOST DAN PRINCIPAL COMPONENT ANALYSIS," *METHOMIKA Jurnal Manajemen Informatika dan Komputerisasi Akuntansi*, vol. 5, no. 1, pp. 42–47, Apr. 2021, doi: 10.46880/jmika.Vol5No1.pp42-47.

[10] S. Devella, Y. Yohannes, and F. N. Rahmawati, "Implementasi Random Forest Untuk Klasifikasi Motif Songket Palembang Berdasarkan SIFT," *JATISI (Jurnal Teknik Informatika dan Sistem Informasi)*, vol. 7, no. 2, 2020, doi: 10.35957/jatisi.v7i2.289.

[11] M. Rizky Mubarok, R. Herteno, I. Komputer Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Lambung Mangkurat Jalan Ahmad Yani Km, and K. Selatan, "HYPER-PARAMETER TUNING PADA XGBOOST UNTUK PREDIKSI KEBERLANGSUNGAN HIDUP PASIEN GAGAL JANTUNG."

[12] M. R. Andryan, M. Fajri, and N. Sulistyowati, "KOMPARASI KINERJA ALGORITMA XGBOOST DAN ALGORITMA SUPPORT VECTOR MACHINE (SVM) UNTUK DIAGNOSIS PENYAKIT KANKER PAYUDARA," *JIKO (Jurnal Informatika dan Komputer)*, vol. 6, no. 1, 2022, doi: 10.26798/jiko.v6i1.500.

[13] W. Apriliah, I. Kurniawan, M. Baydhowi, and T. Haryati, "Prediksi Kemungkinan Diabetes pada Tahap Awal Menggunakan Algoritma Klasifikasi Random Forest," *SISTEMASI*, vol. 10, no. 1, 2021, doi: 10.32520/stmsi.v10i1.1129.

[14] H. H. Sinaga and S. Agustian, "Pebandingan Metode Decision Tree dan XGBoost untuk Klasifikasi Sentimen Vaksin Covid-19 di Twitter," *Jurnal Nasional Teknologi dan Sistem Informasi*, vol. 8, no. 3, 2022, doi: 10.25077/teknosi.v8i3.2022.107-114.
[15] E. Fitri, "Analisis Sentimen Terhadap Aplikasi Ruangguru Menggunakan Algoritma Naive Bayes, Random Forest Dan Support Vector Machine," *Jurnal Transformatika*, vol. 18, no. 1, 2020, doi: 10.26623/transformatika.v18i1.2317.
[16] I. Muslim and K. Karo, "Implementasi Metode XGBoost dan Feature Importance untuk Klasifikasi pada Kebakaran Hutan dan Lahan," *Journal of Software Engineering, Information and Communication Technology*, vol. 1, no. 1, 2020.
[17] A. Syukron and A. Subekti, "Penerapan Metode Random Over-Under Sampling dan Random Forest Untuk Klasifikasi Penilaian Kredit," *Jurnal Informatika*, vol. 5, no. 2, 2018, doi: 10.31311/ji.v5i2.4158.

Conflict of Interest Statement:
The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.