COMPARISON OF NAÏVE BAYES, RANDOM FOREST, AND LOGISTIC REGRESSION ALGORITHMS FOR SENTIMENT ANALYSIS ONLINE GAMBLING

KOMPARASI ALGORITMA NAÏVE BAYES, RANDOM FOREST, DAN LOGISTIC REGRESION UNTUK ANALISIS SENTIMEN JUDI ONLINE

Dwi Nanda Agustia[1], Ryan Randy Suryono[2]
[1,2]Universitas Teknokrat Indonesia
Jl. ZA. Pagar Alam No.9-11, Labuan Ratu, Kec. Kedaton, Kota Bandar Lampung, Lampung 35132
E-mail: dwi _nanda_agustia@teknokrat.ac.id[1], ryan@teknokrat.ac.id [2]

Abstract *- This study aims to compare the performance of Naïve Bayes, Random Forest, and Logistic Regression algorithms for sentiment analysis on the topic of online gambling. The dataset consisted of 4592 entries after preprocessing and applying the SMOTE technique to address class imbalance. The evaluation results show that Random Forest achieved the best performance with an accuracy of 78%, followed by Naïve Bayes and Logistic Regression, both achieving 77%. Random Forest excelled in classifying positive and negative sentiments, while Naïve Bayes demonstrated a significant improvement in recall for neutral sentiment, increasing from 0.45 to 0.82 after the SMOTE application. Logistic Regression showed less optimal performance, particularly for neutral sentiment. This study provides essential guidance for selecting the best algorithms for sentiment analysis in specific domains such as online gambling and highlights the importance of SMOTE in handling imbalanced datasets. The findings of this study can be used by practitioners and policymakers to make more informed decisions in regulating online gambling.*

Keywords – *Sentiment Analysis, Online Gambling, Naïve Bayes, Random Forest, Logistic Regression, SMOTE.*

Abstrak - Tujuan dari penelitian ini adalah untuk membandingkan kinerja algoritma Naive Bayes., Random Forest, dan Logistic Regression dalam analisis sentimen pada topik judi online. Dataset yang digunakan terdiri dari 4592 data setelah melalui proses prapemrosesan dan penerapan teknik SMOTE untuk menangani ketidakseimbangan kelas sentimen. Hasil evaluasi menunjukkan bahwa Random Forest memberikan performa terbaik dengan akurasi 78%, diikuti oleh Naïve Bayes dan Logistic Regression yang masing-masing mencapai akurasi 77%. Random Forest unggul dalam klasifikasi sentimen positif dan negatif, sementara Naïve Bayes menunjukkan peningkatan recall signifikan pada sentimen netral dari 0,45 menjadi 0,82 setelah penerapan SMOTE. Logistic Regression menunjukkan performa kurang optimal, terutama pada sentimen netral. Penelitian ini memberikan panduan penting untuk memilih algoritma terbaik dalam analisis sentimen di domain spesifik seperti judi online dan menyoroti pentingnya SMOTE dalam mengatasi dataset tidak seimbang. Temuan dari penelitian ini bisa digunakan oleh praktisi dan pembuat kebijakan untuk membuat keputusan yang lebih tepat dalam mengatur perjudian online.

Kata Kunci - Analisis Sentimen, Judi Online, Naïve Bayes, Random Forest, Logistic Regression, SMOTE.

I. PENDAHULUAN

Perjudian telah menjadi bagian dari permasalahan sosial yang ada sejak zaman dahulu. Dalam perkembangannya, perjudian kini bertransformasi melalui teknologi menjadi perjudian daring, yang semakin meluas seiring dengan meningkatnya jumlah pengguna perangkat komunikasi elektronik berbasis internet [1]. Perjudian merupakan aktivitas mempertaruhkan sejumlah uang, di mana pemenangnya akan memperoleh seluruh uang taruhan tersebut. Aktivitas ini sering dianggap sebagai ajang mencoba keberuntungan karena hasil permainannya bergantung pada faktor kebetulan. Bagi pemain yang kalah, mereka harus menerima kekalahan dengan konsekuensi kehilangan uang yang telah dipertaruhkan [2]. Dampak judi *online* tidak hanya dirasakan oleh individu yang terlibat secara langsung, tetapi juga memengaruhi keluarga dan komunitas di sekitarnya. Dari sisi ekonomi, banyak keluarga menghadapi tekanan finansial akibat anggota keluarga yang terjebak dalam kecanduan judi online. Secara psikologis, pecandu judi *online* sering kali membuatnya menderita gangguan mental seperti stres, kecemasan, dan depresi. Dorongan untuk terus menang dan kerugian yang terus-menerus dapat meningkatkan tingkat stres, yang pada akhirnya berdampak pada kesejahteraan mental dan emosional individu tersebut [3].

Analisis sentimen digunakan untuk memahami opini atau emosi dalam teks, seperti sentimen positif, negatif, atau netral. Dalam kasus judi *online*, analisis ini mengungkap persepsi masyarakat terhadap aktivitas tersebut, baik sebagai hiburan, ancaman, atau peluang. Penelitian ini membandingkan kinerja algoritma *Naïve Bayes*, *Random Forest*, dan *Logistic Regression* dalam analisis sentimen judi *online*, menggunakan metrik akurasi, presisi, *recall*, dan F1-*score*. Penelitian juga mengevaluasi efektivitas algoritma pada dataset tidak seimbang serta menganalisis faktor-faktor seperti parameter model, *preprocessing* data, dan representasi teks dengan TF-IDF. Penelitian ini diharapkan dapat memberikan panduan dalam memilih algoritma yang paling sesuai untuk analisis sentimen di domain spesifik seperti judi *online*, serta berharap penelitian ini dapat berdampak besar pada kebijakan dan praktik terkait judi *online*. Di tingkat lokal, analisis ini dapat membantu pemerintah atau komunitas memahami dampak judi online dan menyusun regulasi yang lebih efektif. Secara global, hasilnya dapat digunakan untuk memperkuat pengawasan platform judi *online* dan meningkatkan edukasi masyarakat tentang risikonya.

II. SIGNIFIKASI STUDI

A. Penelitian Terdahulu

Beberapa studi yang telah dilakukan sebelumnya telah memberikan kontribusi yang signifikan dalam membentuk dasar pemikiran dan metodologi yang digunakan dalam penelitian ini. Penelitian-penelitian terdahulu berfungsi sebagai referensi yang memperkaya pemahaman dan mendasari analisis yang dilakukan dalam studi ini.

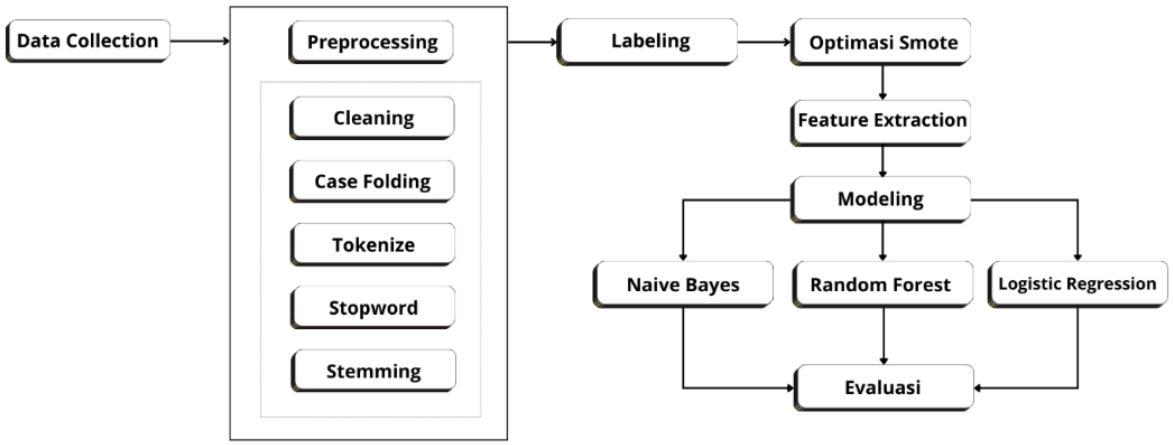
TABEL I
PENELITIAN TERDAHULU

No.	Penulis	Penelitan Terdahulu
1	Andrea Maulana dan Ade Yuliana [4]	Analisis Sentiment Opini Public tentang Judi Online pada Pengguna Aplikasi X menggunakan Algoritma Naïve Bayes dan Support Vector Mechine. Data diperoleh dari platform X. Terungkap bahwa algoritma SVM mencapai akurasi sebesar 98%, yang lebih unggul diatas Naive Bayes dengan nilai sebesar 93%.
2	Robert Antonius, Achmad Rizky Zulkarnain dan Hafiz Irsyad [5]	Pendekatan TF-IDF, SMOTE, dan SVM dalam Klasifikasi Sentimen Masyarakat terhadap Pemblokiran Judi Online. Model dilatih menggunakan dataset yang diseimbangkan dengan metode SMOTE untuk mengatasi ketimpangan klasifikasi, kemudian diberi bobot menggunakan TF-IDF agar lebih menekankan pada kata-kata dengan nilai signifikan tinggi. dengan Model klasifikasi yang dikembangkan menggunakan algoritma Support Vector Machine berhasil mencapai tingkat akurasi sebesar 61,54%, berdasarkan evaluasi menggunakan confusion matrix sebagai tolok ukur.
3	Aji Dewo Pangestu dan Lailan Sofinah [6]	Analisis Sentimen Terkait Judi Online di Media Sosial Instagram Menggunankan Naïve Bayes. Data penelitian dikumpulkan melalui proses *scraping* komentar pada unggahan Instagram, menghasilkan dataset yang terdiri dari 4 komentar. Hasil analisis mengungkapkan bahwa 50% dari komentar mengandung sentimen negatif, sedangkan sentimen positif dan netral masing-masing mencapai 25%. Sentimen negatif ini sebagian besar berisi kritik terhadap isu korupsi dan kemiskinan yang berkaitan dengan judi online.

Berdasarkan penelitian terdahulu, TF-IDF dan SMOTE efektif dalam analisis sentimen. TF-IDF membantu algoritma menyoroti kata-kata penting untuk menganalisis sentimen, sementara SMOTE mengatasi ketidakseimbangan kelas dalam dataset, meningkatkan deteksi sentimen minoritas. Pemilihan metode *Naïve Bayes, Random Forest,* dan *Logistic Regression* didasarkan pada kemampuan masing-masing algoritma dalam menangani masalah ketidakseimbangan kelas dan efisiensi dalam klasifikasi sentimen. Teknik TF-IDF dan SMOTE dipilih untuk meningkatkan kualitas analisis sentimen judi online, mengoptimalkan akurasi dan relevansi hasil.

B. Metode Penelitian

Proses penelitian dilakukan melalui beberapa tahapan, dimulai dengan pengumpulan data menggunakan teknik *scraping* pada *platform* X terkait topik Judi *Online*. Tahap berikutnya adalah *pre-processing* data, yang mencakup langkah-langkah seperti pembersihan data, *case folding*, tokenisasi, penghapusan *stopword*, dan *stemming*. Setelah itu, data diberi label menjadi tiga kategori, yaitu negatif, positif, dan netral. Untuk menjaga keseimbangan data, dilakukan optimasi dengan metode SMOTE. Data yang sudah seimbang kemudian melalui tahap ekstraksi fitur. Selanjutnya, data tersebut dimodelkan menggunakan tiga algoritma: *Naïve Bayes, Random Forest, dan Logistic Regression.* Terakhir, peneliti mengevaluasi hasil performa dari setiap algoritma. Rangkaian tahapan penelitian ini ditampilkan pada gambar1.



Gambar 1. Tahapan Penelitian

1. *Data Collection* (Pengumpulan Data)

Data dikumpulkan dalam rentang waktu 1 Januari 2024 hingga 23 November 2024 menggunakan API dan Token X untuk mengelola data. Pengumpulan data dilakukan dengan beberapa kata kunci yang berhubungan dengan topik, seperti judi *online*, judi slot, dan judol. Jumlah data yang berhasil dihimpun adalah sebanyak 5744 yang kemudian disimpan dalam format csv.

2. *Pre-Processing* (Data Preparation)

Pre-processing Ini adalah langkah pertama dalam pemrosesan data, dan tujuannya adalah untuk menyiapkan data kotor sehingga siap untuk analisis dan pemodelan.. Tahap ini digunakan karena data yang berkualitas dan terstruktur akan memberikan pengaruh positif terhadap hasil analisis serta pengambilan keputusan [7]. Pada penelitian ini, tahap data *preprocessing* mencakup beberapa langkah, yaitu pembersihan data, reduksi data, dan normalisasi data.

a. *Data Cleaning* (Pembersihan Data)
   Pembersihan data adalah proses dalam pengolahan data yang bertujuan untuk memperbaiki atau menangani data yang tidak teratur, duplikat, atau kurang sesuai dalam *dataset*. Data yang hilang dapat ditangani dengan menghapusnya atau menggantinya menggunakan nilai yang relevan. *Outlier* dapat diatasi dengan metode statistik seperti *trimming* atau *winsorizing*, atau dihapus jika diperlukan [8].
b. *Case Folding*
   Case folding adalah tahap dalam preprocessing teks yang bertujuan untuk menyamakan karakter dalam data. Proses ini mengubah semua huruf menjadi huruf kecil, di mana karakter-karakter dari 'A' hingga 'Z' diubah menjadi 'a' hingga 'z' [9].
c. *Tokenizing*
   Tokenizing atau tokenisasi adalah proses memecah kalimat menjadi kata-kata terpisah untuk setiap kata. Tujuan dari proses ini adalah untuk memperoleh potongan kata yang akan dijadikan entitas dan digunakan dalam matriks dokumen yang akan dianalisis [10].
d. *Stopword*
   Stopword merujuk pada kata yang sering muncul dalam teks tetapi tidak terlalu penting dalam analisis, seperti kata penghubung, preposisi, atau kata-kata umum lainnya.
e. *Stemming*
   Stemming adalah proses yang bertujuan untuk menghilangkan infleksi kata sehingga kata-kata tersebut dapat dikembalikan ke bentuk dasarnya. Secara sederhana, *stemming* merupakan proses penghapusan imbuhan pada kata sehingga hanya menyisakan bentuk dasar dari kata tersebut [11].

3. *Labeling* (Pelabelan Data)

Labelling adalah proses menambahkan tanda tertentu pada teks dengan mempertimbangkan emosi termasuk arti di dalamnya. Pada dasarnya, teks dari artikel, komentar, atau media sosial dikelompokkan seperti positif, negatif, atau netral [4].

4. SMOTE

SMOTE merupakan metode yang diterapkanuntuk menangani masalah ketidakseimbangan kelas pada dataset, terutama ketika jumlah sampel pada kelas minoritas sangat terbatas dibandingkan dengan kelas mayoritas. SMOTE dapat mengatasi ketidakseimbangan jumlah sampel antar kelas dengan menghasilkan data sintetis untuk kelas minoritas dengan melakukan replikasi, yang bertujuan untuk meningkatkan jumlah sampel pada kelas tersebut[12]. Teknik ini sangat relevan untuk dataset yang memiliki distribusi sentimen yang tidak merata, seperti yang mungkin terjadi pada diskusi tentang judi online yang bisa lebih banyak mengandung kritik atau peringatan.

5. Feature Extraction

Dalam penelitian ini, fitur teks diekstraksi dengan menggunakan metode TF-IDF. Teknik TF-IDF digunakan untuk memberikan bobot lebih tinggi pada token-token yang sering muncul dalam korpus namun hanya terdapat pada sejumlah kecil dokumen. Dengan langkah tersebut, TF-IDF memperkuat fitur berdasarkan frekuensi kata dalam teks, yang membantu mempercepat dan mempermudah proses klasifikasi[5]. Dalam konteks analisis sentimen judi online, TF-IDF membantu algoritma untuk memberi bobot lebih pada kata-kata yang relevan dan jarang, yang seringkali mengandung makna penting untuk identifikasi sentimen

6. Modeling

Pada tahap ini, dilakukan pemodelan klasifikasi terhadap data *tweet* yang telah melalui serangkaian proses sebelumnya, seperti prapemrosesan data, ekstraksi fitur, penerapan SMOTE, dan pembagian data. Selanjutnya, penulis menggunakan beberapa model klasifikasi yang meliputi, *Naïve Bayes Classifier*, yaitu metode klasifikasi yang didasarkan pada Teorema Bayes dengan asumsi bahwa setiap kondisi atau kejadian dianggap independen satu sama lain. Artinya, jika output tertentu sudah diketahui, probabilitas pengamatan bersama dapat dihitung sebagai hasil kali probabilitas individu. Kelebihan metode ini adalah kemampuannya untuk bekerja dengan baik meskipun hanya menggunakan data pelatihan dalam jumlah kecil. Beberapa kasus dunia nyata yang kompleks, *Naïve Bayes* sering memberikan hasil yang lebih baik dari yang diperkirakan[13]. Selanjutnya *Random Forest*, menggabungkan prediksi dari banyak *decision tree* dalam satu model, sehingga risiko *overfitting*, di mana model terlalu bergantung pada satu dataset dan kurang akurat pada dataset lain dapat diminimalkan. Selain itu, *Random Forest* memiliki berbagai hyperparameter yang dapat disesuaikan secara manual, memungkinkan peningkatan performa sistem dalam penelitian ini[14]. Adapun yang terakhir adalah *Logistic Regression* merupakan salah satu metode statistik yang digunakan untuk melakukan klasifikasi, yaitu memperkirakan probabilitas dari kejadian berdasarkan variabel independen. Probabilitas adalah ukuran kemungkinan terjadinya suatu peristiwa dalam sebuah eksperimen. Nilai probabilitas ini selalu berada dalam rentang antara 0 dan 1[15].

7. Evaluasi

Pengujian model untuk menilai kinerja metode *Naive Bayes, Random Forest*, dan *Logistic Regression* dilakukan untuk mengevaluasi sejauh mana model tersebut mampu memprediksi data dengan akurat. Selanjutnya, evaluasi kinerja model dilakukan menggunakan metode matriks konfusi. Matriks ini memperlihatkan bagaimana model membuat prediksi, serta seberapa sering prediksi tersebut benar atau salah. Dengan matriks konfusi, berbagai metrik seperti *recall*, skor F1, presisi, dan akurasi dapat dihitung, memberikan gambaran yang lebih mendalam tentang performa model dibandingkan jika hanya mengandalkan presisi saja. Model matriks konfusi akan menghasilkan matriks yang terdiri dari *true positive* (TP) atau pasangan positif, dan *true negative* (TN) atau pasangan negatif [4].

III. HASIL DAN PEMBAHASAN

A. *Dataset*

Pengumpulan data dilakukan dalam rentang waktu 1 Januari 2024 hingga 23 November 2024 memanfaatkan API dan Token X untuk mengelola data. Adapun beberapa kata kunci yang berkaitan dengan topik penelitian antara lain, seperti judi online, judi slot, dan judol. Jumlah data yang berhasil dihimpun adalah sebanyak 5744 yang kemudian disimpan dalam format csv. Hasil dari proses crawling data disajikan dalam Tabel 2.

TABEL 2
DATA COLLECTION

No.	Dataset
1	@hariqosatria @NasbiHasan @pco_ri Ternyata presiden sebelumnya gak ada komitmen tegas utk memberantas judi online. Artinya dia sengaja menjerumuskan rakyatnya terjerat judol.
2	@GunRomli @denni_sauya @DivHumas_Polri Sikat jangan segan segan dan jangan Padang siapa berantas judi Online.
3	Polisi Periksa Pejabat di Komdigi Terkait Judi Online Meutya Hafid Tunjuk Brigjen Jadi Dirjen - https://t.co/mgushG7mCs #viral #fyp #x #reader #indonesia https://t.co/0kc9NUOgeN JAKARTA - Upaya Polri melakukan pembersihan di Kementerian Komunikasi dan Digit... https://t.co/rP5FiCfSuA

B. *Pre-Processing* (Data Preparation)

1. *Data Cleaning* (Pembersihan Data)

Tahapan data cleaning yang dilakukan meliputi penghapusan *tag*, *hashtag*, URL, *mention (username)*, tanda baca, angka, emoji, spasi yang berlebihan, serta penghilangan karakter yang diulang secara berlebihan.

TABEL 3
DATA CLEANING

Sebelum	Sesudah
Polisi Periksa Pejabat di Komdigi Terkait Judi Online Meutya Hafid Tunjuk Brigjen Jadi Dirjen - https://t.co/mgushG7mCs #viral #fyp #x #reader #indonesia https://t.co/0kc9NUOgeN JAKARTA - Upaya Polri melakukan pembersihan di Kementerian Komunikasi dan Digit... https://t.co/rP5FiCfSuA	polisi periksa pejabat di komdigi terkait judi online meutya hafid tunjuk brigjen jadi dirjen viral fyp x reader indonesia jakarta upaya polri melakukan pembersihan di kementerian komunikasi dan digit

2. Case Folding

Case folding untuk menyamakan karakter setiap data. Proses ini mengubah semua huruf menjadi huruf kecil, di mana karakter-karakter dari 'A' hingga 'Z' diubah menjadi 'a' hingga 'z' [9]. Seperti kata “Prabowo” menjadi “prabowo”. Seperti ditunjukan pada tabel 4

TABEL 4
CASE FOLDING

Sebelum	Sesudah
Presiden Prabowo perintahkan berantas Judi Online! Sejak awal masa pemerintahan Prabowo menginstruksikan tanpa kompromi Beliau menekankan kerja sama lintas kementerian &amp	presiden prabowo perintahkan berantas judi online sejak awal masa pemerintahan prabowo menginstruksikan tanpa kompromi beliau menekankan kerja sama lintas kementerian amp

3. Tokenizing

Tokenizing atau tokenisasi adalah proses memecah kalimat menjadi kata-kata terpisah untuk setiap kata. Tujuan dari proses ini adalah untuk memperoleh potongan kata yang akan dijadikan entitas dan digunakan dalam *matriks* dokumen yang akan dianalisis [10].

TABEL 5
TOKENIZING

Sebelum	Sesudah
satu hal yang paling saya sesalin dalam hidup saya iya kenal judi online padahal dulu saya selalu benci dan ketawain orang yang sampai kelilit hutang karena judol dan iya sekarang orang yang dulu ketawain itu diri saya sendiri saya kehilangan banyak hal kehilangan pacar yang sangat saya sayang	['satu', 'hal', 'yang', 'paling', 'saya', 'sesalin', 'dalam', 'hidup', 'saya', 'iya', 'kenal', 'judi', 'online', 'padahal', 'dulu', 'saya', 'selalu', 'benci', 'dan', 'ketawain', 'orang', 'yang', 'sampai', 'kelilit', 'hutang', 'karena', 'judol', 'dan', 'iya', 'sekarang', 'orang', 'yang', 'dulu', 'ketawain', 'itu', 'diri', 'saya', 'sendiri', 'saya', 'kehilangan', 'banyak', 'hal', 'kehilangan', 'pacar', 'yang', 'sangat', 'saya', 'sayang']

4. Stopword

Stopword merujuk pada kata yang sering timbul dalam teks tetapi tidak berpengaruh dalam analisis, seperti kata penghubung, preposisi, atau kata-kata umum lainnya. Contohnya adalah "yang", "itu", "saya". Menghapus *stopword* dapat mengurangi kompleksitas data dan memungkinkan analisis lebih fokus pada kata-kata yang lebih penting.

TABEL 6
STOPWORD

Sebelum	Sesudah
satu hal yang paling saya sesalin dalam hidup saya iya kenal judi online padahal dulu saya selalu benci dan ketawain orang yang sampai kelilit hutang karena judol dan iya sekarang orang yang dulu ketawain itu diri saya sendiri saya kehilangan banyak hal kehilangan pacar yang sangat saya sayang	['satu', 'paling', 'sesalin', 'hidup', 'iya', 'kenal', 'judi', 'online', 'padahal', 'dulu', 'selalu', 'benci', 'ketawain', 'kelilit', 'hutang', 'judol', 'iya', 'sekarang', 'dulu', 'ketawain', 'diri', 'sendiri', 'kehilangan', 'banyak', 'kehilangan', 'pacar', 'sangat', 'sayang']

5. Stemming

Stemming bertujuan menghilangkan infleksi kata sehingga kata-kata tersebut dapat dikembalikan ke bentuk dasarnya. Seperti contoh pada table kata “perintahkan” dikembalikan pada kata dasarnya yaitu “perintah”.

TABEL 7
STEMMING

Sebelum	Sesudah
Presiden Prabowo perintahkan berantas Judi Online! Sejak awal masa pemerintahan Prabowo menginstruksikan tanpa kompromi Beliau menekankan kerja sama lintas kementerian &amp	['presiden', 'prabowo', 'perintah', 'berantas', 'judi', 'online', 'sejak', 'awal', 'masa', 'perintah', 'prabowo', 'instruksi', 'kompromi', 'beliau', 'tekan', 'kerja', 'sama', 'lintas', 'menteri', 'amp']

6. Labeling

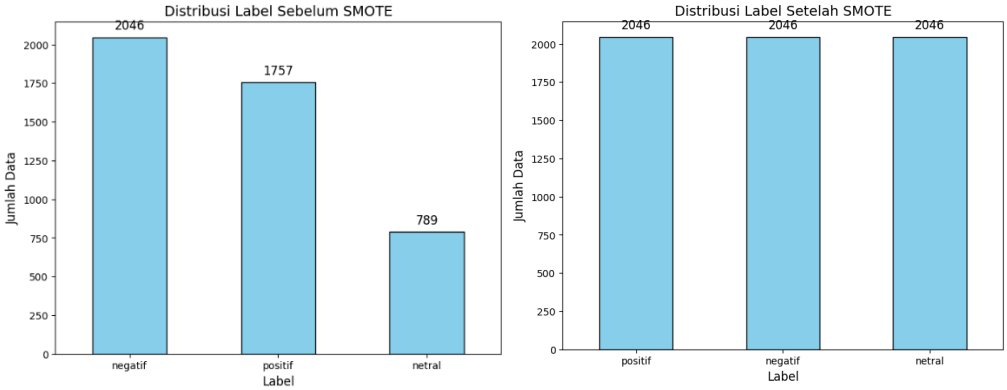
Setelah melalui tahap prapemrosesan, jumlah data ulasan yang awalnya sebanyak 5744 berkurang menjadi 4592 dataset. Tahap berikutnya adalah pelabelan data berdasarkan isi tweet ulasan. Data yang berisi ungkapan dukungan atau saran akan dikategorikan sebagai sentimen positif. Sebaliknya, data yang mengandung ketidakpercayaan, fitnah, hoaks, serta hal-hal negatif terkait judi online, baik dalam bentuk sindiran maupun pernyataan persetujuan terhadap dampak buruknya, akan diklasifikasikan sebagai sentimen negatif. Sementara itu, data yang tidak menunjukkan sentimen

kuat, seperti informasi netral atau pernyataan yang tidak condong ke arah positif maupun negatif, akan dikategorikan sebagai sentimen netral.

TABEL 8
LABELING

No.	Tweet	Sentimen
1	presiden prabowo perintahkan berantas judi online sejak awal masa pemerintahan prabowo menginstruksikan tanpa kompromi beliau menekankan kerja sama lintas kementerian amp	Positif
2	sumpah jaman sekarang ternyata masih banyak yah orang yang tergiur kaya instan kasian banget sama sepupu saya sudah diselingkuhin sama mantan suaminya sekarang nikah kedua kali suami yang tadinya sholeh banget tiba kelilit judol sama pinjol padahal dia isilop	Negatif
3	klik untuk bergabung main di tempat kami pasti di bayar berapapun kemenanggannya klik adalah agen judi online sepakbola slot casino poker dan toto gelap prediksiligaindonesiaklik sepakbola whatsapp	Netral

Pada distribusi *persentase* sentimen, terdapat ketidakseimbangan antara sentimen negatif, positif maupun netral. Jika satu emosi mendominasi dalam kumpulan data, model dapat berfokus pada pelatihan data emosi negatif. Akibatnya, kinerja model dalam mengklasifikasikan emosi positif menjadi kurang optimal. Untuk mengatasi masalah ini, penelitian ini menggunakan teknik SMOTE sebagai solusi untuk mengatasi ketidakseimbangan kelas. Perbandingan data sebelum dan sesudah implementasi SMOTE ditunjukkan pada Gambar 2.



Gambar 2. Presentase Sentimen Tweet Judi Online

Untuk mengatasi ketidakseimbangan data, pada kelas minoritas ditambahkan 289 data sintetis pada sentimen positif, serta 1257 data tambahan lainnya untuk menyamakan jumlah dengan kelas mayoritas, yaitu sentimen negatif. Penambahan ini bertujuan agar algoritma model dapat mempelajari sentimen negatif dan positif secara lebih seimbang. Setelah teknik SMOTE diterapkan, jumlah data untuk sentimen negatif dan positif menjadi setara, yaitu masing-masing sebanyak 2046 data.

C. Evaluasi Algoritma

Setelah melalui serangkaian fase persiapan data, pengujian dilakukan dengan membandingkan algoritma *Naïve Bayes*, *Random Forest*, dan *Logistic Regression*. Proses ini menggunakan pembagian data sebesar 80% untuk training dan 20% untuk testing. Data kemudian dapat dianalisis dan dievaluasi sebelum dan sesudah menerapkan teknik SMOTE.

TABEL 9
DATA COLLECTION

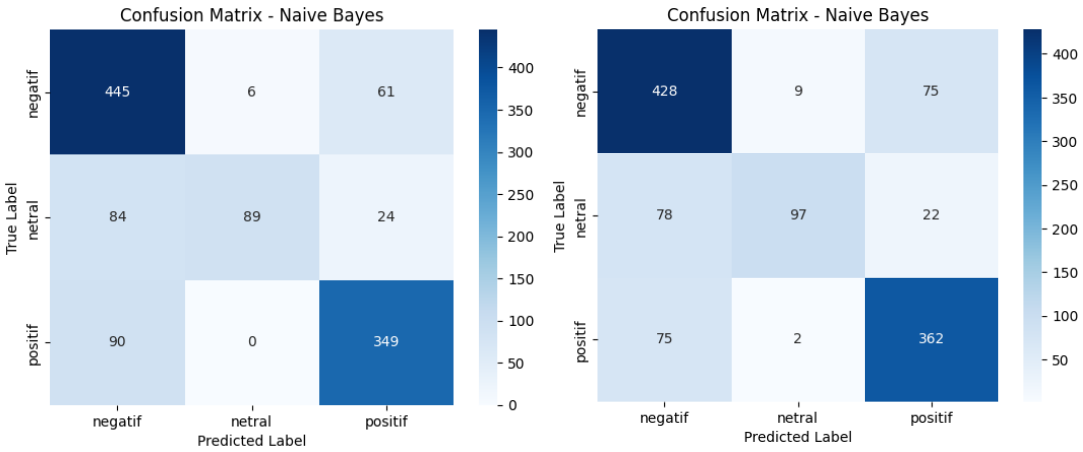
Matriks	NB	NB+ SMOTE	RF	RF+ SMOTE	LR	LR+ SMOTE
Accuracy	0.77	0.77	0.78	0.78	0.78	0.77
Sentimen Negatif						
Precison	0.72	0.74	0.72	0.74	0.71	0.75
Recall	0.87	0.84	0.90	0.87	0.89	0.81
F1-Score	0.79	0.78	0.80	0.80	0.79	0.78
Sentimen Positif						
Precision	0.80	0.79	0.86	0.85	0.84	0.83
Recall	0.79	0.82	0.75	0.76	0.76	0.77
F1-Score	0.80	0.81	0.80	0.81	0.80	0.80
Sentimen Netral						
Precision	0.94	0.90	0.86	0.76	0.90	0.71
Recall	0.45	0.82	0.75	0.59	0.50	0.64
F1-Score	0.80	0.81	0.80	0.81	0.64	0.67

Berdasarkan hasil evaluasi, semua algoritma menunjukkan nilai akurasi yang hampir serupa, yaitu pada kisaran 0.77–0.78, tanpa adanya peningkatan signifikan setelah penerapan teknik SMOTE. Untuk sentimen negatif, penerapan SMOTE meningkatkan nilai precision pada semua algoritma, misalnya *Naïve Bayes* dari 0.72 menjadi 0.74, *Random Forest* dari 0.72 menjadi 0.74, dan Logistic Regression dari 0.71 menjadi 0.75. Namun, nilai recall justru sedikit menurun, terutama pada *Logistic Regression* yang turun dari 0.89 menjadi 0.81. Nilai F1-*Score* pada sentimen negatif tetap stabil di kisaran 0.78–0.80. Pada sentimen positif, *precision* cenderung menurun setelah SMOTE, sedangkan recall meningkat, seperti pada NB yang meningkat dari 0.79 menjadi 0.82. Secara keseluruhan, nilai F1-*Score* tetap konsisten di kisaran 0.80–0.81. Untuk sentimen netral, penerapan SMOTE memberikan peningkatan signifikan pada *recall*, terutama pada NB yang meningkat dari 0.45 menjadi 0.82. Namun, *precision* mengalami penurunan, seperti pada RF menurun dari 0.86 ke 0.76 dan LR yang turun dari 0.90 menjadi 0.71. F1-*Score* pada sentimen netral menunjukkan peningkatan kecil pada NB dan RF setelah SMOTE, tetapi tetap rendah untuk LR.

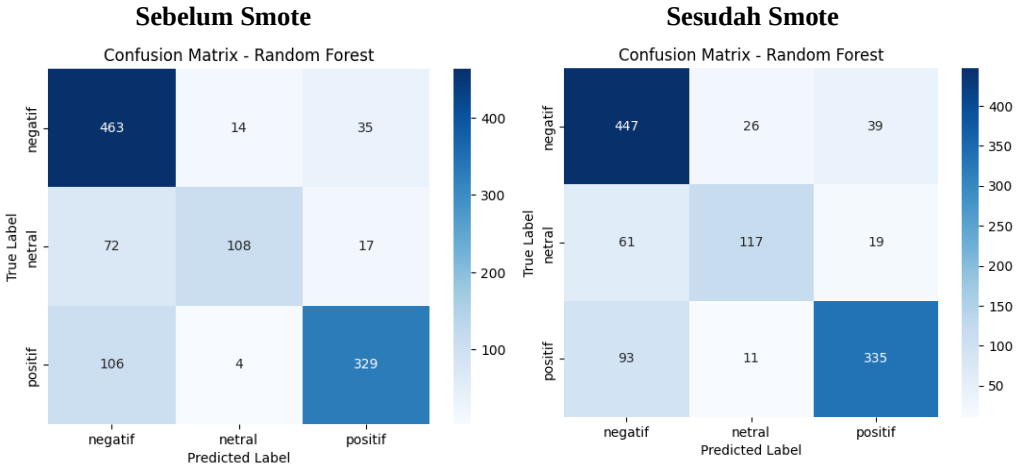
Random Forest konsisten menunjukkan performa terbaik pada sentimen positif dan negatif, sementara Naïve Bayes mengalami peningkatan signifikan pada recall untuk sentimen netral setelah penerapan SMOTE. SMOTE efektif mengatasi ketidakseimbangan data dengan meningkatkan *recall* kelas minoritas, meskipun *precision* cenderung menurun akibat kesalahan klasifikasi. Secara keseluruhan, F1-*Score* tetap stabil, mencerminkan keseimbangan antara *precision* dan *recall*. Dampak SMOTE bervariasi tergantung algoritma dan jenis sentimen yang dianalisis.

Penelitian ini juga melakukan analisis perbandingan menggunakan *confusion matrix* untuk mengevaluasi sejauh mana kedua algoritma mampu mengklasifikasikan data dengan tepat serta mengidentifikasi kesalahan klasifikasi ada. Terlihat pada gambar dibawah yang menunjukan hasil *confusion matrix*

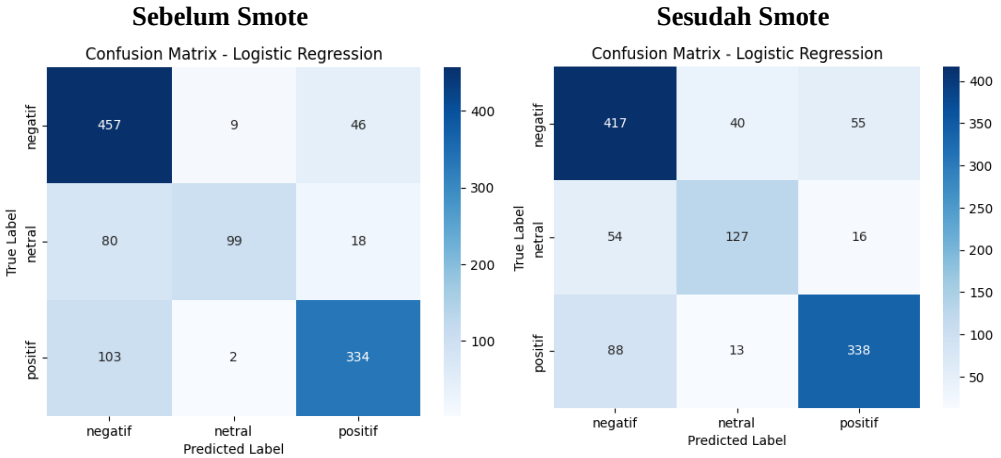
Sebelum Smote **Sesudah Smote**



Gambar 3. *Confusion Matrix Naïve Bayes* Sebelum dan Sesudah Smote



Gambar 4. *Confusion Matrix Random Forest* Sebelum dan Sesudah Smote



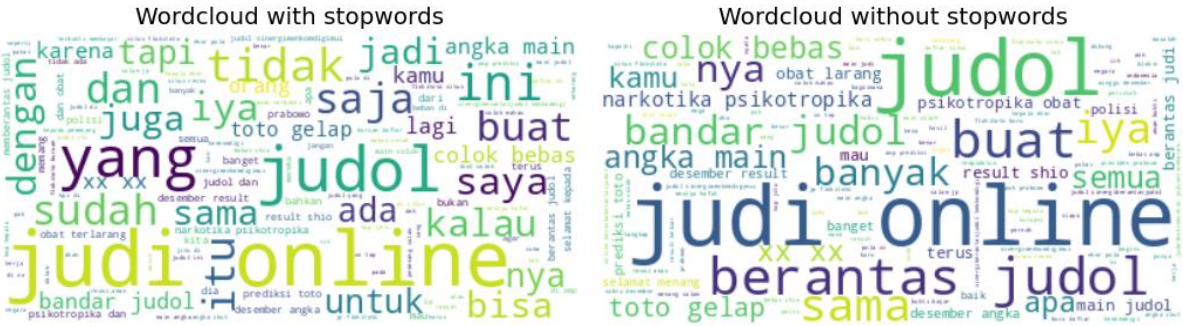
Gambar 5. *Confusion Matrix Regresion linear* Sebelum dan Sesudah Smote

Penerapan SMOTE memberikan pengaruh yang berbeda pada setiap algoritma. *Naïve Bayes* meningkatkan prediksi pada kelas Positif (*True Positif* dari 349 menjadi 362), tetapi mengalami penurunan pada kelas Negatif (*True Negative* dari 445 menjadi 428) serta lebih sering salah mengklasifikasikan kelas Negatif dan Netral setelah SMOTE diterapkan. *Random Forest* menunjukkan hasil yang lebih konsisten, dengan peningkatan pada kelas Positif (*True Positif* dari 329 menjadi 335) dan hanya sedikit penurunan pada kelas Negatif (*True Negative* dari 463 menjadi 447), sehingga menunjukkan stabilitas dalam mengelola *dataset* yang telah seimbang. Sementara

itu, *Logistic Regression* mengalami penurunan performa yang signifikan pada kelas Negatif (*True Negative* dari 457 menjadi 417) dan peningkatan kesalahan klasifikasi pada kelas Positif (*False Positive* dari 2 menjadi 13). Secara keseluruhan, *Random Forest* dianggap sebagai algoritma yang paling andal setelah penerapan SMOTE karena mampu meningkatkan prediksi pada kelas Positif tanpa terlalu banyak mengorbankan akurasi pada kelas lain. Meskipun *Naïve Bayes* menunjukkan perbaikan signifikan pada kelas Positif, penurunan pada kelas Negatif dan kecenderungan membuat kesalahan pada kelas Netral menjadi kelemahan utama. *Logistic Regression*, meskipun sederhana dan mudah diimplementasikan, menunjukkan keterbatasan yang signifikan dalam menangani *dataset* tidak seimbang, bahkan setelah penerapan SMOTE.

D. Hasil Visualisasi Data

Digunakan *wordcloud* untuk menampilkan kata-kata yang paling sering muncul dalam tweet yang berkaitan dengan topik judi *online*. Proses visualisasi ini akan dilakukan dalam dua tahap, yaitu sebelum dan setelah penghapusan *stopwords.*



Gambar 6. *Word Cloud With Stopword* dan *Without Stopword*

Pada Gambar 6, hasil visualisasi *wordcloud* menampilkan kata-kata yang paling sering muncul, di antaranya "judi online" sebagai topik utama penelitian ini, "judol" yang merupakan singkatan dari judi *online*, serta "berantas judol" yang merujuk pada upaya untuk menghentikan aktivitas judi *online*. Selain itu, terdapat pula kata "bandar judol" yang mengacu pada orang yang menyelenggarakan perjudian, serta kata-kata seperti "narkotika", "psikotropika", dan "obat terlarang" yang menggambarkan dampak negatif yang timbul akibat aktivitas perjudian ini.

IV. KESIMPULAN

Penelitian ini membandingkan kinerja tiga algoritma *machine learning*, yaitu *Naïve Bayes, Random Forest,* dan *Logistic Regression*, untuk analisis sentimen pada topik judi *online*. Dataset terdiri dari 5744 data ulasan, yang setelah melalui proses prapemrosesan menjadi 4592 data. Untuk mengatasi ketidakseimbangan kelas sentimen, diterapkan metode SMOTE. Evaluasi dilakukan dengan membandingkan metrik akurasi, presisi, *recall*, dan F1-*score* sebelum dan sesudah penerapan SMOTE. Hasil evaluasi menunjukkan bahwa Random Forest memberikan performa terbaik dengan akurasi 78%, diikuti oleh *Naïve Bayes* dan *Logistic Regression* yang masing-masing mencapai akurasi 77%. *Random Forest* unggul dalam mengklasifikasikan sentimen positif dan negatif secara konsisten, sedangkan *Naïve Bayes* menunjukkan peningkatan *recall* signifikan pada sentimen netral dari 045 menjadi 0,82 setelah SMOTE diterapkan. *Logistic Regression* cenderung mengalami penurunan presisi pada sentimen netral setelah SMOTE, membuatnya kurang optimal dibandingkan dua algoritma lainnya. Dengan demikian, *Random Forest* direkomendasikan sebagai algoritma terbaik untuk analisis sentimen judi *online*, sementara *Naïve Bayes* cocok digunakan untuk fokus pada sentimen netral. Penelitian ini menyoroti pentingnya teknik SMOTE dalam meningkatkan performa algoritma pada dataset yang tidak seimbang, sekaligus memberikan panduan dalam

memilih algoritma yang tepat untuk analisis sentimen di domain spesifik seperti judi *online*. Penelitian ini juga memperkaya pemahaman tentang penerapan *machine learning* dalam analisis sentimen, khususnya untuk isu sosial yang rumit seperti judi *online*. Temuan dari penelitian ini bisa digunakan oleh praktisi dan pembuat kebijakan untuk membuat keputusan yang lebih tepat dalam mengatur perjudian *online*. Kebijakan yang lebih peka terhadap sentimen masyarakat dapat mencakup regulasi yang lebih ketat terhadap platform judi *online,* seperti pembatasan akses atau iklan, serta perlindungan pengguna yang lebih baik.

REFERENSI

[1] R. Sepatia, T. R. Zarzani, and M. Purba, "ANALISIS YURIDIS PERTANGGUNGJAWABAN PIDANA BAGI PEMBUATWEBSITE YANG DIPERGUNAKAN UNTUK PERJUDIAN ONLINE(Analisis Putusan No. 852/Pid. Sus/2020/PN. Mdn)," *J. Rectum*, vol. 4, no. 852, pp. 430–442, 2022.

[2] S. N. I. Mutia Nurdiana, Nurul Aisyah, "Fenomena judi online di daerah jakarta selatan," vol. 2, no. 1, pp. 105–110, 2023.

[3] A. Laras, N. Salvabillah, C. Caroline, J. D. H, F. Dinda, and M. Finanto, "Analisis Dampak Judi Online di Indonesia Fakultas Psikologi ; Universitas Bhayangkara Jakarta Raya," vol. 3, no. 2, pp. 320–331, 2024.

[4] A. Maulana, A. Yuliana, T. Bandung, J. Politeknik, J. Pesantren, and K. Cimahi, "ANALISIS SENTIMEN OPINI PUBLIK TERKAIT JUDI ONLINE MENGGUNAKAN ALGORITMA NAÏVE BAYES DAN SUPPORT VECTOR MECHINE," vol. 12, no. 3, pp. 3706–3714, 2024.

[5] R. Antonius, A. R. Zulkarnain, and H. Irsyad, "Pendekatan TF-IDF , SMOTE , dan SVM dalam Klasifikasi Sentimen Masyarakat terhadap Pemblokiran Judi Online," vol. 2, no. 3, pp. 115–122, 2024, doi: 10.58369/biit.v2i3.65.

[6] L. S. H. Aji Dewo Pangestu, "Analisis Sentimen Terkait Judi Online di Media Sosial Instagram Menggunankan Naïve Bayes," vol. 3, 2025.

[7] U. Suriani, "Penerapan Data Mining untuk Memprediksi Tingkat Kelulusan Mahasiswa Menggunakan Algoritma Decision Tree C4.5," vol. 3, no. 2, pp. 55–66, 2023.

[8] A. Agung, A. Daniswara, and I. K. D. Nuryana, "Data Preprocessing Pola Pada Penilaian Mahasiswa Program Profesi Guru," vol. 05, pp. 97–100, 2023.

[9] R. Y. Lesmana and R. Andarsyah, "Model Klasifikasi Multinomial Naïve Bayes Untuk Analisis Sentiment Terkait Non-Fungible Token," *Inform. J. Tek.*, vol. 14, no. 3, pp. 135–139, 2022.

[10] I. Amelia *et al.*, "Analisis sentimen opini publik terhadap pengambil alihan tmii oleh pemerintah dengan algoritma naïve bayes," vol. 7, no. 2, pp. 142–148, 2023.

[11] G. A. Lustiansyah *et al.*, "Analisis klasifikasi sentimen pengguna aplikasi pedulilindungi berdasarkan ulasan dengan menggunakan metode long short term memory," pp. 630–639, 2022.

[12] C. Meilany, F. Andriani, and T. Kontingensi, "Klasifikasi Waiting Time for Pilot di Pelabuhan Tanjung Perak Menggunakan Metode Regresi Logistik – Synthetic Minority Oversampling Technique (SMOTE)," vol. 12, no. 1, 2023.

[13] B. B. Suherman, "Sistem Pakar Diagnosa Penyakit Dan Hama Pada Tanaman Jagung Menggunakan Metode Naive Bayes," *J. Inform. dan Rekayasa Perangkat Lunak*, vol. 2, no. 3, pp. 390–398, 2021, doi: 10.33365/jatika.v2i3.1251.

[14] Y. S. Triyantono, S. Al Faraby, and M. Dwifebri, "Analisis Sentimen Terhadap Ulasan Film Menggunakan Algoritma Random Forest," *eProceedings Eng.*, vol. 8, no. 4, p. 4136, 2021.

[15] M. I. Gunawan, D. Sugiarto, and I. Mardianto, "Peningkatan Kinerja Akurasi Prediksi Penyakit Diabetes Mellitus Menggunakan Metode Grid Seacrh pada Algoritma Logistic Regression," vol. 6, no. 3, pp. 280–284, 2020.