KLASIFIKASI SENTIMEN MASYARAKAT TERHADAP KINERJA KEMKOMDIGI TERKAIT PERMASALAHAN JUDI ONLINE PADA MEDIA SOSIAL INSTAGRAM MENGGUNAKAN METODE RANDOM FOREST DAN XGBOOST

(Skripsi)

Oleh

IKA RAHMA ALIA
NPM 2117051016

FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS LAMPUNG
BANDAR LAMPUNG
2025

ABSTRAK

KLASIFIKASI SENTIMEN MASYARAKAT TERHADAP KINERJA KEMKOMDIGI TERKAIT PERMASALAHAN JUDI ONLINE PADA MEDIA SOSIAL INSTAGRAM MENGGUNAKAN METODE RANDOM FOREST DAN XGBOOST

Oleh

IKA RAHMA ALIA

Judi online di Indonesia menjadi salah satu isu sosial yang menimbulkan beragam tanggapan masyarakat. Kementerian Komunikasi dan Digital (Kemkomdigi) telah melakukan berbagai upaya, namun kinerjanya tetap menuai kritik maupun apresiasi publik yang banyak disuarakan melalui media sosial, khususnya Instagram. Penelitian ini bertujuan untuk mengidentifikasi sentimen masyarakat terhadap Kemkomdigi serta membandingkan kinerja algoritma *Machine Learning Random Forest* dan *XGBoost* dalam proses klasifikasi. Sebanyak 724 komentar dikumpulkan dari akun Instagram resmi Kemkomdigi dan diberi label manual oleh tiga anotator. Data kemudian melalui tahapan praproses teks (pembersihan, *case folding*, tokenisasi, normalisasi, penghapusan *stopword*, dan *stemming*), representasi fitur dengan *TF-IDF*, serta penyeimbangan kelas menggunakan *Random Oversampling*. Model dibangun dengan data latih dan data uji berbanding 70:30, diuji dengan parameter *default* dan *hyperparameter tuning* menggunakan *GridSearchCV*, serta dievaluasi melalui akurasi, presisi, *recall*, dan *F1-score*. Hasil penelitian memperlihatkan bahwa algoritma *Random Forest* memiliki kinerja lebih unggul dibandingkan *XGBoost*. Setelah dilakukan *tuning*, *Random Forest* mencapai akurasi 71,64% dengan *F1-score* 70,90%, sedangkan XGBoost hanya mencapai akurasi 67,66% dengan *F1-score* 67,85%. Temuan ini menunjukkan bahwa *Random Forest* lebih efektif dalam mengklasifikasikan sentimen masyarakat terkait kinerja Kemkomdigi terhadap isu judi online.

Kata Kunci: Analisis Sentimen, Judi Online, Klasifikasi, *Random Forest*, *TF-IDF*, *XGBoost.*

ABSTRACT

CLASSIFICATION OF PUBLIC SENTIMENT TOWARDS THE PERFORMANCE OF THE MINISTRY OF COMMUNICATION AND INFORMATION TECHNOLOGY REGARDING ONLINE GAMBLING ISSUES ON INSTAGRAM SOCIAL MEDIA USING THE RANDOM FOREST AND XGBOOST METHODS

By

IKA RAHMA ALIA

Online gambling in Indonesia has become a major social issue that generates diverse public responses. The Ministry of Communication and Digital Affairs (Kemkomdigi) has undertaken various measures; however, its performance continues to receive both criticism and appreciation, which are widely expressed on social media, particularly Instagram. This study aims to identify public sentiment toward Kemkomdigi and to compare the performance of two machine learning algorithms, Random Forest and XGBoost, in sentiment classification. A total of 724 comments were collected from Kemkomdigi's official Instagram account and manually labeled by three annotators. The dataset underwent several text preprocessing steps (cleaning, case folding, tokenization, normalization, stopword removal, and stemming), feature representation using TF-IDF, and class balancing with Random Oversampling. The models were trained and tested with a 70:30 ratio, evaluated under both default parameters and hyperparameter tuning via GridSearchCV, and assessed using accuracy, precision, recall, and F1-score. The results show that Random Forest outperformed XGBoost in terms of classification performance. After tuning, Random Forest achieved an Accuracy of 71.64% with an F1-score of 70.90%, while XGBoost only obtained an Accuracy of 67.66% with an F1-score of 67.85%. These findings indicate that Random Forest is more effective in classifying public sentiment regarding Kemkomdigi's performance in addressing the issue of online gambling.

Keywords: *Classification, Online Gambling, Random Forest, Sentiment Analysis, TF-IDF, XGBoost.*

KLASIFIKASI SENTIMEN MASYARAKAT TERHADAP KINERJA KEMKOMDIGI TERKAIT PERMASALAHAN JUDI ONLINE PADA MEDIA SOSIAL INSTAGRAM MENGGUNAKAN METODE RANDOM FOREST DAN XGBOOST

Oleh

IKA RAHMA ALIA

Skripsi

Sebagai Salah Satu Syarat untuk Mencapai Gelar SARJANA ILMU KOMPUTER

Pada

Jurusan Ilmu Komputer
Fakultas Matematika dan Ilmu Pengetahuan Alam

FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS LAMPUNG
BANDAR LAMPUNG
2025

Judul Skripsi	:	**KLASIFIKASI SENTIMEN MASYARAKAT TERHADAP KINERJA KEMKOMDIGI TERKAIT PERMASALAHAN JUDI ONLINE PADA MEDIA SOSIAL INSTAGRAM MENGGUNAKAN METODE RANDOM FOREST DAN XGBOOST**
Nama Mahasiswa	:	Ika Rahma Alia
Nomor Pokok Mahasiswa	:	2117051016
Program Studi	:	Ilmu Komputer (S1)
Fakultas	:	Matematika dan Ilmu Pengetahuan Alam

MENYETUJUI

1. Komisi Pembimbing

Favorisen R. Lumbanraja, Ph.D
NIP. 198301102008121002

2. Mengetahui

Ketua Jurusan Ilmu Komputer

Dwi Sakethi, S.Si., M.Kom.
NIP. 196806111998021001

Ketua Program Studi
S1 Ilmu Komputer

Tristiyanto, M.I.S., Ph.D.
NIP. 198104142005011001

MENGESAHKAN

1. Tim Penguji

   Ketua : **Favorisen R. Lumbanraja, Ph.D.** .....

   Penguji

   Bukan Pembimbing I : **Dr. Aristoteles, S.Si., M.Si.** .....

   Penguji

   Bukan Pembimbing II : **Rico Andrian, S.Si., M.Kom.** .....

2. Dekan Fakultas Matematika dan Ilmu Pengetahuan Alam

Dr. Eng. Heri Satria, S.Si., M.Si.
NIP. 197110012005011002

Tanggal Lulus Ujian Skripsi: 12 Agustus 2025

PERNYATAAN

Saya yang bertanda tangan di bawah ini:

Nama : Ika Rahma Alia

NPM : 2117051016

Menyatakan bahwa skripsi saya yang berjudul "**Klasifikasi Sentimen Masyarakat terhadap Kinerja Kemkomdigi terkait Permasalahan Judi Online pada Media Sosial Instagram Menggunakan Metode Random Forest dan XGBoost**" merupakan karya saya sendiri dan bukan karya orang lain. Seluruh tulisan yang tertulis dalam skripsi ini telah mengikuti kaidah penulisan karya ilmiah Universitas lampung. Apabila di kemudian hari terbukti skripsi saya merupakan hasil penjiplakan atau dibuat orang lain, maka saya bersedia menerima sanksi berupa pencabutan gelar yang telah saya terima.

Bandar Lampung, 28 Agustus 2025

Yang menyatakan

Ika Rahma Alia

NPM. 2117051016

RIWAYAT HIDUP

Penulis bernama Ika Rahma Alia bertempat lahir di Pajar Bulan pada tanggal 23 Oktober 2003, sebagai anak pertama dari dua bersaudara dari pasangan Bapak Miskam dan Ibu Acih. Penulis menyelesaikan pendidikan dasar di SD Negeri 3 Pajar Bulan pada Tahun 2015. Kemudian melanjutkan pendidikan menengah pertama di SMP Negeri 1 Way Tenong yang diselesaikan pada Tahun 2018, lalu menyelesaikan pendidikan menengah atas di SMA Negeri 1 Way Tenong pada Tahun 2021.

Tahun 2021 penulis terdaftar sebagai mahasiswa Jurusan Ilmu Komputer Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Lampung melalui jalur SNMPTN. Selama menjadi mahasiswa, penulis melakukan beberapa kegiatan antara lain.

1. Menjadi anggota Adapter Himpunan Mahasiswa Jurusan Ilmu Komputer pada periode 2021/2022.
2. Menjadi anggota Biro Kesekretariatan Himpunan Mahasiswa Jurusan Ilmu Komputer pada Tahun 2022.
3. Menjadi bendahara Biro Kesekretariatan Himpunan Mahasiswa Jurusan Ilmu Komputer pada Tahun 2023.
4. Mengikuti pelatihan UI/UX.
5. Melaksanakan Kerja Praktik di Badan Kepegawaian Daerah Provinsi Lampung pada periode I Tahun 2024.
6. Melaksanakan KKN di Desa Mengandung Sari Kecamatan Sekampung Udik Kabupaten Lampung Timur pada periode II Tahun 2024.

MOTTO

“Dan janganlah kamu menyerah atas apa yang telah kamu mulai”
(Novel Lautan Selat Gibraltar)

“Dia yang berani hari ini, pernah ketakutan kemarin.
Ambil resikonya, atau kamu tidak akan ke mana-mana.
Fortis fortuna adiuvat, keberuntungan berpihak pada yang berani.”
(*Anonymous*)

PERSEMBAHAN

Alhamdulillahirobbilalamin

Puji dan syukur tercurahkan kepada Allah Subhanahu Wa Ta'ala atas segala Rahmat dan Karunia-Nya sehingga saya dapat menyelesaikan skripsi ini. Shalawat serta salam selalu tercurahkan kepada Nabi Muhammad Shallallahu Alaihi Wasallam.

Kupersembahkan karya ini kepada:

Kedua Orang Tua dan Adik Tercinta

Sumber kekuatan dan cinta yang selalu memberikan pengorbanan, perjuangan, kasih sayang, perhatian, dukungan dan do'a yang tak pernah putus.

Keluarga dan Almh. Nenek Tercinta

Atas segala dukungan, kasih sayang, perhatian dan do'a yang selalu menyertaiku.

Seluruh Keluarga Besar Ilmu Komputer 2021

Yang senantiasa memberikan semangat dan dukungan**.**

Almamater Tercinta, Universitas Lampung dan Jurusan Ilmu Komputer

Tempat bernaung mengemban semua ilmu untuk menjadi bekal kehidupan.

SANWACANA

Puji syukur kehadirat Allah Subhanahu Wa Ta'ala, karena telah memberikan limpahan nikmat, rahmat dan karunia-Nya. Shalawat serta salam semoga senantiasa tercurahkan kepada junjungan Nabi Muhammad SAW, sehingga penulis dapat menyelesaikan skripsi yang berjudul "**Klasifikasi Sentimen Masyarakat terhadap Kinerja Kemkomdigi terkait Permasalahan Judi Online pada Media Sosial Instagram Menggunakan Metode Random Forest dan XGBoost**" dengan baik dan lancar.

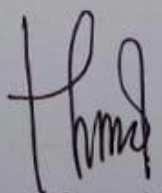
Selesainya skrispsi ini tidak terlepas dari bantuan, bimbingan, dan dukungan dari berbagai pihak. Oleh karena itu, ucapan terima kasih ditujukan kepada:

1. Bapak, Ibu dan keluarga yang selalu mendoakan yang terbaik, memberi dukungan, kasih sayang dan selalu memberikan semangat baik secara moral maupun material dalam menyelesaikan pendidikan dan skripsi ini.
2. Ibu Prof. Dr. Ir. Lusmelia Afriani, D.E.A., I.P.M. ASEAN Eng. selaku Rektor Universitas Lampung.
3. Bapak Dr. Eng. Heri Satria, S.Si., M.Si. selaku Dekan Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Lampung.
4. Bapak Dwi Sakethi, S.Si., M.Kom. selaku Ketua Jurusan Ilmu Komputer FMIPA Universitas Lampung.
5. Ibu Yunda Heningtyas, M. Kom. selaku Sekretaris Jurusan Ilmu Komputer FMIPA Universitas Lampung.
6. Bapak Tristiyanto, S. Kom., M.I.S., Ph.D. selaku Ketua Program Studi Ilmu Komputer FMIPA Universitas Lampung.
7. Ibu Dewi Asiah Shofiana, S.Komp., M.Kom. selaku Pembimbing Akademik, yang senantiasa memberikan dukungan dan arahan dalam pengembangan akademik selama masa studi.

8. Bapak Favorisen R. Lumbanraja, Ph. D selaku Pembimbing Utama dalam penelitian ini yang selalu membimbing, memberikan arahan, masukan dan saran dalam penyelesaian skripsi
9. Bapak Dr. Aristoteles S. Si., M., Si. selaku Pembahas Pertama yang telah memberikan masukan serta saran yang bermanfaat dalam perbaikan skripsi ini.
10. Bapak Rico Andrian S. Si., M. Kom. sebagai Pembahas Kedua yang telah memberikan masukan serta saran yang bermanfaat dalam perbaikan skripsi ini.
11. Seluruh Dosen, Staf dan Karyawan Jurusan Ilmu Komputer yang telah memberikan ilmu, pelajaran dan bantuan terbaik selama penulis menempuh pendidikan di Jurusan Ilmu Komputer Universitas Lampung.
12. Teman seperjuangan semasa kuliah Siska Hermayanti, Aprilia Anggun Sari Rahmawati, Gilang Ramadhan, Sahabat Jannah, teman sepembimbingan dan Kelompok KKN Mengandung Sari yang selalu mendukung, menemani, dan berbagi cerita indah selama masa perkuliahan.
13. Teman-teman Himakom yang sudah mengajarkan banyak hal dalam berorganisasi dan memberikan pengalaman yang berharga.
14. Keluarga Ilmu Komputer 2021 yang telah memberikan pengalaman yang sangat berarti selama menjalankan studi di Jurusan Ilmu Komputer Universitas Lampung.

Penulis menyadari bahwa penyusunan skripsi ini masih jauh dari kata sempurna. Namun penulis sangat mengharapkan skripsi ini dapat bermanfaat bagi para civitas akademik Universitas Lampung pada umumnya dan mahasiswa Ilmu Komputer pada khususnya.

Bandar Lampung, 28 Agustus 2025



Ika Rahma Alia
2117051016

DAFTAR ISI

Halaman

DAFTAR GAMBAR

Gambar Halaman

DAFTAR TABEL

Tabel Halaman

I. PENDAHULUAN

1.1 Latar Belakang

Judi online menjadi fenomena global yang menarik perhatian banyak orang di seluruh dunia, termasuk Indonesia. Perjudian adalah permainan dimana pemain bertaruh bahwa hanya ada satu pilihan yang benar jika mereka memilih satu dari banyak pilihan (Jadidah *et al*., 2023). Masalah utama yang terhubung dengan kegiatan perjudian online meliputi utang, kebocoran informasi pribadi, dan tindakan penipuan yang merugikan para pemain judi online. Menurut laporan dari lembaga Pusat Pelaporan dan Analisis Transaksi Keuangan (PPATK), diperkirakan ada 3,2 juta hingga 4 juta penduduk Indonesia yang terlibat dalam permainan judi online dengan total uang yang digunakan mencapai 327 triliun pada tahun 2023. Sementara itu, pada kuartal pertama 2024, transaksi judi online meningkat hingga 100 triliun (Pusat Pelaporan dan Analisis Transaksi Keuangan, 2024).

Pemerintah Indonesia, melalui Kementerian Komunikasi dan Digital (Kemkomdigi), telah melakukan berbagai upaya untuk menghentikan akses ke situs-situs yang berkaitan dengan judi online melalui pemblokiran, penyuluhan, dan kampanye literasi digital. Kementerian Komunikasi dan Digital (Kemkomdigi) telah memblokir 227.811 konten yang berkaitan dengan judi online terhitung pada sejak 20 Oktober hingga 5 November 2024 pukul 06.00 WIB. (Zulaikha, 2024). Namun, belakangan ini beberapa pegawai di Kementerian Komunikasi dan Digital (Kemkomdigi), yang memiliki kewenangan untuk memblokir situs-situs perjudian justru terlibat dalam kasus judi online. Alih-alih memblokir situs tersebut, oknum-oknum ini diduga melindungi akses ke situs judi online (Bestari, 2024).

Dengan adanya pencapain dan keterlibatan Kemkomdigi dalam kasus judi online, menimbulkan berbagai opini masyarakat terkait integritas Kemkomdigi dalam menjalankan fungsinya sebagai pengawas dan penindak terhadap kasus perjudian online. Komentar positif dan negatif yang berkembang di masyarakat ini terlihat jelas di media sosial, seperti Instagram, di mana pengguna menyuarakan opini mereka terkait kinerja Kemkomdigi.

Media sosial Instagram menjadi salah satu platform yang paling aktif digunakan oleh masyarakat untuk berbagi foto, mengungkapkan pendapat, kritik, dan opini (Luqyana *et al*., 2018). Instagram dipilih karena memiliki jangkauan luas dan memungkinkan interaksi langsung antar pengguna, baik dalam bentuk komentar, unggahan, maupun *hashtag*. Opini yang berkembang di platform ini sangat beragam, mencakup sentimen positif maupun negatif, yang sangat dipengaruhi oleh pemberitaan tentang kinerja Kemkomdigi dalam kasus judi online.

Analisis sentimen masyarakat terhadap kinerja Kemkomdigi terkait permasalahan judi online menjadi sangat penting. Melalui analisis sentimen, tujuan utama adalah mengklasifikasikan teks tersebut menjadi positif, atau negatif. Sehingga, pemahaman yang mendalam tentang opini publik dapat membantu pemerintah dan pembuat kebijakan dalam mengevaluasi efektifitas tindakan yang diambil. Kemajuan dalam bidang kecerdasan buatan dan pembelajaran mesin (*Machine Learning*) telah membuka peluang baru dalam analisis sentimen. Teknik-teknik seperti *Random Forest* dan *XGBoost* memiliki keunggulan dalam menangani data yang kompleks dan besar (Hendrawan, 2022).

Dalam beberapa penelitian sebelumnya, analisis sentimen telah banyak digunakan. Penelitian yang dilakukan oleh Anggraini *et al*., (2024) dalam mengklasifikasikan data komentar menghasilkan nilai akurasi yang cukup baik. Dalam penelitian tersebut data yang digunakan merupakan komentar dari sosial media Reddit yang disebut *GoEmotion*. Data yang digunakan berjumlah 7.325 yang terbagi menjadi 5 kategori yaitu marah 1.778, takut 1.697, gembira 1.697, cinta 1.457, dan sedih 696 data. Data dibagi menjadi 80% data pelatihan dan 20% data uji. Model yang

digunakan yaitu *Random Forest*, *XGBoost*, dan *LightGBM*. Nilai akurasi terbaik dihasilkan *Random Forest* mencapai Akurasi, Presisi, dan *Recall* sebesar 0,85. Untuk model *LightGBM*, memberikan Akurasi 0,70, Presisi 0,71, dan *Recall* 0,70. Sedangkan model *XGBoost* mencapai Akurasi, Presisi, dan *Recall* sebesar 0,74.

Adapun penelitian yang dilakukan oleh Hendrawan (2022) membandingkan tiga algoritma klasifikasi teks, yaitu *Naïve Bayes*, *Support Vector Machine* (SVM), dan *XGBoost,* untuk analisis sentimen masyarakat terhadap produk lokal di Indonesia. Hasil penelitian menunjukkan bahwa kombinasi algoritma *XGBoost* dengan *Word2vec* memberikan hasil terbaik dengan skor F1 sebesar 0,94, lalu *XGBoost* dengan *TF-IDF* yang menghasilkan skor F1 sebesar 0,94. Pada algoritma SVM, penggunaan *TF-IDF* dan *Word2vec* menghasilkan skor F1 sebesar 0,94 dan 0,94. Sedangkan algoritma *Naïve Bayes* menunjukkan performa yang lebih rendah, dengan *F1-score* sebesar 0,92 menggunakan *TF-IDF* dan 0,90 dengan *Word2vec*.

Penelitian ini berfokus pada analisis sentimen masyarakat kinerja Kemkomdigi dalam menangani kasus judi online. Dengan pendekatan *Machine Learning*, metode *Random Forest* dan *XGBoost* digunakan untuk mengklasifikasikan sentimen berdasarkan data komentar di media sosial Instagram. Melalui analisis ini, diharapkan dapat dihasilkan pemahaman yang lebih mendalam tentang perspektif masyarakat terhadap kinerja Kemkomdigi terhadap kasus judi online.

1.2 Rumusan Masalah

Berikut adalah rumusan masalah penelitian ini berdasarkan latar belakang yang telah dijelaskan:

1. Bagaimana mengimplementasikan metode *Random Forest* dan *XGBoost* dalam mengklasifikasikan sentimen masyarakat terhadap kinerja Kemkomdigi terkait permasalahan judi online?
2. Bagaimana hasil perbandingan metode *Random Forest* dan *XGBoost* dalam mengklasifikasikan sentimen masyarakat terhadap kinerja Kemkomdigi terkait permasalahan judi online?

1.3 Batasan Masalah

Batasan masalah dari penelitian ini adalah sebagai berikut:

1. Data yang digunakan dalam penelitian ini terbatas pada komentar dan opini masyarakat yang diperoleh dari media sosial Instagram Kemkomdigi dalam bahasa Indonesia yang berkaitan dengan judi online.
2. Metode yang digunakan yaitu metode *Random Forest* dan *XGBoost*.
3. Klasifikasi sentimen difokuskan pada dua kelas, yaitu positif dan negatif.

1.4 Tujuan Penelitian

Tujuan dari penelitian ini adalah:

1. Mengetahui hasil perbandingan metode *Random Forest* dan *XGBoost* dalam klasifikasi sentimen masyarakat terhadap kinerja Kemkomdigi terkait permasalahan judi online.
2. Mengukur performa metode *Random Forest* dan *XGBoost* dalam klasifikasi sentimen masyarakat terhadap terhadap kinerja Kemkomdigi terkait permasalahan judi online.

1.5 Manfaat Penelitian

Hasil dari penelitian ini diharapkan dapat memberikan manfaat sebagai berikut:

1. Mengetahui perbandingan sentimen masyarakat terhadap kinerja Kemkomdigi terkait permasalahan judi online.
2. Menyediakan dasar bagi penelitian lanjutan yang dapat mengembangkan model klasifikasi sentimen yang lebih optimal.

II. TINJAUAN PUSTAKA

2.1 Penelitian Terdahulu

Sebagai dasar dalam penelitian ini, dilakukan tinjauan literatur terhadap sejumlah studi sebelumnya yang relevan yang dapat dilihat pada Tabel 1.

Tabel 1. Penelitian terdahulu terkait dengan klasifikasi sentimen

No	Penelitian	Data	Metode	Hasil
1	Penerapan Algoritma *Random Forest* Untuk Analisis Sentimen Komentar Di YouTube Tentang Islam fobia (Afdhal *et al*., 2022)	**Komentar Youtube** Jumlah: 1000 Positif: 631 Negatif: 369	*Word Embedding: TF-IDF* *Preprocessing: Cleaning, Case Folding, Tokenizing, Stopword Removal,* dan *Stemming* Model: *Random Forest* *Splitting* Data: *Training*: 90% *Testing*: 10% Evaluasi: *Confusion Matrix*	*Accuracy*: 79% *Precision*: 79% *Recall*: 95% *F1-Score*: 86%

Tabel 1. Lanjutan

No	Penelitian	Data	Metode	Hasil
2	Perbandingan Algoritma *Naïve Bayes*, Svm Dan *XGBoost* Dalam Klasifikasi Teks Sentimen Masyarakat Terhadap Produk Lokal Di Indonesia (Hendrawan, 2022)	**Ulasan Online Produk Di Aplikasi Shopee** Jumlah: 25.851 Baik: 8.141 Buruk: 2.323	*Word Embedding:* *Word2vec* dan *TF-IDF* *Preprocessing:* *Case folding, remove punctuation, stopword removal, word normalize, stemming, tokenization* *Splitting* Data*:* *Training:* 80% *Testing*: 20% Model: *Naïve Bayes*, SVM, dan *XGBoost* Evaluasi: F1-*Score*	*Naïve Bayes* + *Word2vec*: 90% *Naïve Bayes* + *TF-IDF*:91%. SVM + *Word2vec*: 94% SVM + *TF-IDF*: 94% *XGBoost* + *Word2vec*: 94% *XGBoost* + *TF-IDF*: 94%
3	*A Comparative Analysis of Random Forest, XGBoost, and LightGBM Algorithms for Emotion Classification in Reddit Comments* (Anggraini *et al*., 2024)	**Komentar Reddit (GoEmotions)** Jumlah: 7.325 Marah: 1.778 Takut: 1.697 Gembira: 1.697 Cinta: 1.457 Sedih: 696	*Word Embedding:* *TF-IDF* *Preprocessing:* *Case folding, cleaning, selection data, augmentation data, tokenize, stopword removal, stemming, dan label encoding* *Splitting* Data: *Training*: 80%	*Random Forest* *Accuracy*: 86% *Precision*: 86% *Recall*: 86% F1-*Score*: 86% *XGBoost* *Accuracy*: 74% *Precision*: 74% *Recall*: 74% F1-*Score*: 74%

Tabel 1. Lanjutan

No	Penelitian	Data	Metode	Hasil
			***Testing*: 20%** Model: *Random Forest*, *LightGBM* dan *XGBoost* Evaluasi: *Confusion matrix*	***LightGBM:*** ***Accuracy: 71%*** *Precision*: 71% *Recall*: 71% F1-*Score*: 71%
4	Perbandingan Metode *Decision Tree* dan *Support Vector Machine* untuk Analisis Sentimen pada Instagram Mengenai Kinerja PSSI (Asshiddiqi & Lhaksmana, 2020)	**Komentar Instagram** Jumlah: 2.750 Positif: 1.714 Negatif: 1.043	*Word embedding: TF-IDF* *Preprocessing*: *Case folding, data cleaning, tokenizing, word repair, stopword removal,* dan *stemming* Splitting data: *Training:* 80% *Testing*: 20% Model: *Decision Tree* dan *Support Vector* Evaluasi: *Confusion matrix*	*Decision Tree* dengan *TF-IDF*: *Accuracy*: 87% P*recision:* 88% *Recall:* 92% F1-*Score* 89% *SVM* dengan *TF-IDF*: *Accuracy:* 94%, P*recision:* 97%, *Recall:* 94% F1-*Score:* 95%.

Penelitian yang dilakukan oleh Afdhal *et al*. (2022) mengenai analisis sentimen komentar di YouTube tentang islamfobia menghasilkan nilai Akurasi sebesar 79% dengan F1-*Score* sebesar 86%. Data yang digunakan pada penelitian ini yaitu 1.000 komentar yang dikumpulkan dari youtube. Dari data tersebut, 631 komentar merupakan data komentar positif dan 369 merupakan data komentar negatif. Klasifikasi data komentar dilakukan menggunakan model *Random Forest*. Pada

tahap *preprocessing* dilakukan proses *cleaning, case folding, tokenizing, stopword removal,* dan *stemming.* Selain itu penelitian tersebut menggunakan *TF-IDF* untuk menghitung frekuensi kemunculan setiap kata dalam dokumen. Pada penelitian ini dilakukan 125 percobaan terhadap masing-masing *splitting* data dengan mengkomunikasikan nilai dari 3 parameter yaitu *n_estimators, max_dept*h, dan *min_samples_split.* Akurasi tertinggi didapatkan sebesar 79% dengan pembagian data 90% : 10% dengan kombinasi nilai parameter *n_estimators* = 10, *max_depth* = 25 dan *min_samples split* = 10. Dari 100 data uji, model mampu mengklasifikasikan 11 data berlabel negatif dan 68 data berlabel positif. Hasil klasifikasi dengan model tersebut menghasilkan nilai Akurasi sebesar 79%, *Precision* 79%, *Recall* 95%, dan F1-*Score* sebesar 86%.

Penelitian yang dilakukan oleh Hendrawan (2022) membahas tentang kinerja algoritma *Naive Bayes*, SVM, dan *XGBoost* dalam mengklasifikasikan teks sentimen terhadap produk lokal di indonesia. Data yang digunakan yaitu data teks ulasan dan rating pada web *Shopee Marketplace* yang didapatkan melalui proses *scrapping* lalu disimpan dalam format .csv sebanyak 25.581 data. Dari data tersebut, label baik berjumlah 8.141 sedangkan untuk label buruk berjumlah 2.323. Data tersebut melalui 7 tahap *preprocessing* yaitu *case folding, remove punctuation, remove number & short word, word normalize, stopword removal, tokenization,* dan *stemming.* Data dibagi menjadi 80% data latih dan 20% data uji. Evaluasi pada penelitian tersebut menggunakan F1-*Score*. Berdasarkan hasil penelitian tersebut kombinasi *Word2vec + XGBoost* menghasilkan F1-*Score* lebih tinggi yaitu sebesar 94% diikuti dengan *TF-IDF + XGBoost* 94%. Sementara untuk algoritma SVM dengan menggunakan *vector space TF-IDF* menghasilkan 94% dan *Word2vec* 94%. Sedangkan untuk Naïve Bayes memiliki F1-*Score* dengan *TF-IDF* dan 91% dengan *Word2vec* 90%.

Penelitian yang dilakukan oleh Anggraini *et al*. (2024) dalam mengklasifikasikan data komentar menghasilkan nilai Akurasi yang cukup baik. Dalam penelitian tersebut data yang digunakan merupakan komentar dari sosial media Reddit yang disebut *GoEmotion.* Data yang digunakan berjumlah 7.325 yang terbagi menjadi 5

kategori yaitu marah 1.778, takut 1.697, gembira 1.697, cinta 1.457, dan sedih 696 data. Data dibagi menjadi 80% data pelatihan dan 20% data uji. Pada tahap preprocessing data akan melewati beberapa proses seperti *case folding, cleaning, selection data, augmentation data, tokenize, stopword removal, stemming,* dan *label encoding.* Pada penelitian ini, digunakan *GridSearchCV* untuk mendapatkan *hyperparameter* terbaik untuk setiap model. Model yang digunakan yaitu *Random Forest*, *XGBoost*, dan *LightGBM*. Pada tahap evaluasi, digunakan validasi silang (*cross validation*) dan membandingkan dengan *confusion matrix* dari ketiga model. Nilai Akurasi terbaik dihasilkan *Random Forest* dengan parameter *n_estimators* = 1500 dan *max_features* = log2*,* mencapai Akurasi*, precision,* dan *Recall* sebesar 85%. Untuk model *LightGBM*, dikonfigurasi dengan *max_depth* = 2 , *learning_rate* = 0,1, dan *n_estimators* = 500, memberikan Akurasi 70%, *Precision* 71%, dan *Recall* 70%. Sedangkan model *XGBoost* dengan *max_depth* = 2, *learning_rate* = 0,8, dan *n_estimators* = 100, mencapai Akurasi, *precision*, dan *Recall* sebesar 74%.

Penelitian yang dilakukan oleh Asshiddiqi & Lhaksmana (2020) membahas tentang perbandingan metode *Decision Tree* dan *Support Vector Machine* untuk analisis sentimen kinerja PSSI. Data yang digunakan dalam penelitian tersebut adalah komentar instagram yang diperoleh melalui *scrapping* menggunakan aplikasi *Octoparse*. Data yang diperoleh berjumlah 2.750 data, 1.714 merupakan kelas negatif, 1.043 merupakan kelas positif. Data tersebut melalui tahapan *preprocessing* yang terdiri dari *case folding, data cleaning, tokenizing, word repair, stopword removal,* dan *stemming.* Rasio pembagian data pelatihan dan pengujian dibagi menjadi 5 komposisi rasio yaitu 90:10, 80:20, 70:30, 60:40 dan 50:50. Komposisi ini bertujuan mengetahui rasio berapa model bekerja dengan baik. Penelitian melakukan perbandingan pengujian dengan metode pembobotan TF, *TF-IDF*, dan TRF untuk setiap model. Kombinasi nilai Akurasi terbaik yaitu dengan rasio pembagian data 80% data *training* dan 20% data *testing*. *Decision Tree* dengan *TF-IDF* menghasilkan nilai Akurasi 87%, *Precision* 88%, *Recall* 92% dan F1-*Score* 89% sedangkan pada model SVM dengan *TF-IDF* menghasilkan nilai Akurasi 94%, *Precision* 97%, *Recall* 94% dan F1-*Score* 95%.

2.2 Landasan Teori

2.2.1 Analisis Sentimen

Analisis sentimen adalah proses pengidentifikasian dan pengklasifikasian pendapat atau perasaan yang terkandung dalam teks, biasanya untuk menentukan apakah sebuah pernyataan atau komentar memiliki sentimen positif atau negatif (Afdhal *et al*., 2022). Dalam *Natural Language Processing* (NLP), analisis sentimen digunakan untuk mengevaluasi teks dari berbagai sumber, seperti ulasan produk, komentar di media sosial, atau artikel berita untuk mengetahui sikap atau emosi yang diungkapkan oleh penulisnya. Sedangkan secara tidak langsung, analisis sentimen tersebut dapat dilakukan dengan suatu teknik yang disebut dengan *Text Mining.* Fokus utama dari analisis sentimen ini adalah pada pendapat pribadi yang mengekspresikan atau menyiratkan sentimen positif atau negatif (Salim & Syafrullah, 2023).

2.2.2 *Text Mining*

Text mining adalah proses otomatis atau semi-otomatis untuk mengekstrak informasi, pola, dan wawasan yang tersembunyi, tidak diketahui sebelumnya, dan berpotensi bernilai dari data teks tidak terstruktur, seperti dokumen, media sosial, atau pidato (Hassani *et al*., 2020). *Text mining* merupakan salah satu teknik yang dapat digunakan untuk melakukan klasifikasi sentimen, *text clustering*, *information extraction* dan *information retrieval*, dimana *text mining* merupakan variasi dari data mining yang berusaha menemukan pola yang menarik dari sekumpulan data tekstual yang berjumlah besar (Afdhal *et al*., 2022).

2.2.3 *Text Preprocessing*

Text preprocessing adalah kumpulan metode dan teknik yang digunakan untuk memanipulasi dan memproses teks. Tujuan dari *text preprocessing* adalah untuk mengumpulkan informasi penting dari teks dan mengubahnya menjadi format yang dapat diproses oleh komputer. Setiap proses dan urutan yang digunakan

dalam tahap *text preprocessing* tidak diatur. Semuanya tergantung pada hasil yang diinginkan dari data tersebut. *Text preprocessing* adalah proses membersihkan data sebelum diolah (Afdhal *et al*., 2022). Pada tahapan ini terdapat 5 proses, yaitu:

1) *Cleaning*, pada tahap ini dilakukan pemilihan atribut yang akan digunakan, membuang *noise* (angka, tanda baca, emoji, spasi ganda, dan baris enter). Contoh *cleaning* dapat dilihat pada Tabel 2.

   Tabel 2. Contoh *Cleaning*

Teks Awal	Judi online dan KORUPSI adalah perusak bangsa, TOLONG JANGAN KORUPSI 👍
Hasil *Cleaning*	Judi online dan KORUPSI adalah perusak bangsa TOLONG JANGAN KORUPSI

2) *Case Folding*, pada tahap ini dilakukan penyeragaman teks menjadi huruf kecil (*lowercase*). Contoh *case folding* dapat dilihat pada Tabel 3.

   Tabel 3. Contoh *Case Folding*

Teks Awal	Judi online dan KORUPSI adalah perusak bangsa TOLONG JANGAN KORUPSI
Hasil *Case Folding*	judi online dan korupsi adalah perusak bangsa tolong jangan korupsi

3) *Tokenizing*, pada tahap ini dilakukan pemecahan kata pada kalimat. Contoh *tokenizing* dapat dilihat pada Tabel 4.

   Tabel 4. Contoh Hasil *Tokenizing*

Teks Awal	judi online dan korupsi adalah perusak bangsa tolong jangan korupsi
Hasil *Tokenizing*	judi, online, dan, korupsi, adalah, perusak, bangsa, tolong, jangan, korupsi

4) *Stopword Removal*, pada tahap ini dilakukan penghilangan kata yang termasuk kedalam kategori *stopword*. *Stopword* merupakan kata yang sering muncul namun dianggap tidak memiliki arti. Contoh *stopword removal* dapat dilihat pada Tabel 5.

Tabel 5. Contoh Hasil *Stopword Removal*

Teks Awal	judi, online, dan, korupsi, adalah, perusak, bangsa, tolong, jangan, korupsi
Hasil *Stopword Removal*	judi, online, korupsi, perusak, bangsa, korupsi

5) *Stemming*, tahap ini dilakukan untuk menemukan kata dasar dengan menghilangkan semua imbuhan yang menyatu pada kata. Contoh *stemming* dapat dilihat pada Tabel 6.

Tabel 6. Contoh Hasil *Stemming*

Teks Awal	judi, online, korupsi, perusak, bangsa, korupsi
Hasil *Stemming*	judi, online, korupsi, rusak, bangsa, korupsi

2.2.4 *Word Embedding*

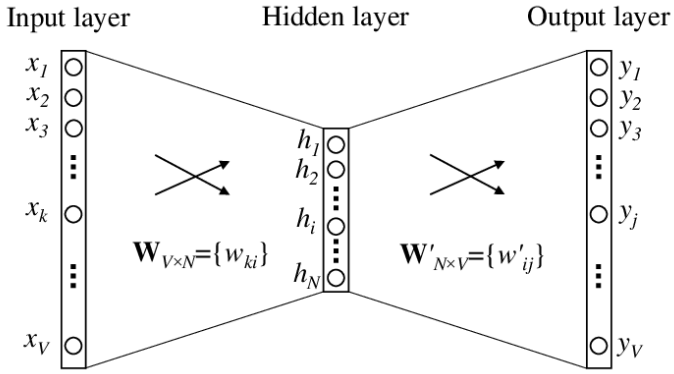
Setelah pembersihan data, langkah penting dalam klasifikasi teks adalah mengubah teks menjadi data numerik melalui proses *embedding* atau vektorisasi (Jilcha & Kwak, 2022). *Word embedding* adalah sebuah fungsi parameter yang memetakan setiap kata ke dalam vektor berdimensi tinggi (Nurdin *et al*., 2020). *Word embedding* merupakan vektor bernilai *real* yang mempresentasikan suatu kata yang dapat mewakili konteks di mana kata tersebut muncul. Berikut ini adalah beberapa metode *word embedding* yang umum digunakan:

1. *Word2Vec*

   Word2vec merupakan salah satu algoritma *word embedding* yang memetakan setiap kata dalam teks ke dalam vektor (Nurdin *et al*., 2020). *Word2vec* mempresentasikan kata dalam bentuk vektor yang mampu menangkap makna

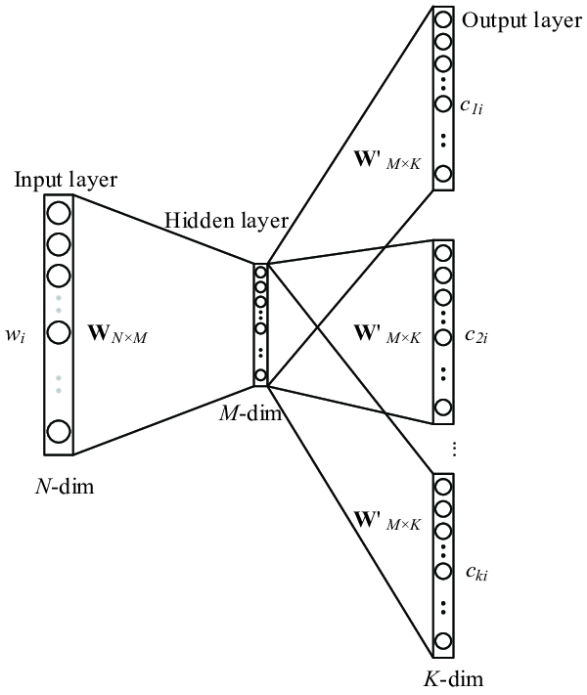
semantik dari sebuah kata. *Word2vec* termasuk implementasi *unsupervised learning* menggunakan jaringan saraf, yang terdiri dari lapisan tersembunyi (*hidden layer*) dan lapisan *fully connected*. *Word2Vec* memanfaatkan informasi lokal dalam bahasa, sehingga semantik yang dipelajari oleh suatu kata dipengaruhi oleh konteks kata-kata di sekitarnya (Purnasiwi *et al*., 2023) . Model ini juga menunjukkan kemampuan dalam mempelajari pola linguistik melalui hubungan linear antara vektor kata.

Secara umum, *Word2vec* adalah teknik yang menggabungkan dua algoritma berbasis prediksi yaitu *Skip-gram* dan *Continuous Bag of Word* (CBOW). CBOW memiliki waktu training lebih cepat dan memiliki akurasi yang sedikit lebih baik untuk *frequent words*. Model CBOW dapat dilihat pada Gambar 1.



Gambar 1. *One Word Context* CBOW (Nurdin *et al*., 2020).

Skip-gram memanfaatkan satu kata untuk memprediksi kata-kata dalam konteks targetnya. *Skip-gram* menunjukkan kinerja yang baik bahkan dengan jumlah data pelatihan yang terbatas dan mampu mempresentasikan kata-kata yang jarang muncul secara efektif. Model *skip-gram* dapat dilihat pada Gambar 2.



Gambar 2. *Skip-Gram* (Nurdin *et al*., 2020).

2. *TF-IDF*

TF-IDF adalah salah satu metode pembobotan sebuah kata didalam sistem pengolahan teks. *TF-IDF* menggabungkan *Term Frequency* (TF) dan *Inverse Document Frequency* (IDF) (Asshiddiqi & Lhaksmana, 2020). TF mengukur seberapa sering suatu kata muncul dalam sebuah dokumen, sedangkan IDF menghitung seberapa sedikit dokumen yang mengandung kata tersebut dalam korpus (Julianti *et al*., 2024). *TF-IDF* membantu memperkuat fitur teks berbasis frekuensi untuk meningkatkan akurasi dalam proses klasifikasi. Hasil akhirnya adalah model pembobotan yang memberikan nilai spesifik untuk setiap token berdasarkan pola penggunaanya dalam data pelatihan. Rumus IF dapat dilihat pada Persamaan 1, IDF dapat dilihat pada Persamaan 2, dan *TF-IDF* dapat dilihat pada Persamaan 3.

$$TF\ (\mathrm{t,d}) = \frac{f(t,d)}{N_d} \tag{1}$$

$$IDF(\mathrm{t}) = \log\left(\frac{N}{df(t)}\right) \tag{2}$$

$$TF\text{-}IDF = TF(t,d)\ \times IDF\ (t) \tag{3}$$

Dengan *TF(t,d)* adalah frekuensi munculnya *term t* pada dokumen *d*, *N* adalah jumlah kumpulan dokumen, *df(t)* adalah jumlah dokumen yang mengandung *term t*.

2.2.5 Judi Online

Judi online adalah salah satu topik yang sedang memanas di media sosial, berkat tersebar luasnya budaya perjudian di Indonesia melalui togel, lottery, taruhan permainan sepak bola, atau taruhan permainan mahjong (Antonius *et al*., 2024). Perjudian secara istilah adalah pertaruhan dengan sengaja yaitu mempertaruhkan satu nilai atau yang dianggap bernilai dengan menyadari adanya resiko dan harapan-harapan tertentu pada peristiwa-peristiwa permainan, pertandingan, perlombaan dan kejadian-kejadian yang tidak atau belum pasti hasilnya. Perjudian di Indonesia sudah ada sejak zaman penjajah Belanda (Jadidah *et al*., 2023).

2.2.6 *Machine Learning*

Machine Learning adalah cabang kecerdasan buatan yang berfokus pada pengembangan algoritma untuk memungkinkan komputer belajar dari data tanpa diprogram secara eksplisit (Retnoningsih & Pramudita, 2020). Dalam *Machine Learning*, model dilatih untuk mengenali pola dari data dan membuat prediksi berdasarkan pengalaman tersebut. Pendekatannya terbagi menjadi tiga kategori utama, yaitu *supervised learning* (dengan data berlabel), *unsupervised learning* (tanpa data berlabel), dan *reinforcement learning* (pembelajaran berbasis umpan balik dari lingkungan) (Roihan *et al*., 2020). Proses ini melibatkan tahapan pengumpulan data, pra-pemprosesan, pelatihan model, evaluasi, dan penerapan model untuk menyelesaikan masalah atau memprediksi data baru.

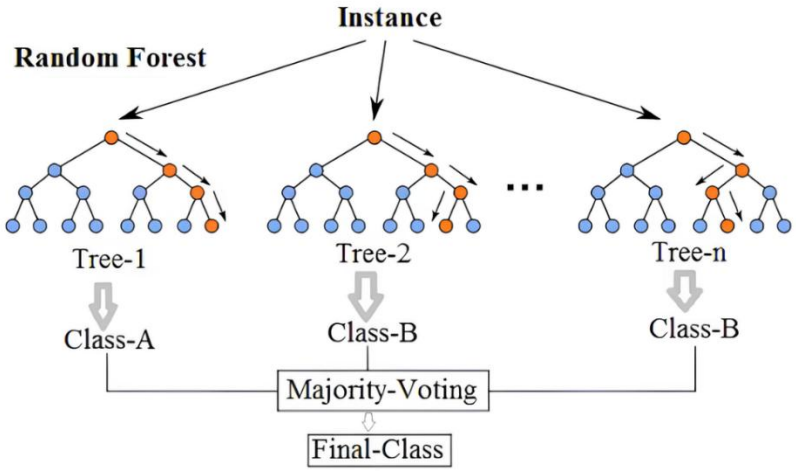
2.2.7 Klasifikasi

Klasifikasi merupakan pengelompokan data ke dalam kategori atau kelas yang telah ditentukan berdasarkan pola atau fitur tertentu dalam pembelajaran mesin

(Maulana *et al*., 2024). Klasifikasi digunakan untuk pendekatan penambangan data yang digunakan untuk memperkirakan keanggotaan kelompok. Klasifikasi termasuk ke dalam *supervised learning*, dimana model dilatih untuk memahami hubungan antara *input* (fitur) dan *output* (label) menggunakan data berlabel (Roihan *et al*., 2020). Setelah pelatihan, model dapat digunakan untuk memprediksi kelas dari data tanpa label. *Support Vector Machines* (SVM), *Decision Trees*, dan *Neural Network* adalah teknik klasifikasi yang umum digunakan. Jenis data yang digunakan dan tujuan aplikasi menentukan keberhasilan teknik ini. Ketidakseimbangan kelas, kebutuhan komputasi, dan relevansi dataset menjadi masalah pada klasifikasi (Hendrawan, 2022).

2.2.8 *Random Forest*

Random Forest adalah algoritma pembelajaran mesin yang menggunakan metode *ensemble* berbasis pohon keputusan (*decision tree*) sebagai *base classifier* yang dibangun dan dikombinasikan, beberapa aspek penting dari metode *Random Forest* diantaranya melakukan sampling terpadu untuk membangun pohon prediksi (Afdhal *et al*., 2022). Setiap pohon keputusan dihasilkan dari sampel data yang diambil secara acak dari kumpulan data pelatihan. Ketika melakukan prediksi, setiap pohon memberikan prediksi dan hasil akhir diambil berdasarkan mayoritas suara dari semua pohon keputusan (Aftari *et al*., 2024). Ilustrasi *Random Forest* dapat dilihat pada Gambar 3.



Gambar 3. Ilustrasi Random Forest (Haidar *et al.*, 2024).

Proses pembelahan simpul diperoleh dengan melakukan suatu perhitungan yang dikenal dengan *Information Gain* dilakukan untuk mencari variabel yang akan dipilih untuk menumbuhkan pohon (Nurfadilla & Faisal, 2022). Untuk menumbuhkan *Information Gain* dilakukan perhitungan *information theory* yaitu *Entropy*. Persamaan *entropy* dapat dilihat pada Persamaan 4.

$$Entropy\ (Y) = \ -\sum_i p\ (c|Y) log_2 p(c|Y) \tag{4}$$

Dimana *Y* merupakan himpunan kasus dan *p(c|Y)* merupakan proporsi nilai *Y* terhadap kelas *c*.
Sedangkan untuk mendapatkan Gainnya menggunakan Persamaan 5.

$$Gain\ (Y, a) = \ Entropy\ (Y) - \sum_{v \in Values(a)} \frac{|Y_v|}{|Y_a|} Entropy\ (Y_v) \tag{5}$$

Dimana nilai *(a)* merupakan semua yang memungkinkan dalam himpunan kasus *a*. *Yv* merupakan subkelas dari *Y* dengan kelas *v* yang berhubungan dengan kelas a. *Ya* merupakan semua nilai yang sesuai dengan nilai *a*.

Algoritma *Random Forest* memiliki beberapa parameter penting yang mempengaruhi performa model. Parameter *Random Forest* dapat dilihat pada Tabel 7.

Tabel 7. Parameter *Random Forest* (Yulianti *et al.*, 2022)

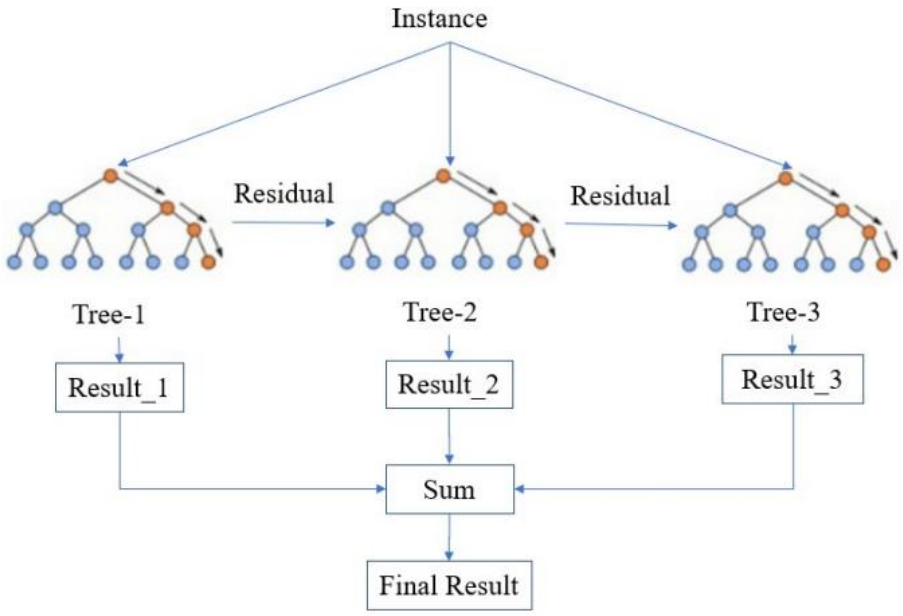
Parameter	Keterangan
n_estimators	Jumlah pohon dalam ensemble.
max_features	Jumlah maksimum fitur yang dipertimbangkan untuk setiap pembagian simpul.
max_depth	Kedalaman maksimum setiap pohon untuk mengontrol kompleksitas model.
min_samples_split	Jumlah minimum sampel yang diperlukan untuk membagi simpul.
min_samples_leaf	Jumlah minimum sampel yang harus ada pada setiap daun pohon.

2.2.9 *XGBoost* (*Extreme Gradient Boosting*)

XGBoost (*Extreme Gradient Boosting*) adalah algoritma pembelajaran mesin berbasis *ensemble* yang dirancang untuk meningkatkan akurasi prediksi melalui teknik *boosting* (Abdurrahman *et al*., 2022). *XGBoost* dikembangkan oleh Tianqi Chen dan pertama kali diperkenalkan pada tahun 2014. Algoritma ini merupakan pengembangan dari metode *Gradient Boosting Machine* (GBM) yang mengoptimalkan proses pelatihan untuk memberikan performa yang lebih baik dengan efisiensi tinggi.

XGBoost menggunakan pendekatan *boosting*, yaitu teknik *ensemble* yang membangun model secara bertahap, di mana model baru dilatih untuk memperbaiki kesalahan dari model sebelumnya. Setiap iterasi dalam *XGBoost* berfokus pada penyesuaian terhadap sisa kesalahan (*residual error*) dari model sebelumnya, dengan tujuan meminimalkan fungsi *loss* secara iteratif. Model akhir merupakan kombinasi dari semua model dalam iterasi, yang digabungkan untuk memberikan prediksi yang lebih akurat.

Fungsi objektif digunakan untuk mengukur seberapa baik model tersebut sesuai dengan data latih. Terdapat 2 bagian penting dalam *objective function* yaitu *training loss* yang digunakan untuk mengukur seberapa prediktif model tersebut sehubungan dengan data latih dan *regularization term* yang digunakan untuk mengontrol kompleksitas model dan membantu untuk menghindari *overfitting* (keadaan ketika model *Machine Learning* terlalu kompleks sehingga tidak dapat mempelajari pola latih yang mengakibatkan tidak dapat menggeneralisasi dengan baik data uji) (Yulianti *et al*., 2022). Ilustrasi *XGBoost* dapat dilihat pada Gambar 4.



Gambar 4. Ilustrasi *XGBoost* (Sinaga & Agustian, 2022).

XGBoost menggunakan perkiraan yang lebih akurat daripada *Gradient Boosting* dengan menggunakan *gradien orde* kedua dan regularisasi tingkat lanjut. Karakteristik yang terpenting dari fungsi objektif terdiri dari 2 bagian yaitu nilai pelatihan yang hilang dan nilai regularisasi seperti pada Persamaan 6.

$$obj(\theta) = L(\theta) + \Omega(\theta) \tag{6}$$

Dimana L adalah fungsi pelatihan yang hilang, dan Ω adalah fungsi regularisasi, dan θ adalah parameter model terkait. Fungsi pelatihan yang hilang secara umum dapat ditulis seperti pada Persamaan 7.

$$L(\theta) = \sum_{i=1}^{n} l\,(y_i, \hat{y}_i) \tag{7}$$

Dimana yi adalah nilai data sebenarnya yang dianggap benar dan $\hat{y}_i$ adalah hasil nilai prediksi dari model, sedangkan n adalah jumlah iterasi nilai dari model. Nilai akurasi *XGBoost* tergantung pada parameter yang digunakan. Parameter yang dapat digunakan untuk mendapatkan hasil akurasi terbaik dapat dilihat pada Tabel 8.

Tabel 8. Parameter *XGBoost* (Yulianti *et al.*, 2022)

Parameter	Keterangan
eta	*Learning rate* pada proses pelatihan
max_depth	Tingkat kedalaman suatu pohon, semakin dalam pohon maka akan semakin kompleks
gamma	Parameter penalty pada *regularization*
min_child_weight	Nilai minimal bobot yang dibutuhkan *child node*
subsample	Jumlah sampel yang digunakan untuk proses pelatihan. Misal 0.5 berarti menggunakan setengah dari data acak dalam membuat tree baru
colsample_bytree	Jumlah sampel kolom untuk membuat *tree* baru

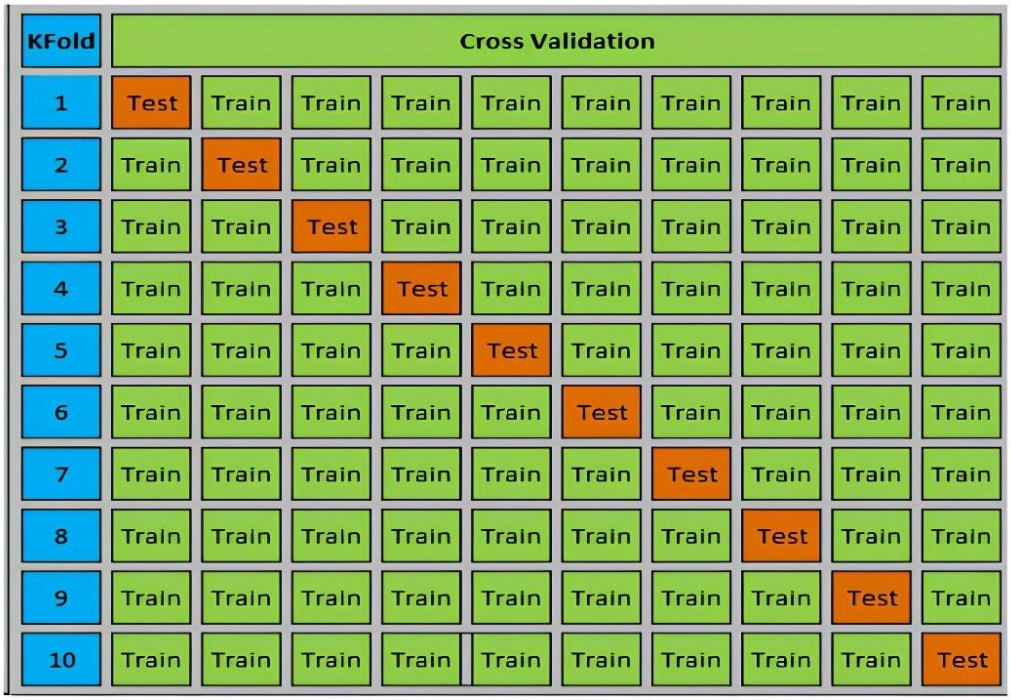
2.2.10 *Split Data*

Split data adalah proses membagi data menjadi beberapa subset atau bagian. Pembagian data adalah salah satu dari beberapa aspek yang mempengaruhi seberapa baik kinerja model klasifikasi pada algoritma (Nugroho, 2022). *Split data* merupakan proses untuk membagi antara data latih, data *validation*, dan data uji. Data latih digunakan untuk membangun model, sementara data uji digunakan untuk mengevaluasi kinerja model (Haidar *et al*., 2024). Data *validation* adalah subset data yang digunakan untuk menyetel parameter model dan mencegah *overfitting* selama pelatihan. *Metode holdout validation* dan *k-fold cross validation* dapat digunakan untuk membagi data latih dan data uji. Proses validasi sangat penting untuk dilakukan, tujuannya agar setiap data memiliki peluang sebagai pelatihan data dan pengujian data.

2.2.11 *K-Fold Cross Validation*

k-fold cross validation adalah teknik yang digunakan untuk model yang melibatkan pemecahan data menjadi beberapa subnet, dan model tersebut akan dilatih dan diuji menggunakan subnet tersebut secara bergantian. *k-fold cross validation* merupakan suatu metode yang biasa digunakan untuk melakukan evaluasi kinerja *classifier*, metode ini dapat digunakan apabila memiliki jumlah data yang sedikit (Lumbanraja *et al*., 2021) . *k-fold cross validation* adalah proses

dimana dataset dipecah menjadi sejumlah *fold* yang digunakan untuk mengevaluasi kemampuan model saat diberikan data baru. Dalam setiap iterasi *k-fold cross-validation*, subdivisi bergantian digunakan sebagai subset pengujian dan pelatihan (Oktafiani *et al*., 2023). Secara umum, pengujian nilai k dilakukan sebanyak 10 kali untuk memperkirakan akurasi estimasi. *k-fold cross validation* digunakan untuk mengukur kinerja model dengan lebih akurat, menghindari *overfitting*, dan memastikan bahwa model memiliki kemampuan generalisasi yang baik (Wijiyanto *et al*., 2024). Skema *k-fold cross validation* dengan nilai *k*=10 dapat dilihat pada Gambar 5.



Gambar 5. Skema *k-fold cross validation* dengan *k*=10 (Wijiyanto *et al*., 2024).

2.2.12 *Oversampling*

Oversampling merupakan salah satu teknik penyeimbangan data (*resampling*) yang dilakukan dengan cara menambahkan jumlah data pada kelas minoritas agar proporsinya menjadi seimbang dengan kelas mayoritas (Diantika, 2023). Teknik ini sangat penting untuk digunakan ketika data yang digunakan tidak seimbang, di mana jumlah sampel dari satu atau lebih kelas jauh lebih sedikit daripada kelas

lainnya. Ketidakseimbangan data dapat menyebabkan model klasifikasi menjadi bias, di mana model lebih cenderung memprediksi kelas mayoritas dan mengabaikan kelas minoritas. Untuk data yang memiliki jumlah instance yang lebih kecil, oversampling cocok digunakan untuk proses penyeimbangan data (Ery, 2024). Beberapa teknik yang sering digunakan adalah *Synthetic Minority Over-sampling Technique* (SMOTE), *Random Oversampling* (ROS), dan *Adaptive Synthetic Sampling* (ADASYN).

2.2.13 *Hyperparameter Tuning*

Hyperparameter tuning adalah proses untuk menentukan kombinasi nilai *hyperparameter* yang paling sesuai agar model dapat memberikan hasil yang optimal (Matin, 2023). Berbeda dengan parameter model yang diperoleh dari proses pelatihan, *hyperparameter* ditetapkan terlebih dahulu dan berfungsi mengontrol perilaku pembelajaran model. Proses *tuning* dilakukan dengan mengevaluasi beberapa konfigurasi yang telah ditentukan hingga ditemukan kombinasi terbaik yang meningkatkan performa model.

Salah satu teknik yang umum digunakan untuk melakukan *tuning* adalah *Grid Search Cross Validation* (*GridSearchCV*). *Grid Search Cross Validation* adalah istilah yang digunakan untuk merujuk teknik *Grid Search* dan *Cross-Validation* yaitu metode pemilihan kombinasi model dan *hyperparameter* (Nugraha & Sasongko, 2022). Metode ini secara sistematis mencoba seluruh kombinasi nilai *hyperparameter* dari suatu ruang pencarian (*parameter grid*) yang telah ditetapkan sebelumnya. Untuk setiap kombinasi, model akan dilatih dan dievaluasi menggunakan teknik *cross-validation*, sehingga performa setiap konfigurasi dapat dinilai secara menyeluruh.

2.2.14 *Confusion Matrix* (Matriks Konfusi)

Confusion matrix adalah pengukuran performa untuk masalah klasifikasi *Machine Learning* dimana keluaran dapat berupa dua kelas atau lebih. *Confusion Matrix*

adalah metode yang digunakan dalam mengukur kinerja dari suatu algoritma klasifikasi yang berbentuk tabel dengan 4 nilai yang merepresentasikan hasil dari klasifikasi (Meilawati & Winiarti, 2022). *Confusion matrix* berupa tabel yang berisi informasi nilai aktual dan prediksi yang dibuat oleh sistem klasifikasi (Suci *et al*., 2022). Ada empat istilah yang merupakan representasi hasil proses klasifikasi pada confusion matrix yaitu *True Positive, True Negative, False Positive*, dan *False Negative*. *Confusion Mat*rix untuk dapat dilihat pada Tabel 9.

Tabel 9. Tabel *Confusion Matrix* (Suci *et al*., 2022)

		Prediksi	
		Positif	*Negatif*
Aktual	Positif	*TP*	FN
	Negatif	FP	TN

Keterangan:

True Positive (TP) = Data positif yang diklasifikasikan dengan benar

False Positive (FP) = Data negatif namun diklasifikasikan sebagai data positif

True Negative (TN) = Data negatif yang diklasifikasikan dengan benar.

False Negative (FN) = Data positif namun diklasifikasikan sebagai data negatif

Untuk mengukur performa *confusion matrix* dapat dilakukan dengan menghitung Akurasi*,* Presisi*, Recall,* dan *F1-score.*

a. Akurasi

Akurasi (*accuracy)* adalah metrik evaluasi yang mengukur seberapa baik model membuat prediksi yang benar dari total prediksi yang dilakukan. Dalam konteks klasifikasi, Akurasi memberikan gambaran mengenai seberapa sering model memprediksi kelas yang benar, baik itu kelas positif maupun negatif. Akurasi merupakan hasil perhitungan ketepatan suatu model dalam mengklasifikasikan data untuk diprediksi dengan benar. Akurasi juga dapat menggambarkan kedekatan nilai prediksi dengan nilai aktual, sehingga menjadi indikator yang penting dalam menilai performa keseluruhan model

klasifikasi (Suci *et al*., 2022). Persamaan Akurasi dapat dilihat pada Persamaan 8 (Arifuddin *et al*., 2024).

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (8)$$

b. Presisi

Presisi (*precision*) adalah metrik evaluasi yang mengukur seberapa baik model membuat prediksi yang benar untuk kelas positif dari total prediksi positif yang dilakukan. Presisi membantu menghitung seberapa sering model memprediksi kelas positif dengan benar, di antara semua prediksi positif yang dibuat oleh model. Persamaan Presisi dapat dilihat pada Persamaan 9 (Arifuddin *et al*., 2024).

$$Precision = \frac{TP}{TP+FP} \quad (9)$$

c. *Recall*

Recall adalah metrik evaluasi yang menggambarkan seberapa baik suatu model dalam mengidentifikasi kelas positif dengan benar. Persamaan *Recall* dapat dilihat pada Persamaan 10 (Arifuddin *et al*., 2024).

$$Recall = \frac{TP}{TP+FN} \quad (10)$$

d. F1-Score

F1-Score merupakan metrik evaluasi yang mencerminkan keseimbangan antara Presisi dan *Recall*. Nilai *F1-Score* akan memberikan informasi tentang seberapa baik model kita dalam menggabungkan kemampuan Presisi dan *Recall*, sehingga kita bisa memahami seberapa efektif model kita dalam melakukan klasifikasi. Persamaan *F1-Score* dapat dilihat pada Persamaan 11 (Astuti *et al*., 2024).

$$F1\text{-}Score = \frac{2 \times Recall \times Precision}{Recall + Precision} \quad (11)$$

III. METODOLOGI PENELITIAN

3.1 Tempat dan Waktu Penelitian

3.1.1 Tempat

Penelitian ini dilakukan di Jurusan Ilmu Komputer, Fakultas Matematika dan Ilmu Pengetahuan Alam yang beralamat di Jl. Soemantri Brojonegoro No. 1 Gedung Meneng, Bandar Lampung.

3.1.2 Waktu

Penelitian ini dilakukan mulai dari bulan November 2024 hingga bulan Juni 2025. Rincian waktu penelitian dapat dilihat pada Tabel 10.

Tabel 10. Tabel Waktu Penelitian

Kegiatan	2024																2025															
	Nov				Desember				Januari				Februari				Maret				April				Mei				Juni				Juli		
	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3
Studi Literatur																																
Pengumpulan Data																																
Penyusunan Bab I-III																																
Preprocessing Data																																
Pelabelan Data dan Ekstraksi Fitur																																
Pembagian Data																																
Pelatihan Model																																
Pengujian Model																																
Perbandingan Hasil Klasifikasi																																
Penyusunan Bab IV-V																																

3.2 Data dan Alat

Adapun data dan alat yang digunakan selama proses penelitian yaitu sebagai berikut.

3.2.1 Data

Dataset ini berisi 724 komentar yang diambil dari unggahan akun Instagram resmi Kementerian Komunikasi dan Digital (Kemkomdigi) yang terkait dengan permasalahan judi online. Data diperoleh melalui proses *scraping* menggunakan *InsC Instagram Comments Picker & Exporter* dengan cara memasukan id postingan. Dataset ini mencakup delapan atribut utama, yaitu *id, username, owner_id, profile_pic_url, text ,createdat, created_at_formated ,dan profile url.* Contoh data penelitian dapat dilihat pada Tabel 11.

Tabel 11. Contoh *Dataset*

id	username	owner id	profile_pic_url	text	created_at	created_at_formated	profile_url
1.81E+16	maimunzakariamahmud	8.56E+09	https://bit.ly/3BBSLKr	Judi online dan KORUPSI adalah perusak bangsa,TOLONG JANGAN KORUPSI	1730387107	10/31/2024, 10:05:07 PM	https://www.instagram.com/maimunzakariamahmud
1.80E+16	hariqosatria	2.16E+09	https://bit.ly/3VLSlYG	Terima kasih untuk seluruh masyarakat yang terus menyampaikan bahaya judi online, pinjol nakal, dll. Keren Kemkomdigi 👍ID	1730387124	10/31/2024, 10:05:24 PM	https://www.instagram.com/hariqosatria

3.2.2 Alat

Berikut alat yang digunakan dalam mendukung penelitian ini:

1. Perangkat Keras (*Hardware)*

   Perangkat keras yang digunakan berupa laptop dengan spesifikasi:

   1) Merk : Asus
   2) Model : LAPTOP-U9I45A9Q
   3) *CPU* : Intel(R) Core i5-1035G1 CPU 1.19 GHz
   4) *RAM* : 8GB

2. Perangkat Lunak (*Software)*

   Perangkat lunak yang digunakan dalam penelitian ini yaitu:

   1) *Google Colab*

      Google Colab, layanan komputasi *cloud* gratis yang disediakan oleh Google, memungkinkan pengguna mengakses sumber daya komputasi yang kuat dan menggunakan lingkungan pengembangan yang nyaman dan terintegrasi dengan alat populer seperti *Jupyter Notebook*. Selain itu, pengguna dapat melakukan pemrosesan data dan pembelajaran mesin, serta menulis, menjalankan, dan berbagi kode Python.

   2) *Python* 3.10.12

      Python adalah bahasa pemrograman komputer yang populer untuk membangun situs web, aplikasi, mengotomatiskan tugas, dan melakukan analisis data. Artinya, ia dapat digunakan untuk membuat berbagai program, bukan hanya untuk menyelesaikan masalah tertentu (Python, 2025)

   3) *Library Pandas* 2.2.2

      Pandas adalah library *Python* yang paling dikenal dan banyak digunakan. Paket ini bisa digunakan untuk menganalisis data dengan cepat, realistis, dan serbaguna. Pandas dapat memakainya untuk mengombinasikan, mengelompokkan, dan mengklasifikasikan data yang berasal dari berbagai

sumber, seperti Excel, SQL databases, CSV, dan sebagainya (NumFOCUS, 2025).

4) *Library Scikit-learn* 1.6.0

Scikit-learn adalah library untuk *machine learning* yang mendukung *supervised* dan *unsupervised learning*. *Scikit-learn* juga menyediakan berbagai alat untuk penyesuaian model, prapemprosesan data, pemilihan model, evaluasi model, dan banyak utilitas lainnya (Scikit-learn, 2025).

5) *Library XGBoost* 2.1.3

XGBoost (*Extreme Gradient Boosting*) adalah library yang digunakan untuk melakukan *boosting* dengan *gradient descent*. Ini adalah teknik *ensemble* yang digunakan untuk meningkatkan akurasi model, terutama dalam tugas klasifikasi dan regresi. *XGBoost* dikenal dengan kecepatan dan kinerjanya yang sangat baik dalam kompetisi *machine learning* (*XGBoost Doc.*, 2022).

6) *Library Imbalanced-learn* (*imblearn*) 0.13.0

Imbalanced-learn adalah *library* yang dibangun di atas *scikit-learn*, digunakan untuk menangani masalah dataset yang tidak seimbang (*imbalanced dataset*). *Library* ini menyediakan berbagai teknik *oversampling* dan *undersampling*, termasuk SMOTE (*Synthetic Minority Over-sampling Technique*) untuk meningkatkan performa model pada data yang tidak seimbang (*Imbalanced-Learn Documentation*, 2024).

7) *Library Nltk* 3.9.1

NLTK adalah *library* untuk pemrosesan bahasa alami (NLP) di *Python*. *Library* ini menyediakan berbagai alat untuk tokenisasi, *stemming*, *lemmatization*, *parsing*, dan analisis sentimen. NLTK dapat digunakan untuk memproses teks dan menganalisis emosi atau opini dalam teks (*Natural Language Toolkit*, 2024).

8) *Library WordCloud*

WordCloud adalah *library* visualisasi dalam *Python* yang digunakan untuk menampilkan kumpulan kata dalam bentuk awan kata. Ukuran setiap kata dalam visualisasi mencerminkan tingkat kepentingan atau frekuensinya dalam data. *Library* ini banyak digunakan dalam analisis teks untuk membantu mengidentifikasi kata-kata kunci secara visual (Pradana, 2020).

9) *Draw.io*

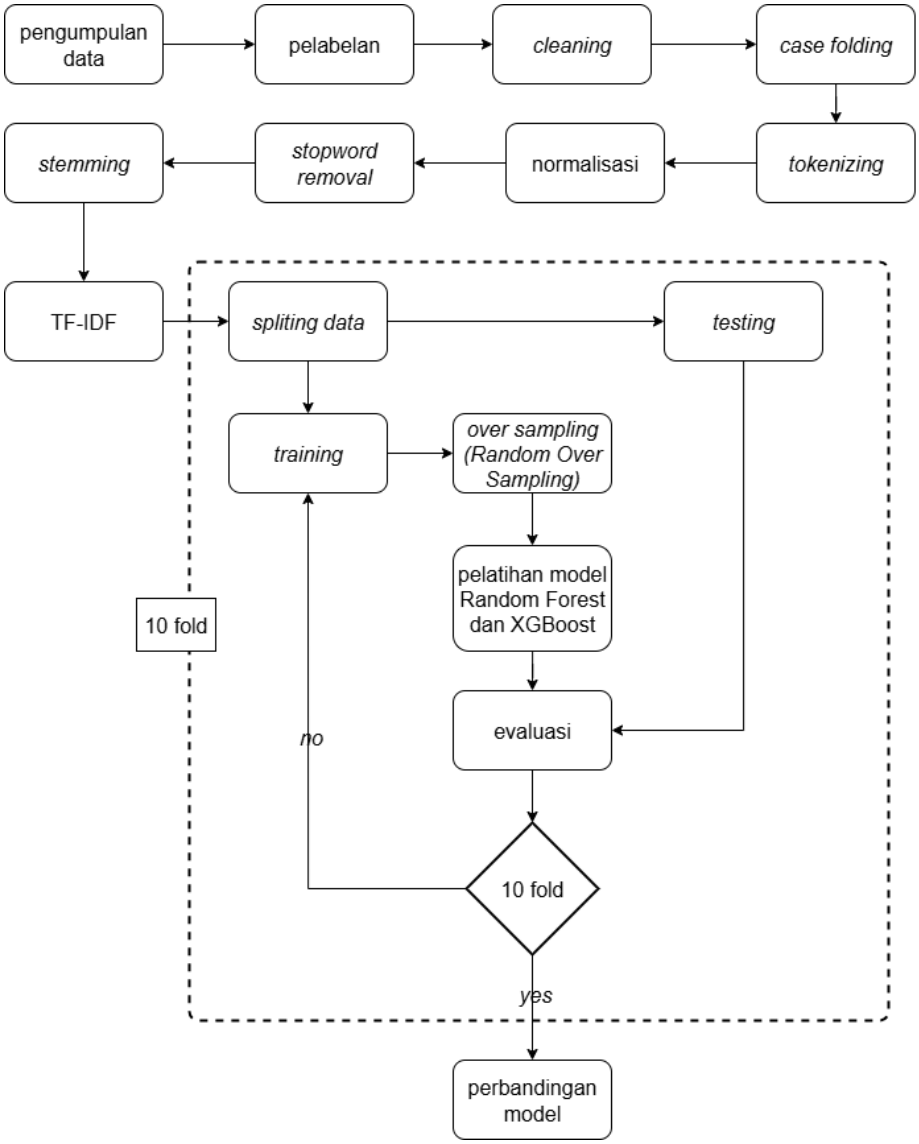
Draw.io adalah aplikasi berbasis web yang digunakan untuk membuat diagram, *flowchart*, diagram alur, peta pikiran, dan banyak lagi. Berbagai alat dan template disediakan oleh aplikasi ini untuk memvisualisasikan konsep dan proses secara jelas dan mudah (Draw.io, 2023).

10) *InsC Instagram Comments Picker & Exporter*

InsC adalah ekstensi *Chrome* yang dirancang untuk mendukung pengumpulan dan pengelolaan data komentar dari platform Instagram. Dengan menggunakan ekstensi ini, pengguna dapat mengambil komentar dari unggahan Instagram secara efisien dan mengekspornya ke format CSV atau *Excel* untuk keperluan analisis lebih lanjut. *InsC* juga menyediakan fitur pemilihan komentar secara acak, yang ideal untuk digunakan dalam aktivitas seperti pengundian hadiah (*giveaway*) (*Instagram Giveaway Picker & IG Comments to Excel Exporter*, 2021).

3.3 Metodologi Penelitian

Tahapan penelitian dapat dilihat pada Gambar 6.



Gambar 6. Tahap Penelitian.Klasifikasi Sentimen

1. Pengumpulan Data

Tahap pertama dalam penelitian ini adalah pengumpulan data. Proses pengumpulan data mengambil dari media sosial Instagram pada komentar foto dengan menggunakan *InsC Instagram Comments Picker & Exporter*. *InsC Instagram Comments Picker & Exporter* adalah alat scraping untuk mendapatkan informasi yang sedang dicari pada suatu postingan instagram. Pengambilan data komentar mempertimbangkan dengan isi dan maksud gambar pada postingan agar relevan dengan penelitian. Komentar diambil pada akun

instagram resmi kemkomdigi yaitu @kemkomdigi yang berkaitan dengan permasalahan judi online. Komentar diambil dari 12 unggahan yang dipublikasikan antara 21 Oktober hingga Desember 2024 yang terdiri delapan atribut yaitu *id*, *username*, *text*, *created_at*, dan *profile_url*.

2. Pelabelan Data

Proses pelabelan data dilakukan untuk menentukan kategori sentimen dari setiap data teks yang digunakan dalam penelitian ini. Pelabelan dilakukan secara manual oleh tiga orang annotator. Setiap annotator memberikan label sentimen secara independen terhadap masing-masing data. Label yang digunakan terdiri dari kategori positif dan negatif. Untuk memastikan konsistensi dan mengurangi subjektivitas, hasil pelabelan akhir ditentukan melalui mekanisme voting. Jika dua dari tiga annotator memberikan label yang sama, maka label tersebut ditetapkan sebagai label akhir data tersebut. Dengan demikian, keputusan mayoritas digunakan untuk meningkatkan validitas hasil pelabelan. Kelas positif diberi label 0 dan kelas negatif diberi label 1.

3. *Preprocessing*

Untuk membuat data lebih mudah diolah oleh model, maka perlu dilakukan *preprocessing* terlebih dahulu. Dalam tahap *preprocessing*, dilakukan penyaringan, penghilangan, dan perbaikan kata pada data ulasan melalui beberapa proses seperti *case folding*, *cleaning* data, pengubahan *slang word*, penghapusan *stop word*, dan *stemming*.

1. *Cleaning* data adalah proses penghapusan bagian tulisan yang maknanya sulit dikuantifikasi untuk pembuatan model seperti angka, emotikon, dan simbol-simbol.
2. *Case folding* merupakan tahap mengubah semua huruf yang terdapat pada dokumen dari huruf kapital menjadi huruf kecil.
3. *Tokenization* adalah proses memecah teks menjadi daftar token-token. Token mencakup satu atau lebih kata dan digunakan untuk menyimpan makna yang ada pada kata dasar yang memiliki lebih dari satu kata.

4. Normalisasi adalah proses mengubah kata-kata yang tidak baku atau *slang* dalam teks menjadi yang baku sesuai dengan kaidah bahasa Indonesia. Normalisasi dilakukan dengan pendekatan otomatis berbasis kamus, yang berisi daftar pasangan kata yang tidak baku dan padanannya yang baku.
5. *Stopword removal* merupakan tahap yang dilakukan untuk menghilangkan kata-kata yang tidak deskriptif. Misalnya kata-kata seperti "dia", "adalah", "semua", "diriku" yang merupakan kata-kata diskrit yang digunakan untuk melancarkan tata bahasa kalimat namun tidak memberi arti signifikan pada kalimat itu sendiri.
6. *Stemming* adalah proses mengubah kata berimbuhan menjadi kata dasar.

4. Pembobotan Kata

Data yang sudah melalui proses *text preprocessing* selanjutnya dihitung berapa banyak frekuensi kemunculan setiap kata dalam dokumen. *TF-IDF* adalah salah satu metode pembobotan sebuah kata didalam sistem pencarian informasi. Hasil akhirnya adalah model pembobotan yang memberikan nilai spesifik untuk setiap token berdasarkan pola penggunaanya dalam data pelatihan.

5. *Splitting data*

Pada langkah ini, data dipisahkan menjadi data latih dan data uji dengan rasio 70% data latih dan 30% data uji. Rasio tersebut dipilih karena keterbatasan jumlah data. Selanjutnya, berdasarkan data latih yang sudah dibentuk melalui proses pemisahan awal dengan data uji, maka data latih tersebut akan dibagi lagi menggunakan *k-fold cross validation* dengan jumlah *k*. Melalui proses *k-fold cross validation* ini akan dihasilkan data latih dan data validasi.

6. Penyeimbangan Data

Penyeimbangan data dilakukan untuk mengatasi masalah ketidakseimbangan jumlah data antar kelas sentimen. Data yang tidak seimbang dapat menyebabkan model cenderung bias terhadap kelas yang dominan dan mengabaikan kelas yang minoritas. Untuk mengatasi hal ini, digunakan teknik *Random Over Sampling*,

yaitu metode yang memperbanyak data pada kelas minoritas dengan cara menduplikasi sampel secara acak hingga jumlahnya seimbang dengan kelas mayoritas. Penyeimbangan data hanya dilakukan pada data latih.

7. Pelatihan Model *Random Forest* dan *XGBoost*

Model dilatih menggunakan dua pendekatan, yaitu dengan *parameter default* dan *parameter hasil tuning*. Proses *tuning* dilakukan untuk meningkatkan performa model melalui pencarian kombinasi parameter terbaik. Pelatihan dilakukan menggunakan *k-fold cross validation* untuk memastikan hasil yang stabil dan mengurangi *overfitting*. Untuk pencarian *hyperparameter* terbaik, digunakan metode *GridSearchCV* yang secara otomatis menguji kombinasi parameter berdasarkan hasil validasi silang.

8. Pengujian Model

Pada tahap pengujian ini, model *Random Forest* dan *XGBoost* yang telah berhasil dibuat dengan data latih selanjutnya diuji menggunakan data uji. Tujuannya adalah untuk mengetahui kemampuan model dalam mengklasifikasikan data baru yang belum pernah dilihat sebelumnya. Pengujian dilakukan pada model dengan parameter default dan model hasil *tuning*, sehingga hasilnya dapat dibandingkan untuk melihat pengaruh *tuning* terhadap kinerja model di dunia nyata.

9. Evaluasi Model

Evaluasi model dilakukan untuk mengukur kinerja hasil klasifikasi sentimen menggunakan beberapa metrik, yaitu *Accuracy*, *Precision*, *Recall*, dan *f1-score*. Keempat metrik ini digunakan agar evaluasi tidak hanya berfokus pada ketepatan keseluruhan, tetapi juga pada kemampuan model dalam mengenali kelas positif dan negatif secara seimbang.

10. Perbandingan Model

Tahapan terakhir adalah melakukan perbandingan hasil klasifikasi dari 2 model yang telah digunakan yaitu *Random Forest* dan *XGBoost*. Perbandingan dilakukan untuk menilai model mana yang memberikan hasil paling optimal dalam mengklasifikasikan data sentimen. Model yang dibandingkan terdiri dari versi parameter default dan parameter hasil *tuning*, untuk melihat pengaruh *hyperparameter* terhadap kinerja masing-masing model.

V. KESIMPULAN DAN SARAN

5.1 Kesimpulan

Berdasarkan penelitian yang telah dilakukan, dapat disimpulkan bahwa:

1. Metode *Random Forest* dan *XGBoost* berhasil diimplementasikan untuk mengklasifikasikan sentimen masyarakat terhadap kinerja Kemkomdigi dalam menangani isu judi online. Proses klasifikasi dilakukan melalui tahapan pelabelan sentimen, preprocessing teks, penyeimbangan data, pelatihan model, dan evaluasi menggunakan metrik Akurasi, Presisi, *Recall*, dan F1-*Score*.

2. Hasil perbandingan menunjukkan bahwa *Random Forest* menunjukkan performa yang lebih unggul dibandingkan *XGBoost* dalam klasifikasi sentimen setelah dilakukan *hyperparameter tuning*. *Random Forest* menghasilkan niai *F1-score* tertinggi yaitu 70,90%, sedangkan XGBoost memiliki nilai 67,85%. Selain itu, *Random Forest* juga menghasilkan jumlah klasifikasi benar lebih banyak, yaitu 144 data, dibandingkan *XGBoost* yang hanya mencapai 136 data. Hasil ini menunjukkan bahwa *Random Forest* lebih efektif dan akurat dalam mengklasifikasikan data sentimen setelah melalui proses penyetelan parameter.

Dengan demikian, dapat disimpulkan bahwa *Random Forest* lebih unggul dalam klasifikasi sentimen masyarakat terhadap isu yang diangkat, khususnya setelah dilakukan optimasi parameter, meskipun masing-masing model memiliki kelebihan pada aspek tertentu.

5.2 Saran

Adapun saran dari penelitian yang telah dilakukan yaitu:

1. Untuk penelitian selanjutnya disarankan untuk menambah jumlah dan variasi data, misalnya dengan menggali data dari platform media sosial lainnya atau menggunakan data dalam rentang waktu yang lebih panjang.
2. Untuk penelitian selanjutnya, disarankan menggunakan metode pelabelan otomatis yang mendukung bahasa Indonesia, seperti IndoBERT atau lexicon-based lokal, agar hasil klasifikasi lebih akurat terhadap konteks bahasa pengguna.
3. Disarankan untuk mengeksplorasi lebih banyak kombinasi hyperparameter atau menggunakan teknik optimasi lain seperti *RandomizedSearchCV* atau *Bayesian Optimization*, agar diperoleh parameter yang benar-benar optimal.
4. Model yang telah dilatih dan dievalasi dalam penelitian ini dapat diintegrasikan ke dalam sistem monitoring opini publik atau dashboard analitik untuk membantu pengambilan keputusan dalam kebijakan publik terhadap isu-isu penting seperti judi online.

DAFTAR PUSTAKA

Abdurrahman, G., Oktavianto, H., & Sintawati, M. (2022). Optimasi Algoritma XGBoost Classifier Menggunakan Hyperparameter Gridesearch dan Random Search Pada Klasifikasi Penyakit Diabetes. *INFORMAL: Informatics Journal*, *7*(3), 193–198. https://doi.org/10.19184/isj.v7i3.35441

Afdhal, I., Kurniawan, R., Iskandar, I., Salambue, R., Budianita, E., & Syafria, F. (2022). Penerapan Algoritma Random Forest Untuk Analisis Sentimen Komentar Di YouTube Tentang Islamofobia. *Jurnal Nasional Komputasi Dan Teknologi Informasi*, *5*(1), 122–130. http://ojs.serambimekkah.ac.id/jnkti/article/view/4004/pdf

Aftari, D. P., Safaat, N. H., Agustian, S., & Afrianty, I. (2024). Perbandingan Performa Klasifikasi Terjemahan Al-Qur'an Menggunakan Metode Random Forest dan Long Short Term Memory. *Journal of Computer System and Informatics (JoSYC)*, *5*(3), 567–577. https://doi.org/10.47065/josyc.v5i3.5156

Anggraini, N., Putra, S. J., Wardhani, L. K., Arif, F. D. U., Hakiem, N., & Shofi, I. M. (2024). A Comparative Analysis of Random Forest, XGBoost, and LightGBM Algorithms for Emotion Classification in Reddit Comments. *Jurnal Teknik Informatika*, *17*(1), 88–97. https://doi.org/10.15408/jti.v17i1.38651

Antonius, R., Zulkarnain, A. R., & Irsyad, H. (2024). Pendekatan TF-IDF, SMOTE, dan SVM dalam Klasifikasi Sentimen Masyarakat terhadap Pemblokiran Judi Online. *Buletin Ilmiah Informatika Teknologi*, *2*(3), 115–122. https://doi.org/10.58369/biit.v2i3.65

Arifuddin, A., Buana, G. S., Vinarti, R. A., & Djunaidy, A. (2024). Performance Comparison of Decision Tree and Support Vector Machine Algorithms for Heart Failure Prediction. *Procedia Computer Science*, *234*, 628–636.

https://doi.org/10.1016/j.procs.2024.03.048

Asshiddiqi, M. F., & Lhaksmana, K. M. (2020). Perbandingan Metode Decision Tree dan Support Vector Machine untuk Analisis Sentimen pada Instagram Mengenai Kinerja PSSI. *E-Proceeding of Engineering*, *22*(10), 9936–9948.

Astuti, K. C., Firmansyah, A., & Riyadi, A. (2024). Implementasi Text Mining Untuk Analisis Sentimen Masyarakat Terhadap Ulasan Aplikasi Digital Korlantas Polri pada Google Play Store. *REMIK: Riset Dan E-Jurnal Manajemen Informatika Komputer*, *8*(1), 383–394.

Bestari, N. P. (2024). *11 Orang Tersangka Judi Online Ditangkap, Termasuk Pegawai Komdigi*. CNBC INDONESIA. https://www.cnbcindonesia.com/tech/20241101120021-37-584821/11-orang-tersangka-judi-online-ditangkap-termasuk-pegawai-komdigi

Diantika, S. (2023). Penerapan Teknik Random Oversampling Untuk Mengatasi Imbalance Class Dalam Klasifikasi Website Phishing Menggunakan Algoritma Lightgbm. *JATI (Jurnal Mahasiswa Teknik Informatika)*, *7*(1), 19–25. https://doi.org/10.36040/jati.v7i1.6006

Draw.io. (2023). *About draw.io*. Draw.Io. https://www.drawio.com/about

Ery, Y. (2024). *Oversampling vs Undersampling dalam Mengatasi Ketidakseimbangan Data*. School of Information Systems. https://sis.binus.ac.id/2024/11/01/oversampling-vs-undersampling-dalam-mengatasi-ketidakseimbangan-data/

Haidar, D., Irawan, B., & Bahtiar, A. (2024). Penerapan Deep Learning Model Random Forest Untuk Prediksi Penerima Bantuan Program Keluarga Harapan (Pkh). *JATI (Jurnal Mahasiswa Teknik Informatika)*, *7*(6), 3564–3571. https://doi.org/10.36040/jati.v7i6.8250

Hassani, H., Beneki, C., Unger, S., Mazinani, M. T., & Yeganegi, M. R. (2020). Text mining in big data analytics. *Big Data and Cognitive Computing*, *4*(1), 1–34. https://doi.org/10.3390/bdcc4010001

Hendrawan, I. R. (2022). Perbandingan Algoritma Naïve Bayes, Svm Dan Xgboost

Dalam Klasifikasi Teks Sentimen Masyarakat Terhadap Produk Lokal Di Indonesia. *Transformasi*, *18*(1), 1–8. https://doi.org/10.56357/jt.v18i1.295

Imbalanced-learn Documentation. (2024). The Imbalanced-Learn Developers. https://imbalanced-learn.org/stable/

Instagram giveaway picker & IG Comments to Excel Exporter. (2021). Chrome Web Store. https://chromewebstore.google.com/detail/insc-instagram-comment-pi/hdfhpnjnlgekgjmniifdieiflhfdkmlk?hl=id

Jadidah, I. T., Milyarta Lestari, U., Alea Amanah Fatiha, K., Riyani, R., & Ariesty Wulandari, C. (2023). Analisis Maraknya Judi Online di Masyarakat. *JISBI: Jurnal Ilmu Sosial Dan Budaya Indonesia*, *1*(1), 20–27.

Jilcha, L. A., & Kwak, J. (2022). Machine learning-based advertisement banner identification technique for effective piracy website detection process. *Computers, Materials and Continua*, *71*(2), 2883–2899. https://doi.org/10.32604/cmc.2022.023167

Julianti, O. N., Suarna, N., & Prihartono, W. (2024). Penerapan Natural Language Processing Pada Analisis Sentimen Judi Online Di Media Sosial Twitter. *JATI (Jurnal Mahasiswa Teknik Informatika)*, *8*(3), 2936–2941. https://doi.org/10.36040/jati.v8i3.9613

Lumbanraja, F. R., Saputra, R. A., Muludi, K., Hijriani, A., & Junaidi, A. (2021). Implementasi Support Vector Machine Dalam Memprediksi Harga Rumah Pada Perumahan Di Kota Bandar Lampung. *Jurnal Pepadun*, *2*(3), 327–335. https://doi.org/10.23960/pepadun.v2i3.90

Luqyana, W. A., Cholissodin, I., & Perdana, R. S. (2018). Analisis Sentimen Cyberbullying pada Komentar Instagram dengan Metode Klasifikasi Support Vector Machine. *Jurnal Pengembangan Teknologi Informasi Dan Ilmu Komputer*, *2*(11), 4704–4713. http://j-ptiik.ub.ac.id

Matin, I. M. M. (2023). Hyperparameter Tuning Menggunakan GridsearchCV pada Random Forest untuk Deteksi Malware. *Multinetics*, *9*(1), 43–50. https://doi.org/10.32722/multinetics.v9i1.5578

Maulana, A., Yuliana, A., Bandung, T., Politeknik, J., Pesantren, J., & Cimahi, K. (2024). *Analisis Sentimen Opini Publik Terkait Judi Online Pada Pengguna Aplikasi X Menggunakan Algoritma Naïve Bayes dan Support Vector Machine*. *12*(3), 3706–3714.

Meilawati, N., & Winiarti, S. (2022). Analisis Sentimen Masyarakat Terhadap Kebijakan Kominfo Tentang Penyelenggara Sistem Elektronik (PSE) Menggunakan Metode Naive Bayes. *Jurnal Sarjana Teknik Informatika*, *10*(3), 172–181.

Natural Language Toolkit. (2024). NLTK Project. https://www.nltk.org/

Nugraha, W., & Sasongko, A. (2022). Hyperparameter Tuning on Classification Algorithm with Grid Search. *Sistemasi*, *11*(2), 391–401. https://doi.org/10.32520/stmsi.v11i2.1750

Nugroho, A. (2022). Analisa Splitting Criteria Pada Decision Tree dan Random Forest untuk Klasifikasi Evaluasi Kendaraan. *JSITIK: Jurnal Sistem Informasi Dan Teknologi Informasi Komputer*, *1*(1), 41–49. https://doi.org/10.53624/jsitik.v1i1.154

NumFOCUS. (2025). *About pandas*. OVHcloud. https://pandas.pydata.org/about/

Nurdin, A., Anggo Seno Aji, B., Bustamin, A., & Abidin, Z. (2020). Perbandingan Kinerja Word Embedding Word2Vec, Glove, Dan Fasttext Pada Klasifikasi Teks. *Jurnal Tekno Kompak*, *14*(2), 74–79. https://doi.org/10.33365/jtk.v14i2.732

Nurfadilla, Z., & Faisal. (2022). Implementasi Data Mining Untuk Memprediksi Kelulusan Mahasiswa Tepat Waktu Menggunakan Random Forest. *Journal of Artificial Intelligence & Data Science*, *01*(1), 127–135.

Oktafiani, R., Hermawan, A., & Avianto, D. (2023). Pengaruh Komposisi Split data Terhadap Performa Klasifikasi Penyakit Kanker Payudara Menggunakan Algoritma Machine Learning. *Jurnal Sains Dan Informatika*, *August*, 19–28. https://doi.org/10.34128/jsi.v9i1.622

Pradana, M. G. (2020). Penggunaan Fitur Wordcloud Dan Document Term Matrix

Dalam Text Mining. *Jurnal Ilmiah Infromatika (JIF)*, *08*(01), 38–43.

Purnasiwi, R. G., Kusrini, & Hanafi, M. (2023). Analisis Sentimen Pada Review Produk Skincare Menggunakan Word Embedding dan Metode Long Short-Term Memory (LSTM). *Innovative: Journal Of Social Science Research*, *3*(2), 11433–11448.

Pusat Pelaporan dan Analisis Transaksi Keuangan. (2024). *Gawat! Jumlah Fantastis Usia Anak Main Judi Online*. PPATK. https://www.ppatk.go.id/news/read/1373/gawat-jumlah-fantastis-usia-anak-main-judi-online.html

Python, F. S. (2025). *Python is powerful... and fast; plays well with others; runs everywhere; is friendly & easy to learn; is Open.* Python. https://www.python.org/about/

Retnoningsih, E., & Pramudita, R. (2020). Mengenal Machine Learning Dengan Teknik Supervised Dan Unsupervised Learning Menggunakan Python. *Bina Insani Ict Journal*, *7*(2), 156–165. https://doi.org/10.51211/biict.v7i2.1422

Roihan, A., Sunarya, P. A., & Rafika, A. S. (2020). Pemanfaatan Machine Learning dalam Berbagai Bidang: Review paper. *IJCIT (Indonesian Journal on Computer and Information Technology)*, *5*(1), 75–82. https://doi.org/10.31294/ijcit.v5i1.7951

Salim, E., & Syafrullah, M. (2023). *Jakarta Barat Menggunakan Algoritme K-Nearest Neighbor*. *20*(1), 58–65. https://kemsalim.space/ulasan_dukcapil/

Scikit-learn. (2025). *scikit-learn Machine Learning in Python*. Scikit-Learn Developers. https://scikit-learn.org/stable/

Sinaga, H. H., & Agustian, S. (2022). Pebandingan Metode Decision Tree dan XGBoost untuk Klasifikasi Sentimen Vaksin Covid-19 di Twitter. *Jurnal Nasional Teknologi Dan Sistem Informasi*, *8*(3), 107–114. https://doi.org/10.25077/teknosi.v8i3.2022.107-114

Suci, A., Nusrang, M., & Aswi, A. (2022). Penerapan Metode Random Forest Untuk Klasifikasi Varian Minuman Kopi di Kedai Kopi Konijiwa Bantaeng.

VARIANSI: Journal of Statistics and Its Application on Teaching and Research, *4*(3), 121–127. https://doi.org/10.35580/variansiunm31

Wijiyanto, W., Pradana, A. I., Sopingi, S., & Atina, V. (2024). Teknik K-Fold Cross Validation untuk Mengevaluasi Kinerja Mahasiswa. *Jurnal Algoritma*, *21*(1), 239–248. https://doi.org/10.33364/algoritma/v.21-1.1618

XGBoost Documentation. (2022). XGBoost Documentation. https://xgboost.readthedocs.io/en/stable/

Yulianti, S. E. H., Oni, S., & Yuana, S. (2022). Penerapan Metode Extreme Gradient Boosting (XGBOOST) pada Klasifikasi Nasabah Kartu Kredit. *Journal of Mathematics: Theory and Applications*, *4*(1), 21–26. https://doi.org/10.31605/jomta.v4i1.1792

Zulaikha, S. (2024). *Kemenkomdigi Berhasil Blokir Ratusan Ribu Konten Judol*. Antaranews. https://www.antaranews.com/berita/4444513/kemenkomdigi-berhasil-blokir-ratusan-ribu-konten-judol