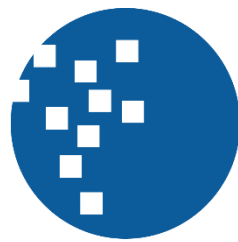


**PEMILIHAN MODEL KLASIFIKASI MACHINE LEARNING
UNTUK MODERASI KOMENTAR JUDI ONLINE DI
YOUTUBE**



UMN
UNIVERSITAS
MULTIMEDIA
NUSANTARA

SKRIPSI

Ray Theodorus
00000054779

**PROGRAM STUDI SISTEM INFORMASI
FAKULTAS TEKNIK DAN INFORMATIKA
UNIVERSITAS MULTIMEDIA NUSANTARA
TANGERANG
2025**

**PEMILIHAN MODEL KLASIFIKASI MACHINE LEARNING
UNTUK MODERASI KOMENTAR JUDI ONLINE DI
YOUTUBE**



Diajukan sebagai Salah Satu Syarat untuk Memperoleh

Gelar Sarjana Komputer

Ray Theodorus

00000054779

**PROGRAM STUDI SISTEM INFORMASI
FAKULTAS TEKNIK DAN INFORMATIKA
UNIVERSITAS MULTIMEDIA NUSANTARA**

TANGERANG

2025

HALAMAN PERNYATAAN TIDAK PLAGIAT

Dengan ini saya,

Nama : Ray Theodorus

Nomor Induk Mahasiswa : **00000054779**

Program Studi : Sistem Informasi

Skripsi dengan judul: PENGEMBANGAN MODEL KLASIFIKASI BERBASIS MACHINE LEARNING UNTUK MODERASI KOMENTAR JUDI ONLINE DI YOUTUBE

Merupakan hasil karya saya sendiri bukan plagiat dari laporan karya tulis ilmiah yang ditulis oleh orang lain, dan semua sumber, baik yang dikutip maupun dirujuk, telah saya nyatakan dengan benar serta dicantumkan di Daftar Pustaka.

Jika di kemudian hari terbukti ditemukan kecurangan/penyimpangan, baik dalam pelaksanaan maupun dalam penulisan laporan karya tulis ilmiah, saya bersedia menerima konsekuensi dinyatakan TIDAK LULUS untuk mata kuliah yang telah saya tempuh.

Tangerang, 4 Juni 2025



(Ray Theodorus)

HALAMAN PENGESAHAN

SKRIPSI dengan judul
PEMILIHAN MODEL KLASIFIKASI MACHINE LEARNING
UNTUK MODERASI KOMENTAR JUDI ONLINE DI YOUTUBE

Oleh

Nama : Ray Theodorus
NIM : 00000054779
Program Studi : Sistem Informasi
Fakultas : Teknik dan Informatika

Telah diujikan pada hari Rabu, 11 Juni 2025
Pukul 10.00 s.d 12.00 dan dinyatakan
LULUS

Dengan susunan penguji sebagai berikut.

Ketua Sidang



Dr. Irmawati, S.Kom., M.M.S.I.
0805097703

Penguji



Johan Setiawan, S.Kom., M.M., MBA
0327106402

Pembimbing



Ririn Ikana Desanti, S.Kom.,
M.Kom.
0313058001

Ketua Program Studi Sistem Informasi



Ririn Ikana Desanti, S.Kom., M.Kom.

iii

Pengembangan Model Klasifikasi..., Ray Theodorus, Universitas Multimedia Nusantara

HALAMAN PERSETUJUAN PUBLIKASI KARYA ILMIAH

Yang bertanda tangan di bawah ini:

Nama : Ray Theodorus

NIM : 00000054779

Program Studi : Sistem Informasi

Jenjang : S1

Judul Karya Ilmiah : PEMILIHAN MODEL KLASIFIKASI
MACHINE LEARNING UNTUK
MODERASI KOMENTAR JUDI ONLINE
DI YOUTUBE

Menyatakan dengan sesungguhnya bahwa saya bersedia* (**pilih salah satu**):

- ☒ Saya bersedia memberikan izin sepenuhnya kepada Universitas Multimedia Nusantara untuk mempublikasikan hasil karya ilmiah saya ke dalam repositori Knowledge Center sehingga dapat diakses oleh Sivitas Akademika UMN/Publik. Saya menyatakan bahwa karya ilmiah yang saya buat tidak mengandung data yang bersifat konfidensial.
- ☐ Saya tidak bersedia mempublikasikan hasil karya ilmiah ini ke dalam repositori Knowledge Center, dikarenakan: dalam proses pengajuan publikasi ke jurnal/konferensi nasional/internasional (dibuktikan dengan *letter of acceptance*) **.
- ☐ Lainnya, pilih salah satu:
 - ☒ Hanya dapat diakses secara internal Universitas Multimedia Nusantara
 - ☐ Embargo publikasi karya ilmiah dalam kurun waktu 3 tahun.

Tangerang, 4 Juni 2025



(Ray Theodorus)

* Pilih salah satu

** Jika tidak bisa membuktikan LoA jurnal/HKI, saya bersedia mengizinkan penuh karya ilmiah saya untuk dipublikasikan ke KC UMN dan menjadi hak institusi UMN.

KATA PENGANTAR

Puji syukur penulis panjatkan ke hadirat Tuhan Yang Maha Esa; berkat rahmat dan kasih karunia-Nya, skripsi berjudul “Pengembangan Model Klasifikasi Berbasis Machine Learning untuk Moderasi Komentar Judi Online di YouTube” dapat diselesaikan tepat sesuai jadwal. Karya ini disusun sebagai salah satu persyaratan kelulusan Program Sarjana (S-1) Sistem Informasi, Fakultas Teknik dan Informatika, Universitas Multimedia Nusantara.

Mengucapkan terima kasih kepada:

1. Bapak Dr. Ir. Andrey Andoko, selaku Rektor Universitas Multimedia Nusantara.
2. Bapak Dr. Eng. Niki Prastomo, S.T., M.Sc., selaku Dekan Fakultas Teknik dan Informatika Universitas Multimedia Nusantara.
3. Ibu Ririn Ikana Desanti, S.Kom., M.Kom. selaku Ketua Program Studi Sistem Informasi Universitas Multimedia Nusantara dan Pembimbing yang telah memberikan masukan, arahan, dan motivasi selama penyusunan skripsi
4. Keluarga yang telah memberikan dukungan berupa doa, semangat, dan dukungan materil sehingga penulis dapat menyelesaikan skripsi ini.

Saya berharap karya ilmiah ini dapat memberi manfaat bagi berbagai pihak—baik sebagai sumber informasi, rujukan, maupun inspirasi dalam melaksanakan penelitian sejenis di masa mendatang.

Tangerang, 4 Juni 2025



Ray Theodorus

PEMILIHAN MODEL KLASIFIKASI MACHINE LEARNING UNTUK MODERASI KOMENTAR JUDI ONLINE DI YOUTUBE

Ray Theodorus

ABSTRAK

YouTube memiliki lebih dari 2,5 miliar pengguna aktif bulanan per 2023, kendati demikian kini di Indonesia kolom komentar pada lama tersebut kerap disalahgunakan untuk promosi judi *online*. Hal ini memberikan pengalaman buruk bagi pengguna serta citra buruk bagi pengelola akun. Saat ini sistem moderasi otomatis bawaan YouTube masih kesulitan mengenali bahasa manipulatif yang digunakan pelaku *spam*. Penelitian ini bertujuan mengembangkan model klasifikasi berbasis *machine learning* yang mampu menyaring komentar promosi judi *online* secara lebih akurat dan adaptif.

Data diperoleh melalui YouTube Data API dari 20 channel terpilih (50 video terakhir per channel; cut-off 9 Maret 2025) dan diberi label menjadi dua kelas: judi *online* (*spam*) dan *non-spam*. Setelah pembersihan teks, tokenisasi TF-IDF, pembagian data latih-uji 80/20, serta undersampling untuk menyeimbangkan kelas, dilakukan benchmarking 11 algoritma. K-Nearest Neighbors (KNN) serta Logistic Regression menunjukkan performa terbaik dan selanjutnya akan dikomparasi melalui optimasi dengan RandomizedSearchCV.

Hasil evaluasi awal menempatkan KNN sebagai model terakurat (*accuracy* 98,48 % dan *F1-score* makro 83,07 %), dan hyperparameter tuning pada KNN meningkatkan *recall* makro menjadi 94 % meski menurunkan *accuracy* menjadi 97 % dan *F1-score* makro menjadi 78,30 %. Sebaliknya, Logistic Regression ter-tuning ($C=5,95$; $class_weight='balanced'$; $penalty='l2'$; $solver='sag'$) meningkatkan *recall* makro menjadi 97 % dengan penurunan *precision* dan *F1-score* yang minimal serta *runtime* hanya 0,05 s. Dengan prioritas meminimalkan *false negatives* sambil menjaga *accuracy* dan *precision*, Logistic Regression ter-tuning dipilih sebagai solusi akhir, membuktikan efektivitas pendekatan *machine learning* dalam moderasi komentar judi online di YouTube.

Kata kunci: Judi online, Klasifikasi teks, Machine learning, Moderasi komentar, YouTube

SELECTION OF MACHINE LEARNING CLASSIFICATION MODEL FOR ONLINE GAMBLING COMMENT MODERATION ON YOUTUBE

Ray Theodorus

ABSTRACT (English)

YouTube had over 2.5 billion monthly active users by 2023, yet in Indonesia, comments are often abused for online gambling promotion, degrading user experience and channel reputations. YouTube's moderation struggles to detect manipulative spam language, motivating this study to develop a more accurate, adaptive ML classifier.

We collected 1 071 028 comments via YouTube Data API from 20 channels (50 latest videos each; cut-off March 9, 2025). After deduplication, 897 894 unique comments remained (17 183 spam, 880 711 non-spam). Texts were cleaned, TF-IDF vectorized (5 000 features), and split 80/20. Classes were balanced via random undersampling, and 11 algorithms benchmarked. K-Nearest Neighbors (KNN) and Logistic Regression (LR) performed best and were optimized with RandomizedSearchCV.

Initial evaluation placed KNN first (accuracy 98.48 %, macro F1 83.07 %). Tuning KNN (ball_tree, leaf_size=42, n_neighbors=2, p=2, weights='distance') raised macro recall to 94 % but dropped accuracy to 97 % and macro F1 to 78.3 %. Tuned LR (C=5.95, class_weight='balanced', penalty='l2', solver='sag') achieved 97 % macro recall with minor precision/F1 losses and runtime 0.05 s. Prioritizing minimal false negatives while preserving accuracy and precision, we select tuned LR as the final solution, proving ML's effectiveness in moderating gambling spam on YouTube.

Keywords: *online gambling, text classification, machine learning, comment moderation, YouTube*

DAFTAR ISI

HALAMAN PERNYATAAN TIDAK PLAGIAT.....	2
HALAMAN PERSETUJUAN.....	3
HALAMAN PERSETUJUAN PUBLIKASI KARYA ILMIAH.....	4
ABSTRAK.....	6
<i>ABSTRACT (English).....</i>	7
DAFTAR ISI.....	8
DAFTAR TABEL.....	10
DAFTAR GAMBAR.....	11
DAFTAR LAMPIRAN.....	12
BAB I	
PENDAHULUAN.....	1
1.1 Latar Belakang.....	1
1.2 Rumusan Masalah.....	4
1.3 Batasan Masalah.....	4
1.4 Tujuan dan Manfaat Penelitian.....	4
1.4.1 Tujuan Penelitian.....	4
1.4.2 Manfaat Penelitian.....	5
1.5 Sistematika Penulisan.....	5
BAB II	
LANDASAN TEORI.....	7
2.1 Penelitian Terdahulu.....	7
2.2 Tinjauan Teori.....	11
2.2.1 Text Classification.....	12
2.2.2 Moderasi Komentar Otomatis.....	12
2.2.3 Undersampling.....	12
2.2.4 Representasi Teks dengan TF-IDF.....	13
2.2.5 Algoritma Machine Learning untuk Klasifikasi.....	14
2.3 Framework dan Algoritma.....	22
2.3.1 RandomizedSearchCV.....	22
2.3.2 Cross-Validation.....	23
2.4 Tools dan software yang digunakan.....	24
2.4.1 Scikit-learn.....	24
2.4.2 Pandas dan NumPy.....	26
2.4.3 SciPy.....	27
2.4.4 XGBoost dan LightGBM Library.....	28
BAB III	

METODOLOGI PENELITIAN.....	31
3.1 Gambaran Umum Objek Penelitian.....	31
3.2 Metode Penelitian.....	32
3.2.1 Alur Penelitian.....	33
3.2.2 Metode Data Mining.....	34
3.3 Teknik Pengumpulan Data.....	40
3.4 Teknik Analisis Data.....	41
BAB IV	
ANALISIS DAN HASIL PENELITIAN.....	44
4.1 Business Understanding.....	44
4.2 Data Acquisition & Understanding.....	45
4.2.1 Pengumpulan Data.....	45
4.2.2 Pelabelan Data.....	47
4.2.3 Exploratory Data Analysis (EDA).....	48
4.3 Modeling.....	51
4.3.1 Data Preprocessing.....	51
4.3.2 Modeling.....	54
4.4 Evaluation.....	56
BAB V	
SIMPULAN DAN SARAN.....	60
5.1. Kesimpulan.....	60
5.2. Saran.....	61
DAFTAR PUSTAKA.....	62
LAMPIRAN.....	70



DAFTAR TABEL

Tabel 3.1 Perbandingan Alat Penelitian	42
Tabel 4.1. Evaluasi Benchmarking Model	57
Tabel 4.2. Perbandingan KNN Baseline Dengan Hyperparameter-Tuned	58



DAFTAR GAMBAR

Gambar 3.1 Alur penelitian	33
Gambar 3.2. Diagram Modelling	38
Gambar 4.1. Contoh Komentar Spam Judi Online pada YouTube	45
Gambar 4.2. Tampilan Lama YouTube API V3	46
Gambar 4.3. Fungsi Untuk Mengambil Video berdasarkan Channel YouTube	46
Gambar 4.4. Fungsi untuk Mengambil Komentar Berdasarkan Video	47
Gambar 4.5. Tampilan Proses Pelabelan Data pada Google Spreadsheet	48
Gambar 4.6. Word Cloud pada Kelas Positif (Kiri) dan Kelas Negatif (Kanan)	49
Gambar 4.7. Distribusi Jumlah Karakter Berdasarkan Kelas	50
Gambar 4.8. Distribusi Jumlah Huruf Berdasarkan Kelas	50
Gambar 4.9. Grafik Distribusi Kelas	51
Gambar 4.10. Grafik Perbandingan Data Sebelum dan Sesudah Penghapusan Duplikat	52
Gambar 4.11. Fungsi Pembersihan Data Teks	53
Gambar 4.12. Jumlah Data per Kelas Sebelum dan Sesudah Undersampling	53
Gambar 4.13. Inisiasi 12 Model Machine Learning untuk Benchmarking	54
Gambar 4.14. Proses Benchmarking Model Yang Telah Di-inisiasi	55
Gambar 4.15. Konfigurasi Hyperparameter-tuning KNN	56



DAFTAR RUMUS

2.1 Rumus Perhitungan TF-IDF	14
2.2 Rumus TF	14
2.2 Rumus IDF	14



UMN
UNIVERSITAS
MULTIMEDIA
NUSANTARA

DAFTAR LAMPIRAN

Lampiran A. Turnitin Similarity Report

71

Lampiran B. Konsultasi Bimbingan

76



UMN

UNIVERSITAS
MULTIMEDIA
NUSANTARA