

**KLASIFIKASI KOMENTAR SPAM JUDI ONLINE BERBASIS
NLP PADA PLATFORM YOUTUBE MENGGUNAKAN
ALGORITMA NAÏVE BAYES DAN SVM**

SKRIPSI

Diajukan untuk memenuhi salah satu syarat mencapai derajat Sarjana
Program Studi Informatika



disusun oleh

GINARIS BUONO

22.11.5068

Kepada

**FAKULTAS ILMU KOMPUTER
UNIVERSITAS AMIKOM YOGYAKARTA
YOGYAKARTA
2026**

**KLASIFIKASI KOMENTAR SPAM JUDI ONLINE BERBASIS
NLP PADA PLATFORM YOUTUBE MENGGUNAKAN
ALGORITMA NAÏVE BAYES DAN SVM**

SKRIPSI

untuk memenuhi salah satu syarat mencapai derajat Sarjana
Program Studi Informatika



disusun oleh

GINARIS BUONO

22.11.5068

Kepada

**FAKULTAS ILMU KOMPUTER
UNIVERSITAS AMIKOM YOGYAKARTA
YOGYAKARTA**

2026

HALAMAN PERSETUJUAN

SKRIPSI

**KLASIFIKASI KOMENTAR SPAM JUDI ONLINE BERBASIS NLP
PADA PLATFORM YOUTUBE MENGGUNAKAN ALGORITMA NAÏVE
BAYES DAN SVM**

yang disusun dan diajukan oleh

GINARIS BUONO

22.11.5068

telah disetujui oleh Dosen Pembimbing Skripsi
pada tanggal 22 Mei 2026

Dosen Pembimbing,



Ike Verawati, S.Kom., M.Kom.
NIK. 190302237

HALAMAN PENGESAHAN
SKRIPSI
KLASIFIKASI KOMENTAR SPAM JUDI ONLINE BERBASIS NLP
PADA PLATFORM YOUTUBE MENGGUNAKAN ALGORITMA NAÏVE
BAYES DAN SVM

yang disusun dan diajukan oleh

GINARIS BUONO

22.11.5068

Telah dipertahankan di depan Dewan Penguji
pada tanggal 22 Mei 2026

Susunan Dewan Penguji

Nama Penguji

Tanda Tangan

Subektiningsih, S.Kom., M.Kom.
NIK. 190302413

Supriatin, A.Md., S.Kom., M.Kom.
NIK. 190302239

Ike Verawati, S.Kom., M.Kom.
NIK. 190302237

Skripsi ini telah diterima sebagai salah satu persyaratan
untuk memperoleh gelar Sarjana Komputer
Tanggal 22 Mei 2026

DEKAN FAKULTAS ILMU KOMPUTER



Prof. Dr. Kusriani, M.Kom.
NIK. 190302106

HALAMAN PERNYATAAN KEASLIAN SKRIPSI

Yang bertandatangan di bawah ini,

Nama mahasiswa : GINARIS BUONO
NIM : 22.11.5068

Menyatakan bahwa Skripsi dengan judul berikut:

KLASIFIKASI KOMENTAR SPAM JUDI ONLINE BERBASIS NLP PADA PLATFORM YOUTUBE MENGGUNAKAN ALGORITMA NAÏVE BAYES DAN SVM

Dosen Pembimbing : Ike Verawati, S.Kom., M.Kom.

1. Karya tulis ini adalah benar-benar ASLI dan BELUM PERNAH diajukan untuk mendapatkan gelar akademik, baik di Universitas AMIKOM Yogyakarta maupun di Perguruan Tinggi lainnya.
2. Karya tulis ini merupakan gagasan, rumusan dan penelitian SAYA sendiri, tanpa bantuan pihak lain kecuali arahan dari Dosen Pembimbing.
3. Dalam karya tulis ini tidak terdapat karya atau pendapat orang lain, kecuali secara tertulis dengan jelas dicantumkan sebagai acuan dalam naskah dengan disebutkan nama pengarang dan disebutkan dalam Daftar Pustaka pada karya tulis ini.
4. Perangkat lunak yang digunakan dalam penelitian ini sepenuhnya menjadi tanggung jawab SAYA, bukan tanggung jawab Universitas AMIKOM Yogyakarta.
5. Pernyataan ini SAYA buat dengan sesungguhnya, apabila di kemudian hari terdapat penyimpangan dan ketidakbenaran dalam pernyataan ini, maka SAYA bersedia menerima SANKSI AKADEMIK dengan pencabutan gelar yang sudah diperoleh, serta sanksi lainnya sesuai dengan norma yang berlaku di Perguruan Tinggi.

Yogyakarta, 22 Mei 2026

Yang Menyatakan,



Ginaris Buono

HALAMAN PERSEMBAHAN

Dengan penuh rasa syukur kepada Tuhan Yang Maha Esa, skripsi ini penulis persembahkan kepada:

1. Kedua orang tua dan keluarga tercinta yang senantiasa memberikan doa, dukungan, motivasi, serta pengorbanan selama proses perkuliahan hingga penyusunan skripsi ini selesai.
2. Dosen pembimbing dan seluruh dosen Universitas AMIKOM Yogyakarta yang telah memberikan ilmu, arahan, dan bimbingan selama masa studi.
3. Sahabat dan rekan-rekan yang telah memberikan bantuan, semangat, serta dukungan dalam proses penyusunan skripsi ini.
4. Semua pihak yang telah membantu dan mendukung penulis, baik secara langsung maupun tidak langsung, sehingga skripsi ini dapat diselesaikan dengan baik.

KATA PENGANTAR

Puji syukur penulis panjatkan kehadiran Tuhan Yang Maha Esa, yang telah melimpahkan rahmat, hidayah, serta karunia-Nya, sehingga penulis dapat menyelesaikan laporan tugas akhir ini dengan lancar. Tugas akhir ini disusun untuk memenuhi salah satu syarat kelulusan Program Sarjana pada Fakultas Ilmu Komputer Universitas Amikom Yogyakarta.

Selain itu penulis dengan segala kerendahan hati ingin menyampaikan rasa terima kasih kepada semua pihak yang telah berjasa memberikan dukungan dan bantuan untuk menyelesaikan laporan tugas akhir ini. Untuk itu penulis mengucapkan terima kasih kepada :

1. Prof. Dr. M. Suyanto, M.M. selaku Rektor Universitas Amikom Yogyakarta
2. Prof. Dr. Kusrini, M.Kom. selaku Dekan Program Fakultas Ilmu Komputer
3. Eli Pujastuti, M.Kom. selaku Ketua Program Studi Informatika
4. Ike Vewawati, S.Kom., M.Kom. selaku dosen pembimbing yang memberikan arahan, saran, dan motivasi terhadap penulis
5. Kedua orang tua, keluarga besar, dan teman-teman tercinta yang memberikan semangat dan doa kepada penulis.
6. Seluruh staff dan rekan kerja pada Direktorat Administrasi Akademik Kemahasiswaan

Yogyakarta, 20 Mei 2026

Ginaris Buono

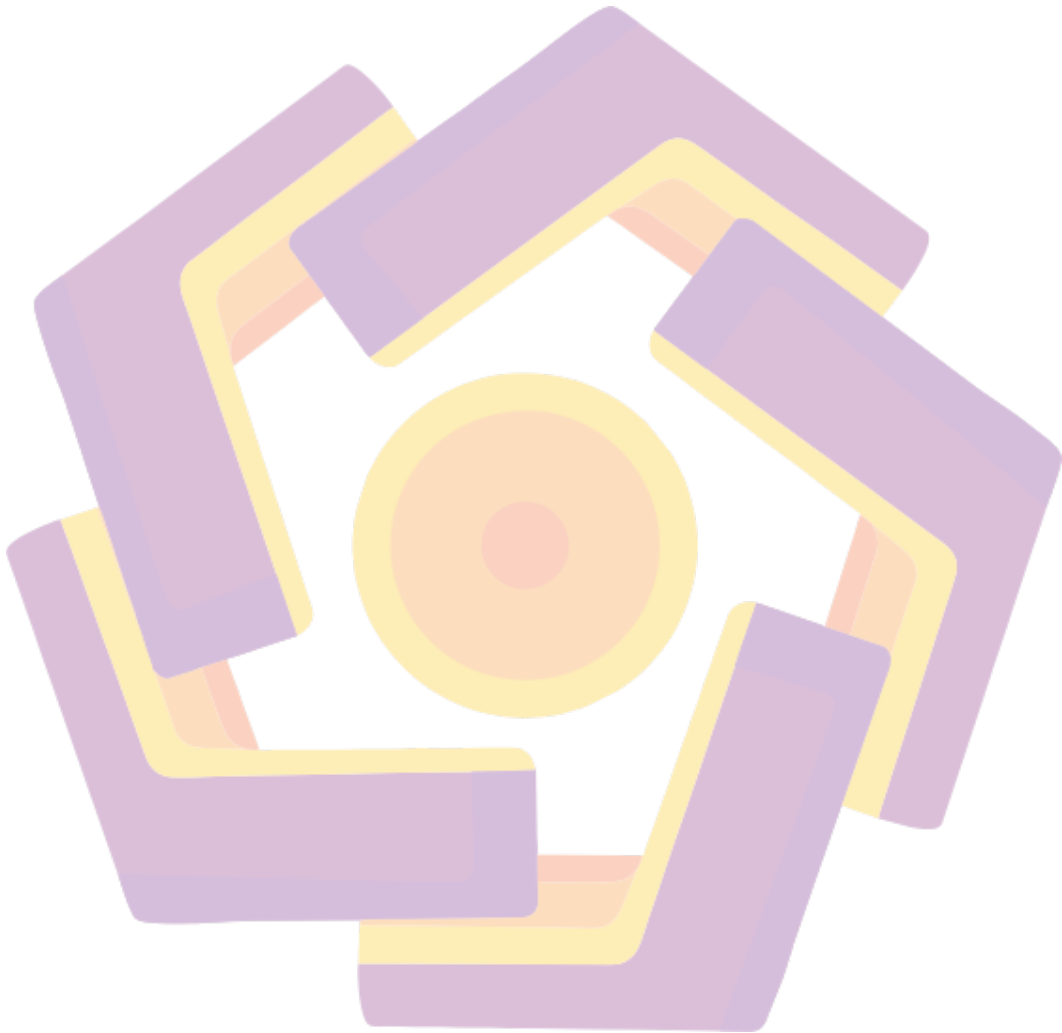
DAFTAR ISI

HALAMAN JUDUL	ii
HALAMAN PERSETUJUAN.....	iii
HALAMAN PENGESAHAN	iv
HALAMAN PERNYATAAN KEASLIAN SKRIPSI	v
HALAMAN PERSEMBAHAN	vi
KATA PENGANTAR	vii
DAFTAR ISI.....	viii
DAFTAR TABEL.....	xii
DAFTAR GAMBAR	xiii
INTISARI	xiv
<i>ABSTRACT</i>	xv
BAB I PENDAHULUAN.....	1
1.1 Latar Belakang	1
1.2 Rumusan Masalah	3
1.3 Batasan Masalah	4
1.4 Tujuan Penelitian	5
1.5 Manfaat Penelitian	5
1.6 Sistematika Penulisan	6
BAB II TINJAUAN PUSTAKA	8
2.1 Studi Literatur	8
2.2 Dasar Teori.....	15
2.2.1 Klasifikasi Teks	15

2.2.2	Natural Language processing (NLP).....	15
2.2.3	Text Normalization pada Media Sosial.....	16
2.2.4	Ruled-Based Entity Detection.....	17
2.2.5	Representasi Teks (Bag-of-Words dan TF-IDF).....	17
2.2.6	Machine Learning	19
2.2.7	Algoritma Naïve Bayes.....	19
2.2.8	Algoritma SVM	21
2.2.9	Confusion Matrix	22
2.2.10	Evaluasi Model Klasifikasi	23
2.2.11	Imbalanced Dataset dan Data Balancing	25
BAB III METODE PENELITIAN		27
3.1	Objek Penelitian.....	27
3.2	Alur Penelitian	28
3.2.1	Akuisisi Data.....	30
3.2.2	Data Preprocessing.....	30
3.2.3	Pelabelan Data	31
3.2.4	Data Balancing.....	31
3.2.5	Pembagian Dataset.....	32
3.2.6	Ekstraksi Fitur.....	32
3.2.7	Pemodelan Machine Learning	34
3.2.8	Evaluasi Model	37
3.2.9	Hasil dan Analisis	37
3.2.10	Kesimpulan	38
3.3	Alat dan Bahan.....	38
3.3.1	Data Penelitian	38

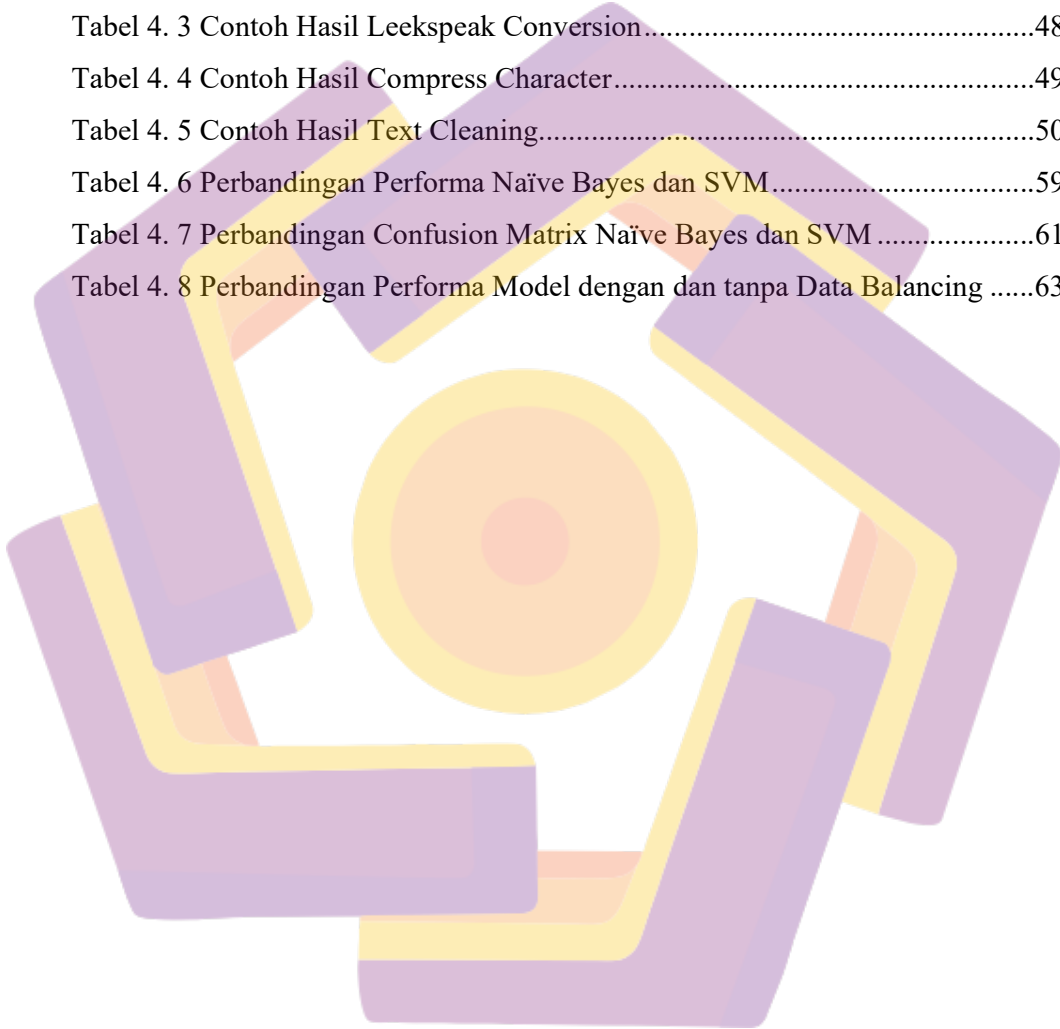
3.3.2	Alat/instrumen.....	40
BAB IV HASIL DAN PEMBAHASAN		43
4.1	Akuisisi Data.....	43
4.2	Preprocessing Data.....	45
4.2.1	Unicode Normalize	46
4.2.2	Case Folding	47
4.2.3	Leetspeak Conversion	48
4.2.4	Compress Character	48
4.2.5	Text Cleaning.....	49
4.3	Pelabelan Data	50
4.4	Data Balancing.....	53
4.5	Pembagian Dataset.....	54
4.6	Ekstraksi Fitur TF-IDF.....	55
4.7	Pemodelan Machine Learning	57
4.7.1	Naïve bayes.....	58
4.7.2	SVM.....	58
4.8	Evaluasi Model	58
4.8.1	Analisis Performa Model	58
4.8.1	Analisis Berdasarkan Confusion Matrix	60
4.9	Hasil Analisis.....	63
4.9.1	Analisis Pengaruh Data Balancing.....	63
4.9.2	Perbandingan Naïve Bayes dan SVM.....	64
BAB V PENUTUP		65
5.1	Kesimpulan	65
5.2	Saran	65

REFERENSI	67
-----------------	----



DAFTAR TABEL

Tabel 2. 1 Tabel Keaslian Penelितain	11
Tabel 4. 1 Contoh Hasil Unicode Normalize	46
Tabel 4. 2 Contoh Hasil Case Folding	47
Tabel 4. 3 Contoh Hasil Leekspeak Conversion	48
Tabel 4. 4 Contoh Hasil Compress Character	49
Tabel 4. 5 Contoh Hasil Text Cleaning	50
Tabel 4. 6 Perbandingan Performa Naïve Bayes dan SVM	59
Tabel 4. 7 Perbandingan Confusion Matrix Naïve Bayes dan SVM	61
Tabel 4. 8 Perbandingan Performa Model dengan dan tanpa Data Balancing	63



DAFTAR GAMBAR

Gambar 3. 1 Alur Penelitian	29
Gambar 3. 2 Contoh Struktur Data	39
Gambar 3. 3 Distribusi Dataset.....	40
Gambar 4. 1 Contoh Hasil Crawling.....	44
Gambar 4. 2 Daftar Situs Pada Site Root.....	51
Gambar 4. 3 Distribusi Hasil Pelabelan	52
Gambar 4. 4 Hasil Proses Data Balancing	53
Gambar 4. 5 Hasil Distribusi Data Training dan Testing	55
Gambar 4. 6 Hasil Dimensi Matrix TF-IDF	57
Gambar 4. 7 Hasil Evaluasi Model Naïve Bayes.....	59
Gambar 4. 8 Hasil Evaluasi Model SVM	59
Gambar 4. 9 Confusion Matrix Naïve bayes	60
Gambar 4. 10 Confusion Matrix SVM	61
Gambar 4. 11 Perbandingan Semua Metrik	63

INTISARI

Perkembangan media sosial, khususnya platform YouTube, menyebabkan meningkatnya aktivitas interaksi digital melalui kolom komentar. Namun, tingginya aktivitas tersebut juga memunculkan permasalahan berupa penyebaran komentar spam promosi judi online yang dapat mengganggu kenyamanan pengguna serta menimbulkan dampak sosial dan ekonomi yang merugikan. Karakteristik komentar spam yang menggunakan variasi penulisan seperti Unicode dekoratif, leetspeak, dan pengulangan karakter menyebabkan proses deteksi menjadi lebih kompleks. Oleh karena itu, diperlukan sistem klasifikasi otomatis yang mampu mendeteksi komentar spam secara efektif menggunakan pendekatan Natural Language Processing (NLP) dan machine learning.

Penelitian ini menggunakan dataset sebanyak 10.946 komentar YouTube yang diperoleh melalui YouTube Data API. Tahapan penelitian meliputi preprocessing teks berupa Unicode normalization, case folding, leetspeak conversion, compress character, dan text cleaning. Pelabelan data dilakukan secara otomatis menggunakan pendekatan rule-based entity detection berbasis pola nama situs judi online. Selanjutnya, representasi teks dilakukan menggunakan metode TF-IDF dan proses klasifikasi menggunakan algoritma Naïve Bayes dan Support Vector Machine (SVM). Penelitian juga menerapkan teknik data balancing menggunakan Random UnderSampling untuk menganalisis pengaruh ketidakseimbangan kelas terhadap performa model.

Hasil penelitian menunjukkan bahwa algoritma SVM memiliki performa lebih baik dibandingkan Naïve Bayes dalam mengklasifikasikan komentar spam judi online berdasarkan metrik accuracy, precision, recall, dan F1-score. Selain itu, penerapan data balancing terbukti mampu meningkatkan kemampuan model dalam mendeteksi kelas spam. Penelitian ini diharapkan dapat menjadi referensi dalam pengembangan sistem moderasi otomatis komentar media sosial serta mendukung upaya deteksi spam berbasis teks pada platform digital.

Kata kunci: NLP, spam komentar, judi online, Naïve Bayes, SVM

ABSTRACT

The rapid growth of social media, particularly the YouTube platform, has increased digital interaction activities through comment sections. However, this high level of activity has also led to the spread of online gambling promotional spam comments, which may disrupt user experience and create negative social and economic impacts. The characteristics of spam comments, such as the use of decorative Unicode characters, leetspeak, and repetitive characters, make the detection process more challenging. Therefore, an automatic classification system is needed to effectively detect spam comments using Natural Language Processing (NLP) and machine learning approaches.

This study used a dataset of 10,946 YouTube comments collected through the YouTube Data API. The research stages included text preprocessing consisting of Unicode normalization, case folding, leetspeak conversion, character compression, and text cleaning. Data labeling was performed automatically using a rule-based entity detection approach based on online gambling site name patterns. Furthermore, text representation was conducted using the TF-IDF method, while the classification process utilized the Naïve Bayes and Support Vector Machine (SVM) algorithms. This study also applied a data balancing technique using Random UnderSampling to analyze the effect of class imbalance on model performance.

The results showed that the SVM algorithm achieved better performance than Naïve Bayes in classifying online gambling spam comments based on accuracy, precision, recall, and F1-score metrics. In addition, the implementation of data balancing was proven to improve the model's ability to detect spam classes. This research is expected to serve as a reference for developing automatic moderation systems for social media comments and to support text-based spam detection on digital platforms.

Keywords: NLP, spam comments, online gambling, Naïve Bayes, SVM