Jurnal Manajemen Informatika & Sistem Informasi (MISI)
Volume 8, Nomor 2, Juni 2025.
ISSN 2614-1701 (media cetak)
ISSN 2614-3739 (media online)
DOI : 10.36595/misi.v5i2
http://e-journal.stmiklombok.ac.id/index.php/misi

IMPLEMENTASI CHATBOT PADA TELEGRAM UNTUK MENDETEKSI SPAM MENGGUNAKAN METODE XGBOOST

Fery Juliyanto[1], Husain[2], Lilik Widyawati[3], Khairan Marzuki[4], I Made Yadi Dharma[5]
[123,]Program Studi Teknologi Informasi, [45,]Program Studi Ilmu Komputer Universitas Bumigora,

Jl. Ismail Marzuki No.22, Universitas Bumigora, Cilinaya, Cakranegara, Mataram 83127
[1] xniners2222@gmail.com, [2] husain@universitasbumigora.ac.id, [3] lilikwidya@universitasbumigora.ac.id,
[4] khairan.marzuki @universitasbumigora.ac.id, [5] yadi_dharma@universitasbumigora.ac.id,

Abstract

The increasing use of the Telegram application as a communication medium has driven the need for an automated system capable of detecting spam messages in real time, as spam messages can disrupt user comfort and even cause harm. This study aims to design and implement a chatbot on Telegram to identify spam messages using the XGBoost method. The development process follows the CRISP-DM steps, which include business understanding, data preparation, modeling, evaluation, and deployment. The dataset is divided into two parts—training data and testing data—to evaluate the model's performance. The XGBoost model is trained using the training data and tested with the testing data across three proportion split scenarios: 70% - 30%, 80% - 20%, and 90% - 10%, in order to calculate the accuracy in classifying messages as spam or non-spam (ham). The results show that the chatbot can accurately recognize spam messages and take actions such as issuing warnings, deleting spam messages, and blocking senders if the same spam message is repeated more than twice. Therefore, the system can maintain communication comfort on the Telegram platform in an automated and flexible manner. The highest accuracy achieved across all testing scenarios is 0.9656 or 96.57%. The average precision for ham and spam messages is 96% and 97%, respectively; recall is 96% and 97%; and the F1-score is 96% for ham and 97% for spam. The confusion matrices for each scenario show accuracies of 94.86% (70% - 30%), 94.42% (80% - 20%), and 94.87% (90% - 10%). The main contribution of this study is the development of an XGBoost-based chatbot system integrated into the Telegram application, capable of real-time detection and handling of spam messages. This system helps create a safer, more comfortable communication environment that is automatically protected from spam disruptions.

Keywords : *Detection, XGBoost, CRISP-DM,, Spam, Confusion Matrix.*

Abstrak

Meningkatnya pemakaian aplikasi Telegram sebagai sarana komunikasi mendorong kebutuhan akan sistem otomatis yang bisa mendeteksi pesan spam secara langsung, karena pesan spam dapat mengganggu kenyamanan pengguna dan bahkan menyebabkan kerugian. Studi ini bertujuan untuk merancang dan menerapkan chatbot di Telegram untuk mengidentifikasi pesan spam dengan metode XGBoost. Proses pengembangan mengikuti langkah-langkah CRISP-DM, yang mencakup business understanding, data preparation, modeling, evaluation, dan deployment. Dataset terbagi menjadi dua bagian, yaitu data pelatihan dan data pengujian, untuk mengukur kinerja model. Model XGBoost dilatih menggunakan data training dan diuji dengan data testing pada tiga skenario pembagian proporsi, yaitu 70% - 30%, 80% - 20%, dan 90% - 10%, guna menghitung tingkat akurasi dalam mengklasifikasikan pesan sebagai spam atau non spam (ham). Hasil penerapan menunjukkan bahwa chatbot mampu mengenali pesan spam dengan baik dan dapat melakukan tindakan seperti memberikan peringatan, menghapus pesan spam, serta memblokir pengirim jika ada pengulangan pesan spam yang sama lebih dari dua kali. Oleh karena itu, sistem ini dapat mempertahankan kenyamanan komunikasi di platform Telegram dengan cara yang otomatis dan fleksibel. Tingkat akurasi tertinggi yang diraih dari semua pengujian adalah 0.9656 atau 96.57%. Rata-rata precision untuk pesan ham dan spam adalah 96% dan 97%, recall masing-masing

Jurnal Manajemen Informatika & Sistem Informasi (MISI)
Volume 8, Nomor 2, Juni 2025.
ISSN 2614-1701 (media cetak)
ISSN 2614-3739 (media online)
DOI : 10.36595/misi.v5i2
http://e-journal.stmiklombok.ac.id/index.php/misi

sebesar 96% dan 97%, serta F1-score masing-masing 96% dan 97%. Matriks kebingungan pada setiap skenario menunjukkan akurasi yaitu 94.86% (70% - 30%), 94.42% (80% - 20%), dan 94.87% (90% - 10%). Kontribusi utama dari studi ini adalah hadirnya sistem chatbot berbasis XGBoost yang terintegrasi dalam aplikasi Telegram, yang dapat mendeteksi dan menangani pesan spam secara langsung. Sistem ini membantu menciptakan suasana komunikasi yang lebih aman, nyaman, dan secara otomatis terhindar dari gangguan spam.

Kata kunci : *Deteksi, XGBoost, CRISP-DM,, Spam, Confusion Matrix.*

1. PENDAHULUAN

Perkembangan teknologi informasi (TI) yang pesat telah menjadikannya bagian tak terpisahkan dari kehidupan organisasi, individu, dan masyarakat. TI tidak hanya menjadi sarana pendukung aktivitas, tetapi juga berperan strategis dalam menyelesaikan berbagai permasalahan secara efisien dan dapat diandalkan. [1] Pemanfaatan TI, khususnya melalui internet, telah membawa dampak signifikan pada struktur organisasi, otoritas, peran kerja, hingga sistem kepemimpinan di berbagai sektor. [1]

Dalam lingkungan organisasi modern, TI menjadi pilihan utama dalam merancang sistem informasi yang kuat untuk menciptakan keunggulan kompetitif. Investasi TI biasanya bertujuan meningkatkan kinerja baik di level individu maupun institusi. [2]

Salah satu wujud aplikasi TI dalam komunikasi adalah Telegram, sebuah platform pesan instan yang dapat diakses secara lintas perangkat, baik melalui aplikasi maupun web, dengan fungsi serupa seperti WhatsApp dan Facebook Messenger.

Pesan, dalam konteks komunikasi, merupakan informasi yang dikirimkan dari pengirim ke penerima, yang dapat berupa pengetahuan, saran, hiburan, atau informasi penting[3]. Namun, tidak semua pesan bersifat positif. Pesan spam—yakni pesan yang tidak diinginkan dan tidak bermakna—umumnya berisi promosi, penipuan, atau informasi tidak penting yang dikirim secara massal [4]. Spam telah menjadi isu global. Truecaller mencatat lebih dari 31,3 miliar pesan spam tersebar di seluruh dunia sepanjang tahun 2020, termasuk penipuan, layanan keuangan palsu, dan telemarketing yang agresif [5]

Keberadaan chat spam sangat mengganggu maka dibutuhkan pengembangan berupa bot untuk mendeteksi dan memfilter pesan yang masuk dari orang yang tidak diketahui, spam itu sendiri masih sering terjadi diaplikasi telegram yang sangat merugikan bagi pengguna yang aktif diaplikasi telegram, maka penulis mengembangkan chatbot diaplikasi telegram

Para ahli teknologi berpendapat bahwa chatbot adalah salah satu teknologi yang sangat baik dalam hal perkembangan. Ini terlihat dari banyaknya chatbot yang beroperasi di beragam platform. Sebagai contoh, pada messenger Facebook, per April 2017, ada lebih dari 100.000 chatbot yang aktif setiap bulannya. .[6]

Penelitian ini mengembangkan chatbot berbasis algoritma XGBoost (Extreme Gradient Boosting), yang merupakan pengembangan dari *gradient tree boosting*. XGBoost dikenal efektif dalam menangani data berskala besar dengan akurasi dan kemampuan interpretasi yang tinggi. [7] Dibandingkan algoritma lain seperti SVM, BPNN, dan Random Forest, XGBoost memiliki keunggulan dalam generalisasi model dan efisiensi[8].

Contoh Kasus pesan spam pada 16 Mei 2021 pembobolan m-banking akibat link phising yang merugikan korban mencapai puluhan juta yang dialami oleh sahabat dari pemilik akun tiktok @callme_syu yang di mana menjelaskan bahwa Tindakan pembobolan akun m-banking temannya dimulai setelah korban mendapatkan pesan yang berisi permintaan pembaruan biaya transaksi lewat link phising dari pelaku yang berpura-pura sebagai bank B*I. Korban kemudian meng-klik link tersebut dan mengisi beberapa pertanyaan terkait data sensitif. Akibatnya, pelaku berhasil meretas akun Mobile Banking korban, sementara korban kehilangan uangnya hingga puluhan juta dalam sekejap. Ia mengetahui hal tersebut saat menerima bukti transaksi uang. [9]

Jurnal Manajemen Informatika & Sistem Informasi (MISI)
Volume 8, Nomor 2, Juni 2025.
ISSN 2614-1701 (media cetak)
ISSN 2614-3739 (media online)
DOI : 10.36595/misi.v5i2
http://e-journal.stmiklombok.ac.id/index.php/misi

2. TINJAUAN PUSTAKA

2.1 Implementasi Penelitian

Implementasi merupakan proses pelaksanaan yang bersifat sistemik, melibatkan tindakan terencana dan terstruktur dalam mencapai tujuan yang ditetapkan. Dalam konteks pengembangan sistem, implementasi berarti menerapkan desain teknis dan algoritma ke dalam sistem nyata yang dapat dijalankan.[10].

2.2 Chatbot

Chatbot adalah aplikasi kecerdasan buatan (AI) yang dirancang untuk mensimulasikan percakapan dengan pengguna secara otomatis melalui teks atau suara. Performa chatbot sangat dipengaruhi oleh metode pelatihan dan teknik pemrosesan bahasa alami (NLP) yang digunakan [11] Mengembangkan chatbot untuk klasifikasi pesan promosi berbasis Rasa NLU, sementara studi ini mengimplementasikan chatbot berbasis XGBoost untuk mendeteksi spam pada Telegram, yang bertujuan meningkatkan presisi deteksi dan kemampuan pengambilan tindakan otomatis.

2.3 Python

Python adalah bahasa pemrograman serbaguna yang populer dalam bidang analisis data dan pembelajaran mesin karena sintaksisnya yang sederhana dan ekosistem pustaka yang kuat seperti Scikit-learn, NLTK, dan XGBoost. [12] Banyak penelitian sistem cerdas memanfaatkan Python untuk membangun pipeline NLP, model klasifikasi, dan deployment sistem secara integratif.

2.4 Bot Telegram

Telegram Bot merupakan API yang memungkinkan pengembang menciptakan interaksi otomatis di dalam platform Telegram. Bot ini mendukung berbagai fungsi, termasuk penerimaan pesan, pengolahan data teks, serta tindakan otomatis seperti menghapus atau membalas pesan. [13] Penelitian oleh Reynaldi et al. (2020) menggunakan Bot Telegram untuk Media Sosialisasi Keselamatan Berkendara, sementara dalam studi ini, bot dikembangkan untuk mendeteksi pesan spam secara real-time dengan menggunakan xgboost model.

2.5 NLP

NLP merupakan cabang ilmu komputer yang berfokus pada pemahaman konsep dan arti dari bahasa manusia. Meski manusia cukup terampil dalam memahami sintaksis linguistik dan aturan tata bahasa serta hubungan spasial yang tersirat, komputer mengalami kesulitan signifikan dalam memproses permintaan bahasa alami.[14] Tujuan NLP ialah merancang dan menciptakan aplikasi yang mendukung interaksi manusia dengan mesin serta perangkat lainnya melalui pemanfaatan bahasa alami [15]. Dalam studi oleh Aprian Putra dan et al. n.d. metode NLP sebagai media pencarian informasi menggunakan react. sedangkan Studi ini menggunakan NLP sebagai praproses sebelum klasifikasi menggunakan XGBoost untuk hasil yang lebih akurat.

2.6 TF-IDF

TF-IDF (Term Frequency–Inverse Document Frequency) adalah metode pembobotan kata yang populer dalam ekstraksi fitur pada data teks. TF-IDF menilai tingkat kepentingan suatu kata dalam dokumen relatif terhadap seluruh korpus [16]. Pada studi sebelumnya, seperti oleh Septiani et al. (2023), menggunakan TF-IDF untuk analisis dalam temu kembali informasi pada dokumen text. Sedangkan Penelitian ini mengkombinasikan TF-IDF dengan XGBoost untuk meningkatkan akurasi klasifikasi pesan spam.

2.7 Metode XGBoost

XGBoost merupakan algoritma ensemble yang mengembangkan pendekatan gradient tree boosting. Algoritma ini terkenal karena kemampuannya dalam menangani dataset besar dan kompleks secara efisien serta akurat [7]. XGBoost memiliki beberapa keunggulan dibandingkan algoritma pembelajaran mesin lainnya. Ia memiliki kemampuan untuk memodelkan sistem dan dikenal karena akurasi prediksi dan kemampuan interpretasinya yang unggul. Dibandingkan dengan metode pembelajaran mesin seperti Support Vector Machine, Back Propagation Neural Network, dan Random Forest, XGBoost memiliki generalisasi dan interpretabilitas yang lebih baik [8].

Jurnal Manajemen Informatika & Sistem Informasi (MISI)
Volume 8, Nomor 2, Juni 2025.
ISSN 2614-1701 (media cetak)
ISSN 2614-3739 (media online)
DOI : 10.36595/misi.v5i2
http://e-journal.stmiklombok.ac.id/index.php/misi

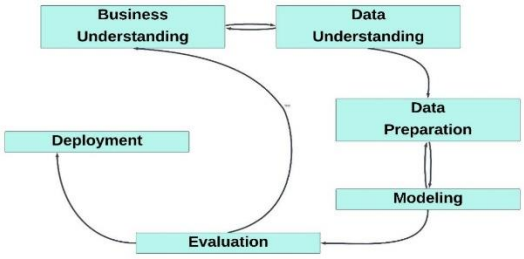
Dalam studi oleh Wang et al. (2022), XGBoost menghasilkan akurasi tinggi dalam Algoritma Pembelajaran Mesin XGBoost untuk Prediksi Hasil pada Perdarahan Subarachnoid Aneurisma. Sedangkan Studi ini memanfaatkan XGBoost untuk klasifikasi pesan spam, dengan keunggulan pada performa prediktif dan fleksibilitas dalam penanganan fitur teks

2.8 CRISP-DM

CRISP-DM (Cross Industry Standard Process for Data Mining) adalah sebuah standar untuk pemrosesan data mining yang telah dirancang agar data yang tersedia melalui setiap tahapan yang terstruktur, terdefinisi dengan baik, dan efisien. Di samping menerapkan model dalam proses penambangan data, pilihan algoritma berpengaruh besar terhadap perbandingan kinerja metode data mining.[17]. CRISP-DM banyak digunakan dalam penelitian aplikasi berbasis data karena bersifat fleksibel dan dapat diadaptasi untuk berbagai jenis proyek, termasuk pengembangan sistem deteksi spam seperti dalam penelitian ini.

3. METODOLOGI PENELITIAN

Pada penelitian ini membahas tentang bagaimana kinerja CRIPS-DM dalam upaya pemecahan masalah implementasi chatbot untuk deteksi spam diaplikasi telegram, secara garis besar proses dimulai dari *Business Understanding* kemudian *Data Understanding* setelah itu Data Understanding kemudian proses *data Preparation* yang dimana pada proses ini terdapat berbagai tahapan di dalam *Data Preparation*, setelah tahapan tersebut selanjutnya ke proses *Modeling* setelah proses modeling selesai maka selanjutnya ke tahapan *Evalution* dan tahapan yang terakhir adalah *Deployment.*



Gambar 1. Metodologi Penelitian

3.1. Business Understanding

Business Understanding menjadi dasar perancangan untuk memenuhi kebutuhan penelitian dan hasil dari tahapan ini dibagi menjadi 4 yaitu masalah penelitian, Solusi penelitian, tujuan penelitian dan Kriteria penelitian berhasil atau gagal.

1. **Masalah Penelitian**
   Pesan spam merupakan salah satu masalah yang sering terjadi yang mengakibatkan kenyamanan pengguna terganggu.
2. **Solusi Penelitian**
   Mengembangkan chatbot diaplikasi Telegram yang dapat mendeteksi dan memfilter pesan spam secara otomatis.
3. **Tujuan Penelitian**
   Untuk meningkatkan kenyamanan user menggunakan telegram yang terjaring dari pesan spam yang mengganggu bahkan sampai merugikan pengguna.
4. **Kriteria penelitian berhasil**
   Tanda penelitian berhasil adalah model XGBoost mencapai akurasi diatas 90% dan chatbot dapat memproses pesan baru kurang dari 3 detik

3.2. Data Understanding

Pada proses *data understanding* penelitian ini menggunakan bahasa pemrograman python dengan memanfaatkan dataset penelitian sebelumnya sebanyak 1133 data berjenis text yang terdiri dari pesan ham 559 dan spam 574 yang di labelkan 0 sebagai pesan ham atau non spam dan 1 sebagai pesan spam, penulis juga menambahkan sedikit data yang ada di telegram maupun pesan dibuat secara manual untuk data yang akan diuji sehingga total keseluruhan data adalah 1163 data.

3.3. Data Preparation

Tahapan ini merupakan tahapan dari proses analisis sentimen, pada penelitian ini dilakukan proses pengolahan data untuk melabelkan pesan SPAM maupun HAM. Pada tahap data Preparation, dilakukan berbagai proses

Jurnal Manajemen Informatika & Sistem Informasi (MISI)
Volume 8, Nomor 2, Juni 2025.
ISSN 2614-1701 (media cetak)
ISSN 2614-3739 (media online)
DOI : 10.36595/misi.v5i2
http://e-journal.stmiklombok.ac.id/index.php/misi

preprocessing teks, yang meliputi *case folding, tokenisasi, stopword removal,* dan *stemming*, yang bertujuan untuk memperbaiki kualitas data masukan ke model *XGBoost*.

3.4. Modeling

Pada fase peneliti menentukan teknik pemodelan untuk menemukan pola yaitu dengan algoritma xgboost sehingga dapat digunakan sebagai informasi dan pengetahuan bagi pengguna. Kelebihan dari algoritma xgboost adalah memberikan kinerja yang sangat baik dalam hal akurasi dibandingkan dengan algoritma lain. Pemodelan di bangun menggunakan *Visual Studio Code* dengan berbagai operator pendukung yang tersedia. Pengembangan model ChatBot pada aplikasi telegram yang menggunakan algoritma xgboost untuk mendeteksi pesan masuk yang berlabelkan Spam dan Ham atau non spam

3.5. Evaluation

Pada tahapan *Evaluation* penulis melibatkan beberapa pengukuran performa yang didapatkan untuk memastikan akurasi, presisi, recall, dan matrik lainnya dengan menggunakan *confusion matrix* dalam mendeteksi pesan spam dan ham. Pada tahapan ini proses mengevaluasi kinerja model dengan memberikan gambaran rinci tentang prediksi yang benar (*True Positives, True Negatives*) dan kesalahan (*False Positives, False Negatives*) dari *confusion matrix*.

1. TP (True Positive)
   Merepresentasikan jumlah model memprediksi positif dan memang benar-benar positif sesuai dengan label aslinya.
2. TN (True Negative)
   Merepresentasikan jumlah model memprediksi negatif dan memang benar-benar negatif sesuai dengan label aslinya.
3. FP (False Positive)
   Merepresentasikan jumlah model memprediksi positif, tetapi sebenarnya negatif kesalahan prediksi positif.
4. FN (False Negative)
   Merepresentasikan jumlah model memprediksi negatif, tetapi sebenarnya positif kesalahan prediksi negatif.

Berikut Kode bertujuan untuk mengambil nilai *True Positive (TP), True Negative (TN), False Positive (FP),* dan *False Negative (FN)* dari *confusion matrix.*

Tabel 1. Code Conf Matrix

TP = conf_matrix[1, 1]	FP = conf_matrix[0, 1]
TN = conf_matrix[0, 0]	FN = conf_matrix[1, 0]

Perhitungan matrik seperti *accuracy, precision, recall,* dan *specificity* adalah untuk mengevaluasi kinerja model dalam.

$$Akurasi = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Presisi = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$Spesifisitas = \frac{TN}{TN + FP}$$

$$F1\text{-}Score = 2 \cdot \frac{precision \,.\, Recall}{precision + Recall}$$

Jika diimplementasikan ke dalam kode program python maka menggunakan perintah seperti berikut :

```
accuracy = (TP + TN) / (TP + TN + FP + FN)
precision = TP / (TP + FP)
recall = TP / (TP + FN)
specificity = TN / (TN + FP)
f1_score = 2 * (precision * recall) / (precision + recall)
```

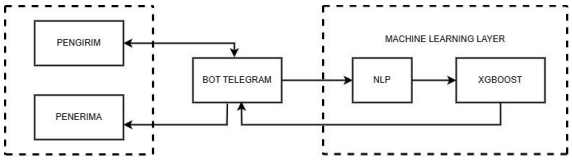
3.6. Deployment

Deployment merupakan tahap akhir dari metodologi *CRISP-DM*, model yang telah dilatih dan dievaluasi akan diimplementasikan. Pada penelitian ini, model XGBoost yang telah dikembangkan diintegrasikan dengan **Bot Telegram** untuk mendeteksi pesan spam secara otomatis. Bot ini mampu memantau pesan masuk dan memberikan respons sesuai dengan hasil prediksi model.

Jurnal Manajemen Informatika & Sistem Informasi (MISI)
Volume 8, Nomor 2, Juni 2025.
ISSN 2614-1701 (media cetak)
ISSN 2614-3739 (media online)
DOI : 10.36595/misi.v5i2
http://e-journal.stmiklombok.ac.id/index.php/misi

3.7. Processing alur Program

Pada proses ini alur kerja sistem yang melibatkan bot telegram dan menggunakan lapisan machine learning untuk mendeteksi spam yang melibatkan metode *Natural Language Processing (NLP)* dan menggunakan *XGBoost* sebagai model machine learning. Diagram pada gambar 3 bertujuan untuk menjelaskan alur proses dalam sistem chatbot pada telegram untuk mendeteksi spam dengan menggunakan metode *Xgboost* sebagai model machine learning.



Gambar 2. Processing Alur Program

4. HASIL DAN PEMBAHASAN

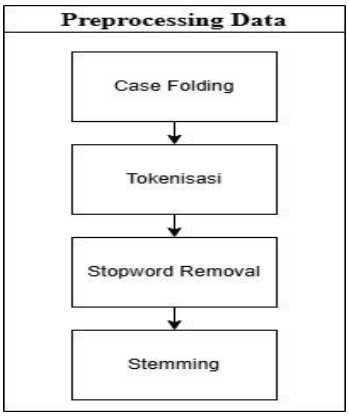
4.1. Data Preparation

Pada tahap ini, bertujuan untuk meningkatkan kualitas data masukan ke model *XGBoost*.

Alur Kerja untuk Deteksi Spam

Input: "Selamat Anda telah memenangkan hadiah utama, segera ambil!"

Preprocessing:



Gambar 3. Processing Data

Case Folding

Proses mengubah semua huruf kapital menjadi huruf kecil, untuk menyamakan format teks agar tidak ada perbedaan akibat penggunaan huruf kapital.

Case Folding

Tabel 2. Case Folding

Input	Output
Promo Special Untuk Anda, Segera Claim Untuk Mendapatkannya!	"promo special untuk anda, segera claim untuk mendapatkannya!"

Tokenisasi

Tabel 3. Tokenisasi

Input	Output
Selamat Anda telah memenangkan hadiah utama, segera ambil!	promo', 'special', 'untuk', 'anda', 'segera', 'claim', 'untuk', 'mendapatkannya'

Stopword Removal

Tabel 4. Stopword Removal

Input	Output
Selamat Anda telah memenangkan hadiah utama, segera ambil!	“ 'promo', 'special', 'segera', 'claim', 'mendapatkannya' "

Stemming

Tabel 5. Stemming

Input	Output
Selamat Anda telah memenangkan hadiah utama, segera ambil!	“ 'promo', 'special', 'segera', 'claim', 'dapat' “

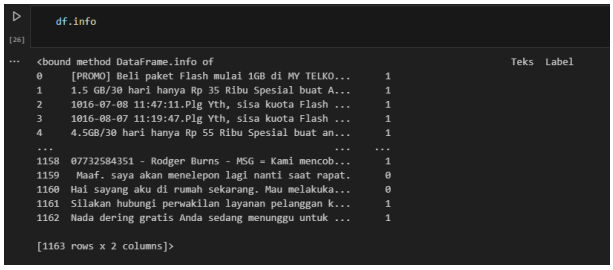
4.2. Modeling

Pada tahapan ini metode xgboost digunakan sebagai algoritma untuk mengimplementasi chatbot yang bertujuan untuk mendeteksi pesan yang masuk termasuk spam atau ham.

Load Dataset

Load dataset yang sudah disiapkan sebelumnya, dan berikut adalah contoh datasetnya. Menampilkan jumlah dataset yang sudah di load

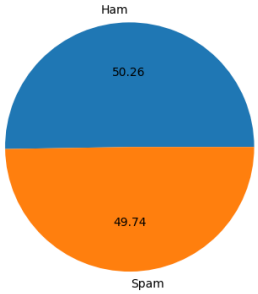
Jurnal Manajemen Informatika & Sistem Informasi (MISI)
Volume 8, Nomor 2, Juni 2025.
ISSN 2614-1701 (media cetak)
ISSN 2614-3739 (media online)
DOI : 10.36595/misi.v5i2
http://e-journal.stmiklombok.ac.id/index.php/misi



Gambar 4. Jumlah Dataset

Perbandingan Jumlah pesan Spam dan Ham (Non Spam)

Perbandingan jumlah pesan Spam dan Ham Perbandingan jumlah data antara pesan Spam dan pesan Ham yang dimana menunjukkan 0 sebagai pesan ham dan 1 sebagai spam atau bisa juga di lihat pada gambar di bawah, yang menunjukkan pesan ham ada dataset berjumlah 50,26% dan pesan spam pada dataset berjumlah 49,74%



Gambar 5. Perbandingan Jumlah Pesan

Pemisahan Data

Pisahkan data menjadi 2 fitur X = df["Teks"] Y = df["Label"]. X Biasanya digunakan untuk melatih model yang biasanya berisikan pesan yang akan diproses untuk memprediksi apakah pesan itu termasuk spam atau ham sedangkan nilai Y untuk menyimpan label atau target yang di prediksi yang berisikan tentang kategori pesan berupa spam atau ham

Training data

Membagi dataset menjadi data latih dan data uji. Berikut penjelasan setiap elemen:
Data Latih: Digunakan untuk melatih model.
Data Uji: Dimanfaatkan untuk menilai kinerja model pada data yang belum pernah diakses sebelumnya.
Setelah data dibagi maka training data pengujian menjadi 70% - 30%, 80% - 20% dan 90% - 10%
Training data 70% - 30%
814 Jumlah sampel dalam data pelatihan 70% dari total data
350 Jumlah sampel dalam data pengujian 30% dari total data
Training data 80% - 20%
931 Jumlah sampel dalam data pelatihan 80% dari total data
233 Jumlah sampel dalam data pengujian 20% dari total data
Training data 90% - 10%
1047 Jumlah sampel dalam data pelatihan 90% dari total data
117 Jumlah sampel dalam data pengujian 10% dari total data

Akurasi

Rasio Jumlah prediksi yang benar terhadap total jumlah prediksi sebesar 0.944206008583691 atau 94%

Presisi

Mengukur seberapa banyak prediksi "Spam" yang benar-benar spam dan total jumlah yang di dapat adalah 0.959349593495935 atau 95% total presisi

Recall

Recall mengukur seberapa banyak spam yang berhasil dideteksi sebagai spam. Total jumlah yang didapatkan adalah 0.9365079365079365 atau 93% total recall

F1 – Score

F1-Score merupakan rata-rata dari presisi dan recall. Berguna untuk dataset tidak seimbang. Dengan total yang didapatkan sebesar 0.9477911646586346 atau 94% total F1-Score

Cross - Validation

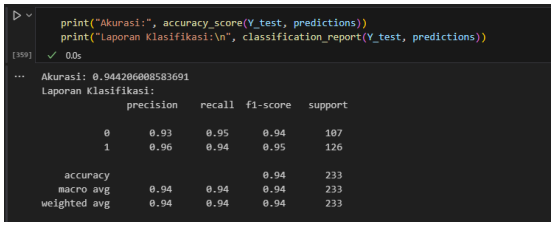
Untuk menghindari overfitting, lakukan validasi silang (cross-validation) pada data latih. Dengan total yang didapatkan sebesar

Jurnal Manajemen Informatika & Sistem Informasi (MISI)
Volume 8, Nomor 2, Juni 2025.
ISSN 2614-1701 (media cetak)
ISSN 2614-3739 (media online)
DOI : 10.36595/misi.v5i2
http://e-journal.stmiklombok.ac.id/index.php/misi

0.947369328963257 atau 94% total dari *Cross-validation Accuracy*

Laporan Klasifikasi

menampilkan laporan yang merangkum kinerja model yang lebih lengkap.

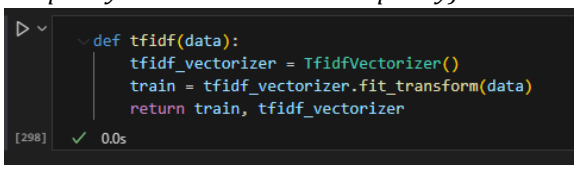


Gambar 6. Laporan Klasifikasi

Laporan ini memberikan informasi yang lebih lengkap dibandingkan dengan hanya menggunakan ***confusion matrix***, karena mengukur kinerja model pada setiap kelas secara terpisah dan memberi gambaran yang lebih jelas tentang kemampuan model.

TF – IDF

Mengubah data teks menjadi representasi numerik menggunakan teknik *TF-IDF (Term Frequency-Inverse Document Frequency)*



Gambar 7. TF – IDF

Normalisasi text: Menghapus kata-kata umum (*stop words*) yang tidak relevan dalam analisis bahasa dan menyamakan kata dengan kapitalisasi yang berbeda dan Membuat kosakata berdasarkan apa yang muncul dalam data latih. berarti hanya kata-kata yang ditemukan di data latih yang akan dipertimbangkan ketika mengubah teks menjadi fitur numerik,

Melatih Model XGBoost

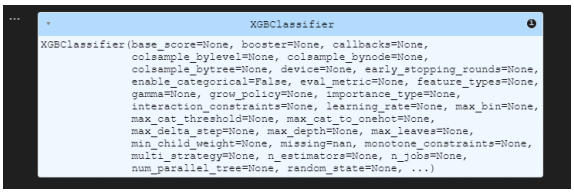
Melatih model *XGBoost* untuk melakukan klasifikasi pada data yang telah disiapkan.

def train_model() : Untuk membungkus langkah-langkah pelatihan model dalam sebuah fungsi dan pemanggilan Kembali

use_label_encoder=False **:** menghindari penggunaan *label encoder* bawaan XGBoost

XGBClassifier(): Membuat sebuah instance model klasifikasi berbasis *XGBoost*.

model.fit(X_train, Y_train): Melatih *XGBoost* menggunakan data pelatihan



Gambar 8. XGBClassifier

4.3. Evaluation

Pada tahapan *Evaluation* penulis melibatkan beberapa pengukuran performa yang didapatkan untuk memastikan akurasi, presisi, recall, dan matrik lainnya dalam mendeteksi pesan spam dan ham. Mengevaluasi kinerja model klasifikasi dengan memberikan gambaran rinci tentang prediksi yang benar (*True Negatives, True Positives,*) dan kesalahan (*False Negatives, False Positives,*).

Tabel 6. Skor

Akurasi	**0.9656652360515021**
Presisi	0.9682539682539683
Recall	0.9682539682539683
F1-Score	0.9682539682539683

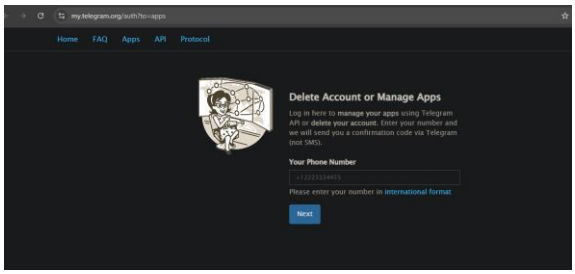
4.4. Deployment

Pada proses ini model yang telah di bangun akan diimplementasikan menggunakan metode xgboost ke dalam Aplikasi telegram. Membuat instance dari *TelegramClient* menggunakan library *Telethon* Tujuan utamanya adalah menginisialisasi klien Telegram, setelah itu Menginisialisasi stemmer yang digunakan untuk pemrosesan *NLP* ke dalam bahasa Indonesia. Langkah berikutnya adalah implementasi ke bot telegram. Implementasikan model yang sudah dilatih ke dalam chatbot Telegram. Gunakan API Telegram untuk menghubungkan chatbot dengan

Jurnal Manajemen Informatika & Sistem Informasi (MISI)
Volume 8, Nomor 2, Juni 2025.
ISSN 2614-1701 (media cetak)
ISSN 2614-3739 (media online)
DOI : 10.36595/misi.v5i2
http://e-journal.stmiklombok.ac.id/index.php/misi

model XGBoost sehingga pesan pengguna bisa diklasifikasikan secara real-time.

Untuk mendapatkan *API_ID* dan *API_HASH* kunjungi halaman resmi dari telegram *https://my.telegram.org/auth?to=apps* kemudian masukan nomor telepon dengan masukan format *+62.*



Gambar 9. Tampilan Utama

Setelah melakukan semua tahapan maka akan diarahkan ke APP configuration.
Dimana nantinya akan mendapatkan *App api_id*, dan *App api_hash*.

Gambar 10. Tampilan Configuration

Setelah mendapatkan *App api_id*, dan *App api_hash* buat file yang bernama *.env Environment Variable*

Gambar 11. Buat File .ENV

membuat file *.env* yang berfungsi untuk memuat *API_ID* dan *API_HASH* dari file .env yang dibuat tadi

```
11  # Load API_ID dan API_HASH dari file .env
12  load_dotenv()
13  API_ID = os.getenv("API_ID")
14  API_HASH = os.getenv("API_HASH")
```

Gambar 12. Load.ENV

4.5. Hasil Implementasi

Berikut adalah cara kerja bot yang mendeteksi adanya pesan baru yang masuk dari pengirim pesan dimana bot memiliki beberapa tindakan yang pertama adalah jika pesan spam yang dikirim terdeteksi sebagai spam sebanyak 1 kali maka tindakan bot hanya mengirimkan peringatan 1 kali kirim, Jika pengirim mengirimkan pesan spam yang sama maka bot akan mengirimkan peringatan ke dua untuk tidak mengirim pesan spam yang sama sebanyak 3 kali atau akun pengirim terblokir, jika pesan spam yang sama dikirim 3 kali maka akun pengirim akan terblokir otomatis oleh bot. Jika pengirim mengirimkan pesan spam berbeda maka bot akan mendeteksi ulang kurang dari 4 detik maka bot juga akan memberikan peringatan ulang 1 kali kirim

Tabel 7. Tindakan Bot

NO	Isi Pesan	Jumlah Pengiriman	Deteksi	Tindakan Bot
1	"Dapatkan Hadiah Gratis"	1	Spam	Peringatan 1 kali kirim
2	"Dapatkan Hadiah Gratis"	2	Spam	Peringatan ulang (2 kali Kirim)
3	"Dapatkan Hadiah Gratis"	3	Spam	Dihapus dan pengiri m diblokir.
4	"Ikuti Event Gratis"	1	Spam	Peringatan 1 kali Kirim
5	"Promo Spesial Dari XL"	1	Spam	Peringatan 1 kali Kirim
6	"Ikuti Event Gratis"	2	Spam	Peringatan ulang (2 kali Kirim)
7	"Promo Spesial Dari XL"	2	Spam	Peringatan ulang

Jurnal Manajemen Informatika & Sistem Informasi (MISI)
Volume 8, Nomor 2, Juni 2025.
ISSN 2614-1701 (media cetak)
ISSN 2614-3739 (media online)
DOI : 10.36595/misi.v5i2
http://e-journal.stmiklombok.ac.id/index.php/misi

				(2 kali Kirim)
8	"Ikuti Event Gratis"	3	Spam	Dihapus dan pengirim diblokir.
9	"Promo Spesial Dari XL"	3	Spam	Dihapus dan pengirim diblokir.
10	"Selamat Pagi"	1	Ham	Tidak ada

4.6. Analisis Hasil

Hasil analisa yang didapatkan dari pembangunan sistem implementasi chatbot pada telegram untuk mendeteksi spam dengan menggunakan metode XGBoost menunjukkan kinerja metode xgboost bagus untuk melakukan pendeteksian menangani text dengan memanfaatkan dataset yang didapatkan dari kaggle.

1. Analisis Kinerja Bot

1.1 Akurasi Deteksi

Model XGBoost mampu mengidentifikasi pesan spam dengan cukup akurat berdasarkan data pelatihan yang digunakan. Tidak ditemukan kasus false positive (pesan biasa terdeteksi spam) selama pengujian berlangsung dan respons bot terhadap spam cukup cepat kurang dari 4 detik.

1.2 Penanganan Spam

Bot mengelola spam dengan baik memberi peringatan sebelum bot melakukan pemblokiran yang bertujuan untuk menghindari pemblokiran secara langsung agar tidak salah sasaran.

Implementasi ini menunjukkan bahwa XGBoost dan telegram dapat digunakan untuk mendeteksi dan mengelola pesan secara realtime, meskipun berjalan dengan baik model ini belum mendukung deteksi spam yang menggunakan multi bahasa.

2. Ringkasan Hasil

Bot telegram menggunakan metode XGBoost berhasil menjalankan tugasnya mendeteksi, memperingatkan, dan memblokir secara otomatis, Bot berhasil mengidentifikasi pesan spam kurang dari 4 detik dan memblokir pengirim spam dengan akurasi tinggi dengan hasil yang sesuai dengan tujuan penelitian, bot juga memiliki tindakan yang bervariasi sesuai tingkat pelanggaran yang dilakukan oleh pengirim spam (peringatan - Blokir).

5. KESIMPULAN DAN SARAN

Berdasarkan hasil penelitian evaluasi kinerja metode XGBoost dalam implementasi chatbot di telegram yang diterapkan menunjukkan performa yang baik dengan matrik seperti akurasi, presisi, spesifisitas recall, dan F1-score yang melibatkan training data pengujian menjadi 70% - 30%, 80% - 20% dan 90% - 10%. Implementasi XGboost menghasilkan Accuracy 94%, Precision 95%, Recall 94% , Specificity 95%, F1-Score 94% pada data train dan testing 70% - 30%, pada data train dan testing 80% - 20% menghasilkan Accuracy 94%, Precision 95%, Specificity 95% Recall 93%, F1-Score 94%, dan pada data train dan testing 90% - 10% menghasilkan Accuracy 94%, Precision 95%, Recall 95%, Specificity 94% F1-Score 95%. Metode XGBoost terbukti efektif untuk mendeteksi pesan spam pada Telegram. Algoritma ini menangani data yang tidak seimbang dan memberikan akurasi tinggi, XGBoost menjadi pilihan yang tepat untuk tugas klasifikasi spam.

Penggunaan NLP, seperti tokenisasi, stemming, stopword removal, dan ekstraksi fitur seperti TF-IDF, membantu mengonversi data teks menjadi format numerik yang bisa dipakai oleh model XGBoost. Integrasi tahapan pemrosesan teks menggunakan Natural Language Processing (NLP) seperti *tokenization*, *stemming*, *stopword removal*, serta ekstraksi fitur menggunakan TF-IDF, terbukti membantu model dalam memahami dan mengklasifikasikan isi pesan dengan lebih akurat.

Penelitian ini memberikan kontribusi ilmiah melalui pembangunan sistem pendeteksi pesan spam secara otomatis yang terintegrasi dalam chatbot Telegram, dengan memanfaatkan algoritma XGBoost dan teknik Natural Language Processing (NLP) seperti tokenisasi, stemming, stopword removal, dan TF-IDF. Sistem yang dihasilkan mampu mengklasifikasikan pesan dengan akurasi tinggi dan dapat dijadikan sebagai referensi dalam pengembangan sistem keamanan komunikasi digital berbasis chatbot di berbagai platform.

Jurnal Manajemen Informatika & Sistem Informasi (MISI)
Volume 8, Nomor 2, Juni 2025.
ISSN 2614-1701 (media cetak)
ISSN 2614-3739 (media online)
DOI : 10.36595/misi.v5i2
http://e-journal.stmiklombok.ac.id/index.php/misi

Adapun saran Setelah melakukan penelitian, penulis menyadari adanya banyak kekurangan dan tantangan, oleh sebab itu peneliti merekomendasikan beberapa hal yang dapat diperbaiki lebih lanjut sebagai berikut:

5.1 Peningkatan Kualitas Data
Lakukan pengumpulan data yang lebih beragam untuk menangkap berbagai jenis spam, termasuk pesan multi-bahasa atau dengan pola teks yang kompleks , Perbarui dataset secara berkala untuk mengakomodasi pola baru yang digunakan oleh spammer, Kumpulkan data yang lebih beragam, termasuk pesan dalam berbagai bahasa dan pola spam yang kompleks dan Kumpulkan data yang lebih beragam, termasuk pesan dalam berbagai bahasa dan pola spam yang kompleks.

5.2 Peningkatan NLP
Gunakan model NLP yang lebih canggih seperti BERT atau GPT untuk meningkatkan pemahaman konteks dan nuansa pesan dan Tambahkan fitur analisis semantik untuk mendeteksi spam yang menggunakan sinonim atau frasa tidak langsung.

5.3 Optimasi Model XGBoost
Lakukan hyperparameter tuning secara menyeluruh untuk memastikan performa optimal model XGBoost dan Bandingkan dengan metode lain seperti Random Forest, Support Vector Machines (SVM), atau deep learning untuk melihat kemungkinan peningkatan akurasi.

5.4 Evaluasi Berkala
Lakukan evaluasi berkala terhadap performa model menggunakan data terbaru untuk memastikan sistem tetap relevan dan efektif.
Gunakan matrik evaluasi tambahan seperti Matthews Correlation Coefficient (MCC) untuk mengukur keseimbangan performa pada dataset tidak seimbang.

5.5 Penambahan Fitur
Kembangkan kemampuan chatbot untuk mendeteksi jenis konten lain yang mengganggu, seperti tautan berbahaya atau file berbahaya dan Sediakan opsi untuk mendeteksi spam dengan parameter khusus, seperti grub bahkan sampai wilayah.

DAFTAR PUSTAKA:

[1] F. Agustika, S. Siregar, D. Obara, D. Vipparamarta, and M. Manajemen, "Suatu Kajian Teori," vol. 9, no. 1, 2023, doi: 10.57249/jbk.

[2] M. A. Machmudi, "Peran Teknologi Informasi dalam Usaha Meraih Kesempatan Masa Depan Organisasi," 2019.

[3] M. Fajar Bahari, "Analisa Dan Implementasi Keamanan Pesan Chatting Menggunakan Algoritma Challenge Response," *Jurnal Sains Dan Teknologi Informasi*, vol. 1, no. 2, 2022, doi: 10.47065/jussi.v1i2.1442.

[4] R. Chandra and D. C. Juan, "Deteksi Pesan Spam pada Forum Daring Menggunakan Metode Naïve Bayes," vol. 7, no. 2, pp. 369–373, 2022, doi: 10.32493/informatika.v7i2.16411.

[5] kim fai kok, "Truecaller Insights: 20 Negara Teratas yang Terkena Panggilan & SMS Spam pada Tahun 2019," 3 Desember 2019.

[6] Ariyan Zubaidi and Ramdani, "LAYANAN DAN INFORMASI AKADEMIK BERBASIS BOTTELEGRAM DI PROGRAM STUDI TEKNIK INFORMATIKA UNIVERSITAS MATARAM," *LAYANAN DAN INFORMASI AKADEMIK BERBASIS BOTTELEGRAM DI PROGRAM STUDI TEKNIK INFORMATIKA UNIVERSITAS MATARAM*, vol. 1, no. 1, pp. 1–8, 2019.

[7] E. H. Yulianti, O. Soesanto, and Y. Sukmawaty, "Penerapan Metode Extreme Gradient Boosting (XGBOOST) pada Klasifikasi Nasabah Kartu Kredit," *JOMTA Journal of Mathematics: Theory and Applications*, vol. 4, no. 1, 2022.

[8] R. Wang, J. Zhang, B. Shan, M. He, and J. Xu, "XGBoost Machine Learning Algorithm for Prediction of Outcome in Aneurysmal Subarachnoid Hemorrhage," *Neuropsychiatric Disease and Treatment*, vol. 18, pp. 659–667, 2022, doi: 10.2147/NDT.S349956.

[9] Ni Luh Putu Kintania Marsha D., "Viral Pembobolan M-Banking Akibat Link Phishing: Uang Rp 16,4 Juta Raib ," 25 Mei 2022.

[10] I. Magdalena, A. Salsabila, D. A. Krianasari, and S. F. Apsarini, "IMPLEMENTASI MODEL PEMBELAJARAN DARING PADA MASA PANDEMI COVID-19 DI KELAS III SDN SINDANGSARI III," 2021. [Online]. Available: https://ejournal.stitpn.ac.id/index.php/pandawa

[11] A. A. Chandra, V. Nathaniel, F. R. Satura, F. Dharma Adhinata, and P. Studi, "Pengembangan Chatbot Informasi Mahasiswa Berbasis Telegram dengan Metode Natural Language Processing."

[12] M. Riziq sirfatullah Alfarizi, M. Zidan Al-farish, M. Taufiqurrahman, G. Ardiansah, and M. Elgar, "PENGGUNAAN PYTHON SEBAGAI BAHASA PEMROGRAMAN UNTUK MACHINE LEARNING DAN DEEP LEARNING," 2023.

[13] M. Reynaldi, S. al Khairi, N. Geraldi Hendarman, and F. I. Nugroho, "Sistem Informasi Berbasis Bot Telegram Sebagai Media Sosialisasi Keselamatan Berkendara," 2020.

[14] I. Huda, "IMPLEMENTASI NATURAL LANGUAGE PROCESSING (NLP) UNTUK APLIKASI PENCARIAN LOKASI."

[15] J. Aprian Putra dan and A. Budi, "PENERAPAN NATURAL LANGUAGE PROCESSING DALAM APLIKASI CHATBOT SEBAGAI MEDIA PENCARIAN INFORMASI DENGAN MENGGUNAKAN REACT (STUDI KASUS: INSTITUT BISNIS DAN INFORMATIKA KWIK KIAN GIE)".

[16] D. Septiani and I. Isabela, "SINTESIA: Jurnal Sistem dan Teknologi Informasi Indonesia ANALISIS TERM FREQUENCY INVERSE DOCUMENT FREQUENCY (TF-IDF) DALAM TEMU KEMBALI INFORMASI PADA DOKUMEN TEKS," no. Vol. 1 No. 2 (2022): SINTESIA, Oct. 2023.

[17] M. A. Hasanah, S. Soim, and A. S. Handayani, "Implementasi CRISP-DM Model Menggunakan Metode Decision Tree dengan Algoritma CART untuk Prediksi Curah Hujan Berpotensi Banjir," 2021. [Online]. Available: http://jurnal.polibatam.ac.id/index.php/JAIC