

AI-Powered Spam Detection Using Machine Learning

Gongu Navya Sri

Department of Computer

Science and Engineering

Vardhaman College of Engineering

Hyderabad-Telangana, 501218, India

navyasrigongu@gmail.com

Gunti Rahul

Department of Computer

Science and Engineering

Vardhaman College of Engineering

Hyderabad-Telangana, 501218, India

rahulgunti2006@gmail.com

M. Shantesh

Department of Computer

Science and Engineering

Vardhaman College of Engineering

Hyderabad-Telangana, 501218, India

Shanteshyadav055@gmail.com

Abstract: With the rise of online communication, spam messages and emails have become a big problem. They waste our time and sometimes include harmful links, scams, or unwanted advertisements. This project is about building an AI-based spam detection system that can automatically identify and filter spam messages.

The system is trained using examples of spam and normal messages. It uses machine learning algorithms like Naive Bayes, SVM, and Decision Trees to recognize patterns in the messages and decide whether a message is spam or safe. Over time, it can learn from new messages and adapt to new types of spam.

Using this system, people can save time, avoid threats, and keep their inbox or messaging apps clean. It can be used in email services, messaging apps, or social media platforms. The system can also give insights about spam trends, helping organizations prevent spam more effectively.

Key Words: Artificial Intelligence, Machine Learning, Spam Detection, Email Spam Filtering, Text Classification, Automated Spam Detection System

I. INTRODUCTION

The rapid rise of online communications has resulted in an unprecedented level of connectivity, while significantly increasing the volume of unwanted communications. With increased use of email and direct messaging, there has been a corresponding rise in spam, and researchers continue to report that spam messages comprise a large percentage of email traffic. Sometimes referred to as "junk mail," spam content—commonly used for advertising, phishing, or other malicious purposes—poses very real risks to security that go beyond annoyance. Spam content enables fraudulent campaigns (e.g. phishing and identity theft), increases distrust in online systems, and fills user's inboxes and networks. In fact, research suggests that more than 50% of email is spam and can not only waste users' time but also put their devices at risk from malware and data theft. Overall, this lull in the existence of spam is a clear threat to productivity, privacy, and network security.

In response to the growing threat, Artificial Intelligence (AI), specifically machine learning (ML) and natural language

processing (NLP), has become a critical solutions approach in cyber security. Several modern approaches to AI learn autonomously from very large corpora of messages to identify the subtle differences that distinguish legitimate content from spam. In practice, AI spam filters utilize ML models (Naive Bayes, SVM, or neural networks) along with NLP feature extraction to establish very high levels of accuracy in the classification process. For example, as opposed to providing a classified list of known spam content or known spam sources, an AI system automatically detects if a piece of content is spam by crawling the web and crawling social media in addition to looking at email streams to find and highlight suspicious links or repeated keywords. These more advanced approaches factor into addressing real-time detection: One research reported an NLT + Deep-learning phishing filter identified 97.5% of phishing attacks, which is higher than traditional frameworks based on rules or even simple ML models. Spam filters powered by artificial intelligence (AI) are now commonplace on a variety of platforms. In email services, social networks, e-commerce sites, and instant-messaging applications, AI-based spam filters detect and eliminate spam or phishing emails before the consumer receives them. These tools help improve data privacy and security, thereby allowing users and businesses to communicate more securely, by preventing malicious and irrelevant content from reaching everyone. Since spammers constantly evolve their tactics, having artificial intelligence (AI) that can adapt is essential. Modern spam detection solutions include continuous learning and updating so that new forms of spam can be identified in real time. In

practice, these spam filters operate with low latency, and can easily scale to fit the needs of small organizations or international enterprises: one hybrid model, for example, allows for automatic updates when new threats emerge, and is simultaneously low-cost, and learns and scales to any sized organization. In conclusion, the demands of today's spam environment requires automated, scalable, and intelligent responses – and this is where AI, machine learning (ML), and natural language processing (NLP) are uniquely suited.

II. LITERATURE REVIEW

Early spam detection systems heavily relied on blacklists, whitelists, and rule-based methodologies. While these methods provided some degree of protection, they came with significant limitations. Rule-based systems required frequent manual updates to keep up with evolving spam tactics, leading to a high maintenance burden. Blacklists only blocked known spammers, while whitelists restricted communication to predefined senders, ultimately hindering broader communication. These approaches were predominantly reactive and lacked the adaptability to handle innovative tactics employed by spammers determined to bypass filtering technologies.

The advent of machine learning brought a transformative shift to spam detection, enabling more adaptive and efficient solutions. Fundamental techniques like Naive Bayes and Support Vector Machines (SVMs) have been widely adopted due to their proven efficacy in text classification tasks. Naive Bayes, being a probabilistic approach, is both simple and effective, providing a strong baseline for spam detection when combined with features like

bag-of-words or TF-IDF. On the other hand, SVMs excel by identifying optimal hyperplanes to separate data points in high-dimensional spaces, offering robustness against overfitting and exceptional performance with textual data. Significant research has been dedicated to applying artificial intelligence (AI) and machine learning (ML) models to address spam detection challenges. For instance, Odeh and Al Hattab (2023) conducted an extensive review of AI applications in social systems for spam detection, highlighting their growing prominence. Similarly, studies by Anuja et al. (2024) and Goswami et al. (2024) underscore the broad scope of utilizing AI and ML techniques for combating online spam. Collectively, this body of research demonstrates that machine learning has become an essential tool in the fight against spam.

Natural Language Processing (NLP) plays a pivotal role in transforming raw textual data into numerical features interpretable by machine learning models while preserving the nuances of human language. Without effective NLP methodologies, the subtle differences between legitimate messages and spam would be difficult to capture. Key feature extraction methods include:

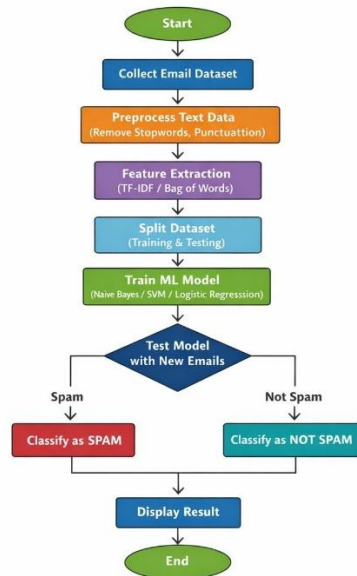
I. TF-IDF (Term Frequency-Inverse Document Frequency): This statistical technique measures a word's relevance in a document based on its frequency within that document relative to its occurrence across an entire document set. Higher weights are assigned to words that appear frequently in a single text but rarely across the collection, effectively isolating common indicators of spam.

II. Word Embeddings: Advanced methods like Word2Vec or Glove transform words into dense, low-dimensional vectors within continuous vector spaces. These representations capture semantic relationships and contextual information between words, enabling models to consider meaning beyond simple word presence. Similar words tend to have comparable vector representations, allowing models to effectively generalize to new spam variations. Kotevski (2025) explores this concept further in a "spam detection pipeline using AI and NLP," illustrating a systematic workflow from raw text processing to classified outputs.

In recent years, spam detection has shifted towards leveraging deep learning architectures that excel at identifying complex patterns and hierarchical structures within textual data. Models such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), including Long Short-Term Memory (LSTM) networks, can directly learn intricate representations from raw text or word embeddings, often surpassing traditional machine learning methods when applied to large, complex datasets. Research on deep learning applied to adjacent cybersecurity areas, such as phishing detection, has shown promising outcomes (Dey, 2023; Lamina et al., 2024; Enitan, 2023). Given that phishing represents a specific type of malicious spam, these approaches are highly relevant for general spam detection. Additionally, hybrid models combining various AI approaches have been explored to capitalize on the strengths of different techniques. For example, Douse et al. (2020) studied how integrating machine learning with deep learning enhances spam detection

capabilities. This continuous progress reflects the dynamic nature of innovations in the field.

III. FLOWCHART



IV. PROPOSED SOLUTION

To improve the proposed AI-powered spam detection system, Edge AI is used to allow spam detection directly on the user's device. Instead of sending emails or messages to cloud servers for checking, the trained machine learning model runs locally on the user's mobile or desktop. With this setup, spam is detected right away, even when the device is not connected to the internet. Since no user data leaves the device, privacy is greatly increased, making the system more secure and easier to use. This feature helps the spam detection system work faster, reduces the need for an internet

connection, and keeps sensitive messages private. It is especially helpful for mobile users who need quick spam filtering without delays from the network.

Future versions can also integrate federated learning, where multiple devices collaboratively improve the model without sharing raw user data. This ensures better accuracy while maintaining strong privacy protection.

Additionally, the system can be expanded beyond email to detect spam in SMS, social media messages, and instant messaging platforms, making it a unified on-device spam protection solution. With continuous improvements in Edge AI frameworks, the proposed system can become more efficient, energy-aware, and widely applicable in real-world communication systems.

Benefits:

- Faster spam detection with low delay
- Works even without an internet connection
- Better privacy and security for user data.

V. Code

app.py

```
import streamlit as st

from model import predict_spam

st.title("AI Spam Detector")

msg = st.text_area("Enter Message")

if st.button("Check"):

    st.write(predict_spam(msg))
```

```
#streamlit run app.py
```

Model.py

```
import pandas as pd

import nltk

import re

from nltk.corpus import stopwords

from sklearn.feature_extraction.text import
TfidfVectorizer

from sklearn.model_selection import
train_test_split

from sklearn.linear_model import
LogisticRegression

from sklearn.preprocessing import
LabelEncoder

from sklearn.metrics import
accuracy_score, classification_report

# Download stopwords (only first time)

nltk.download('stopwords')

# Load dataset

data = pd.read_csv("SpamN.csv",
encoding="latin-1")[['v1', 'v2']]

data.columns = ['label', 'message']

# Text cleaning

def clean_text(text):

    text = text.lower()

    text = re.sub(r'^[a-z]', '', text)

    words = text.split()

    words = [w for w in words if w not in
stopwords.words('english')]

    return " ".join(words)
```

```
data['clean_message'] =
data['message'].apply(clean_text)

# Feature extraction

tfidf =
TfidfVectorizer(max_features=3000)

X =
tfidf.fit_transform(data['clean_message']).t
oarray()

# Encode labels

encoder = LabelEncoder()

y = encoder.fit_transform(data['label'])

# Train-test split

X_train, X_test, y_train, y_test =
train_test_split(

    X, y, test_size=0.2, random_state=42
)

# Train model

# model = MultinomialNB()

# model.fit(X_train, y_train)

model =
LogisticRegression(max_iter=1000)

model.fit(X_train, y_train)

# Improve Spam Recall (Advanced)

# Try Logistic Regression:

# from sklearn.linear_model import
LogisticRegression

# model =
LogisticRegression(max_iter=1000)

# model.fit(X_train, y_train)

# Evaluate

y_pred = model.predict(X_test)
```

```

print("Accuracy:", accuracy_score(y_test,
y_pred))

print(classification_report(y_test, y_pred))

# Prediction function

def predict_spam(text):

    text_lower = text.lower()

    spam_keywords = [

        'free', 'win', 'winner', 'prize', 'lottery',
        'jackpot',

        'cash', 'reward', 'bonus', 'giveaway',
        'claim', 'money',

        'urgent', 'immediately', 'limited time',
        'hurry',

        'last chance', 'today only', 'expires',

        'click', 'tap', 'open', 'link', 'subscribe',

        'download', 'verify', 'confirm',

        'account', 'bank', 'debit', 'credit', 'card',
        'pin', 'otp',

        'suspended', 'blocked', 'security alert',

        'offer', 'discount', 'deal', 'promotion',
        'special offer',

        'exclusive', 'trial', 'free trial',

        'guaranteed', 'risk-free', 'congratulations',
        'selected'

    ]

    count = sum(1 for word in
spam_keywords if word in text_lower)

    if count >= 2:

        return "Spam "

    cleaned = clean_text(text)

    vector =
tfidf.transform([cleaned]).toarray()

```

```

    result = model.predict(vector)

    return "Spam " if result[0] == 1 else
"Not Spam "

if __name__ == "__main__":

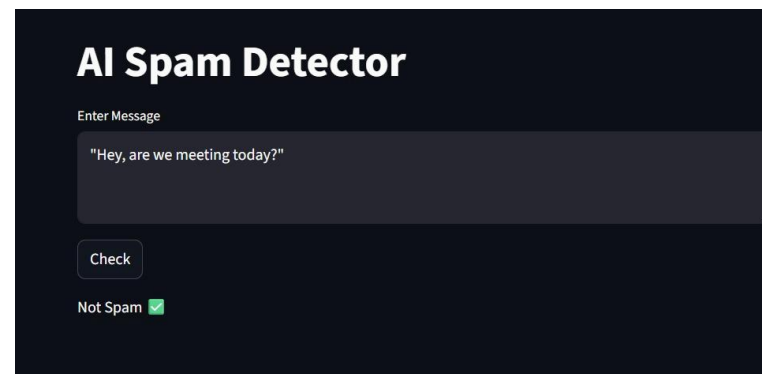
    # Test manually

    #print(predict_spam("Congratulations!
You won a free ticket"))

    print(predict_spam("Hey, are we
meeting today?"))

```

VI. Result



VII. CONCLUSION

This project successfully presents an intelligent spam detection system that helps identify unwanted messages in an effective and reliable manner. By using machine learning techniques, the system is able to analyze message content and accurately classify emails as spam or non-spam, reducing manual effort and improving user experience.

The inclusion of Edge AI makes the system more practical by allowing spam detection directly on user devices. This approach ensures faster processing, works even without internet connectivity, and protects user privacy by keeping personal data on

the device itself. The use of multiple models and feedback-based learning further improves accuracy and adaptability.

Overall, the proposed solution offers a secure, efficient, and user-friendly method for spam detection. With future enhancements, this system can be extended to other communication platforms, making it a valuable tool for safe and trustworthy digital communication.

VIII. FUTURE SCOPE

The spam detection system developed in this project can be improved and expanded in several ways in the future. As communication platforms continue to grow, the system can be extended to work not only with emails but also with SMS, messaging apps, and social media platforms, providing broader protection against unwanted messages.

Future versions of the system can be made more personalized by learning user preferences, allowing important messages to be identified more accurately while reducing false spam alerts. The model can also be updated regularly to handle new and evolving spam techniques, ensuring long-term effectiveness. With the advancement of device hardware, the system can support faster processing while consuming less power, making it suitable for continuous use on mobile devices. Additional language support can also be introduced so that spam messages in multiple regional and international languages can be detected effectively.

Overall, the project has strong potential for real-world application, and with further enhancements, it can become a reliable

solution for secure and efficient digital communication.

IX. REFERENCES

1. Odeh, A. H., & Al Hattab, M. (2023, November). AI Methods Used for Spam Detection in Social Systems- An Overview. In 2023 Tenth International Conference on Social Networks Analysis, Management and Security (SNAMS) (pp. 1-8). IEEE.
2. Anuja, K., Dnyaneshwari, C., Sharda, M., Mansi, S., & Rina, S. (2024). Spam Spyder (Spam Detection using MI & AI). International Journal of Trend in Scientific Research and Development, 8(5), 999-1007.
3. Dey, S. (2023). AI-powered phishing detection: Integrating natural language processing and deep learning for email security.
4. Lamina, O. A., Ayuba, W. A., Adebisi, O. E., Michael, G. E., Samuel, O. O. D., & Samuel, K. O. (2024). Ai-Powered Phishing Detection and Prevention. Path of Science, 10(12), 4001-4010. Appendices
5. Goswami, A., Patel, R., Mavani, C., & Mistry, H. K. (2024). Identifying Online Spam Using Artificial Intelligence. International Journal on Recent and Innovation Trends in Computing and Communication, 12(2), 548-55.
6. Alqarni, A. (2025). How Generative AI Transforms Spam Detection. In Tech Fusion in Business and Society (pp. 3-11). Springer, Cham.

7. Kasa, A. S. The Power of AI in Detecting Spam Emails.
8. Douzi, S., AlShahwan, F. A., Lemoudden, M., & El Ouahidi, B. (2020). Hybrid email spam detection model using artificial intelligence. *International Journal of Machine Learning and Computing*, 10(2).
9. Enitan, O. I. (2023). An AI-Powered Approach to Real-Time Phishing Detection for Cybersecurity. *International Journal*, 12(6).
10. Kotevski, A. (2025). Spam detection pipeline using AI and NLP. Preface to Volume 5 Issue 1 of the *Journal of University of Information Science and Technology "St. Paul the Apostle"*– Ohrid, 5(1), 16.