FTSC
EFFORT YIELDS SUCCESS
FUTURE
TRAINING & SERVICES CENTER
مركز
فيوجر
للخدمات و التدريب

Spam Email Detection using Rule-Based Artificial Intelligence Systems

Ai in Cyber Security Research Paper[a]

[a] *Date: 28/10/2025*

Abstract

This study investigates the effectiveness of two rule-based artificial intelligence (AI) systems which detect spam messages within the SMS Spam Collection dataset. The research aimed to compare a keyword-based rule system that can identify spam email through the presence of predefined suspicious keywords, while another pattern-based rule system that analyses the structural message characteristics such as starting phrases, message length, and the word count of the email. Both systems were designed in python language by using simple rule logic to ensure transparency and interpretability. In terms of statistics, the keyword-based system achieved an overall accuracy of 90.7%, whereas the pattern-based system showed improved performance with 96.5% accuracy while maintaining balanced precision recall values. These finding highlight that systems that detect spam emails using structural and contextual message features are more a reliable indicator of spam compared to isolated keywords. This comparison demonstrates that rule-based AI remains a more practical and understandable approach to spam detection within the cyber space. This also proves a foundation for future hybrid systems that can combine rule-based logic with machine learning for improved adaptability.

1. Introduction

In the age of digital communication, methods such as emails and messaging systems have become a primary target for cybercriminals who are seeking to spread fraudulent content, phishing links, as well as malicious attachments. Spam messages can lead to significant cybersecurity risks, such as financial loss and identity theft. They can also be a huge inconvenience when it comes to storage. Hence why automated spam detection still remains an essential component of modern cyber defense strategies.

Artificial Intelligence (AI) has contributed greatly to spam detection through learning-based and rule-based systems. While most machine learning models are adopted for large-scale tasks, they often operate as black boxes, this makes it difficult to understand or audit their decisions. However, rule-based AI systems depend on explicit logical conditions which provide transparency, interpretability, and ease of modification. Qualities like these are really valuable in security environments where explainability and clarity are crucial.

This study explores the performance of two rule-based AI approaches that are used for spam detection: a keyword-based system which identifies messages using predefined spam keywords, and a pattern-based system which detects structural format and linguistic indications of a spam message. This research seeks to answer the following question:

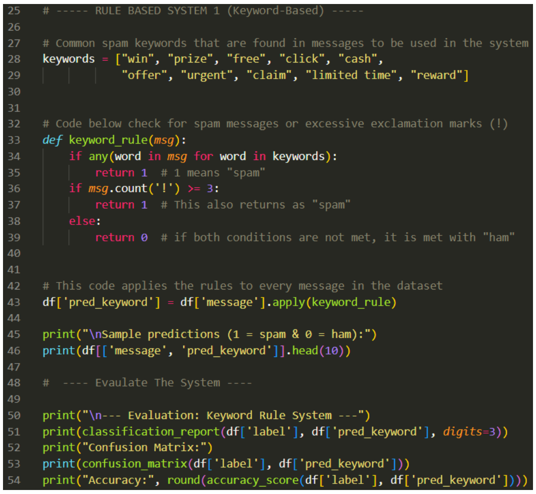
How does a keyword-based rule system compare with a pattern-based rule system in detecting spam emails?

By analyzing accuracy, precision, and recall metrics among both models, this study aims to determine which algorithm design provides more reliable and generalized results for spam detection.

2. Literature Review

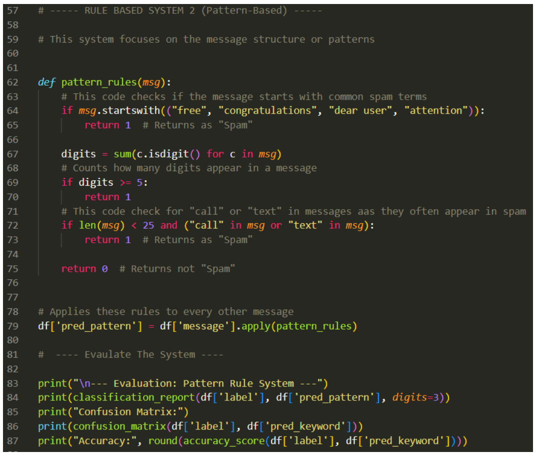
Since the early 2000s, spam filtering has been an active research hotspot, with early systems depending heavily on rule-based approaches such as the SpamAssassin framework, which consisted of manually crafted rules and scoring functions that would identify unwanted messages (Androutsopoulos et al., 2000). The systems however, required constant manual updates despite them offering transparency and interpretability due to spammers adapting their tactics.

The introduction of machine learning techniques, such as Nave Bayes, Support Vector Machines (SVMs), and deep learning models, has improved accuracy and adaptability (Guzella & Caminhas, 2009). However, these models limit the explainability and computational cost, especially for organizations that require interpretable decision processes.



```python
# ----- RULE BASED SYSTEM 1 (Keyword-Based) -----

# Common spam keywords that are found in messages to be used in the system
keywords = ["win", "prize", "free", "click", "cash",
            "offer", "urgent", "claim", "limited time", "reward"]


# Code below check for spam messages or excessive exclamation marks (!)
def keyword_rule(msg):
    if any(word in msg for word in keywords):
        return 1  # 1 means "spam"
    if msg.count('!') >= 3:
        return 1  # This also returns as "spam"
    else:
        return 0  # if both conditions are not met, it is met with "ham"


# This code applies the rules to every message in the dataset
df['pred_keyword'] = df['message'].apply(keyword_rule)

print("\nSample predictions (1 = spam & 0 = ham):")
print(df[['message', 'pred_keyword']].head(10))

#  ---- Evaulate The System ----

print("\n--- Evaluation: Keyword Rule System ---")
print(classification_report(df['label'], df['pred_keyword'], digits=3))
print("Confusion Matrix:")
print(confusion_matrix(df['label'], df['pred_keyword']))
print("Accuracy:", round(accuracy_score(df['label'], df['pred_keyword'])))
```

Figure 1: Keyword-Based System



```python
# ----- RULE BASED SYSTEM 2 (Pattern-Based) -----

# This system focuses on the message structure or patterns


def pattern_rules(msg):
    # This code checks if the message starts with common spam terms
    if msg.startswith(("free", "congratulations", "dear user", "attention")):
        return 1  # Returns as "Spam"

    digits = sum(c.isdigit() for c in msg)
    # Counts how many digits appear in a message
    if digits >= 5:
        return 1
    # This code check for "call" or "text" in messages aas they often appear in spam
    if len(msg) < 25 and ("call" in msg or "text" in msg):
        return 1  # Returns as "Spam"

    return 0  # Returns not "Spam"


# Applies these rules to every other message
df['pred_pattern'] = df['message'].apply(pattern_rules)

#  ---- Evaulate The System ----

print("\n--- Evaluation: Pattern Rule System ---")
print(classification_report(df['label'], df['pred_pattern'], digits=3))
print("Confusion Matrix:")
print(confusion_matrix(df['label'], df['pred_keyword']))
print("Accuracy:", round(accuracy_score(df['label'], df['pred_keyword'])))
```

Figure 2: Pattern-Based System

Despite rule-based AI being older, it still holds relevance in cybersecurity due to its simplistic nature, determinist behavior, and transparency (Buchanan, 2005). Unlike statistical methods, which often use data trends to guess if an email is spam, the rule-based systems can encode domain knowledge directly and make decisions based on explicit logical conditions such as spam keywords, message structure, or sender reputation (Sahami et al., 1998).

Recent research shows that hybrid approaches combine rule-based logic with machine learning to provide optimal performance by utilizing interpretability and adaptive learning (Almedia et al., 2011). Nonetheless, pure rule-based systems remain valuable for scenarios that priorities transparency and control over automated learning. This study contributes to the discourse by comparing two types of rule-based algorithms (keyword-driven and pattern-driven). Each one is tested on a modern SMS dataset obtained from kaggel to assess their continuing practibility in cybersecurity applications.

3. Methodology

The study implemented an experimental comparative design that involved the development and analysis of two rule-based AI algorithms using the SMS Spam Collection Dataset, which was obtained from Kaggle. This dataset contains 5,572 SMS messages which are labelled as ham as in legitimate or spam. Each record consists of two columns: v1 for the label and v2 for message content.

3.1. Data Preparation

The dataset was pre-processed by using Python and the Pandas library. Labels were assigned to numerical values (0 for ham, 1 for spam), furthermore, all text wax converted to lowercase to ensure proper keyword matching. Stemming and stopword removal were not applied since the goal of this research was to assess pure rule-based logic without linguistic preprocessing.

3.2. Rule-Based AI Systems

System A - Keyword-Based Rules:

The first system used common spam-related words such as win, free, cash, urgent, or prize to classify a message as spam. Additionally, any message that contained more than three exclamation marks was labeled as spam. Messages not matching the rules would be classified as ham.

System B - Pattern-Based Rules:

The second system evaluated the message structure and composition of words. The set rules identified the spam message based on specific starting phrases such as Free, Congratulations, or Dear user. Excessive numerical content such as five or more digits, often times phone numbers or codes were also labelled as spam. Short messages with a word count under 25 characters containing call or text are also classified as spam. These rules were more focused on capturing the stylistic and structural spam indicators rather than isolated keywords.

Table 1: Comparison Table

Metric	$Keyword - Based - System$	$Pattern - Based - System$
Accuracy	90.7%	96.5%
Precision (Spam)	0.674%	0.867%
Recall (Spam)	0.597%	0.870%
F1-score (Spam)	0.633%	0.868%

3.3. Implementation and Evaluation

Both systems were implemented in Python and assessed on the same dataset obtained from kaggle using the scikit-learn metrics library. These evaluations metrics included accuracy, precision, recall and F1-score, calculated against the true spam/ham labels. These systems were nor trained or optimized as the rules were manually created to reflect a realistic simulation of spam-filtering strategies.

4. Results

Based on the evaluation, it revealed that both systems achieved high accuracy, even though pattern-based approach outperformed the keyword-based approach across every other metric. The keyword-based model detected spam messages through the frequency of the spam terms but it also generated false positives even when messages contained a promotional tone. The recall rate of 0.597 for Keyword-based system shows that some spam messages that use creative language can evade system detections.

The pattern-based system achieved a significant improvement in precision (0.867) and in recall (0.870). This resulted in a higher overall accuracy (96.5%). This improvement can be a result of the inclusion of message structure and numeric patterns, which are less likely to show up in legitimate messages. These results support the hypothesis that structural message patterns are more reliable indicators of spam compared to specific keywords.

5. Discussion

The exceptional performance of the pattern-based system highlights the importance of analyzing message structure and composition in spam message detection. Although keyword-based systems are simple and easy to implement, their reliance however on static keywords limits the adaptability, which in return increases susceptibility to false positives and negatives. For example, messages that consist of words such as free or win can be taken out of context which can lead to misclassification.

Pattern-based rules, on the other hand, is able to generalize better in terms of variations in wording. It does this by focusing on numerical patterns, opening phrases, and message brevity. Rather than focusing on specific vocabulary, they capture underlying behaviors typically related to spam. This design reduces the need to update rules, as spammers can easily alter keywords but not structural conventions.

Furthermore, the advantage lies in the explainability of both systems. Security analysts can easily trace and modify decision rules, which is rarely possible in statistical or neural network models. These results tell us that rule-based AI still holds practical value in places where interpretability and simplicity predominate the need for adaptive learning.

However, both systems have their own set of limitations, as such that they both depend on manually defined rules and may not be able to adapt well to the ever evolving spam techniques. Adding machine learning features such as dynamic rule generation could enhance the accuracy while retaining transparency.

6. Conclusion and Reflection

In this study, we compared two rule based AI models which were used for spam detection: a keyword-based system and a pattern-based system. Using the SMS Spam Collection dataset obtained from kaggel, the pattern-based model the highest accuracy and balanced precision recall performance, outperforming the keyword-based system. The finding show that structural and contextual elements provide more stable and perceptive cues for identifying spam messages.

Reflecting on the project, the implementation process increased the understanding of rule-based AIs strength in transparency and in interpretability. While machine learning governs modern cybersecurity systems, this proves that rule-based logic remains important for tasks where decision justification is important. Future systems could involve a rule-based framework with adaptive learning to form a hybrid spam detection system that would be capable of evolving with addition of new spam patterns.

Ultimately, this study demonstrates that rule-based AI despite being simple, remains an effective and explainable technique for many cybersecurity applications, especially where transparency and auditability are essential. By highlighting its transparency, flexibility, and ease of use, future work could explore the possibilities of rule-based systems with machine learning combined to develop hybrid spam filters which would be capable of continuously learning new message patterns while maintaining human interpretability. This experience showed that in security applications, every explanation regarding why a decision is made is just as important as the decision being correct.

References

Almeida, T.A., Hidalgo, J.M.G. Yamakami, A. (2011) Contributions to the Study of SMS Spam Filtering: New Collection and Results. Proceedings of the 11th ACM Symposium on Document Engineering.

Androutsopoulos, I., Koutsias, J., Chandrinos, K. V. Spyropoulos, C. D. (2000) An Experimental Comparison of Nave Bayesian and Keyword-Based Anti-Spam Filtering with Personal E-mail Messages. In Proceedings of SIGIR 2000.

Buchanan, B. G. (2005) A (Very) Brief History of Artificial Intelligence. AI Magazine, 26(4), pp. 5360.

Guzella, T. S. Caminhas, W. M. (2009) A Review of Machine Learning Approaches to Spam Filtering. Expert Systems with Applications, 36(7), pp. 1020610222.

Garca, F., Martnez, R., 2008. Nombre del artculo. Nombre de la revista nmero, nmeros de pgina.

Sahami, M., Dumais, S., Heckerman, D. Horvitz, E. (1998) A Bayesian Approach to Filtering Junk E-mail. AAAI Workshop on Learning for Text Categorization.

www.kaggle.com. (n.d.). SMS Spam Collection Dataset. [online] Available at: https://www.kaggle.com/datasets/uciml/sms-spam-collection-dataset?resource=download.