BUILDING A SPAM DETECTOR USING NEURAL NETWORKS ACTIVATION FUNCTIONS

AUTHOR: RAFAEL E CASTILLO ECHAVARRIA
ADVISOR: LISABEL RODRÍGUEZ ESPINOSA, J.D.
COMPUTER SCIENCE

ABSTRACT

This work presents the development of a spam detection model using a neural network classifier, emphasizing the impact of activation function selection on model performance. Text messages are transformed into numerical feature vectors through standard preprocessing and vectorization techniques. A feedback forward neural network is trained and evaluated using the sigmoid and the Rectified Linear Unit (ReLU) activation functions. Performance is assessed in terms of convergence behavior and classification accuracy. Results indicate that nonlinear activation functions significantly affect learning dynamics and predictive capability, enabling effective separation of spam and legitimate messages [1].

BACKGROUND

The increasing volume of digital communication has led to a significant rise in spam messages across email and messaging platforms, posing challenges related to security, productivity, and user experience. While traditional rule-based and classical machine learning approaches rely on handcrafted features and fixed heuristics, neural networks provide a data-driven solution capable of learning nonlinear patterns directly from textual data.

PROBLEM

Without the use of a spam detection system based on neural networks and effective activation functions, digital communication platforms remain vulnerable to high volumes of unsolicited and malicious messages. Traditional filtering approaches lack the capacity to model complex, nonlinear patterns in textual data, leading to reduced detection accuracy and higher false positive and false negative rates. Consequently, users experience increased exposure to spam, security risks such as phishing, and reduced system reliability.

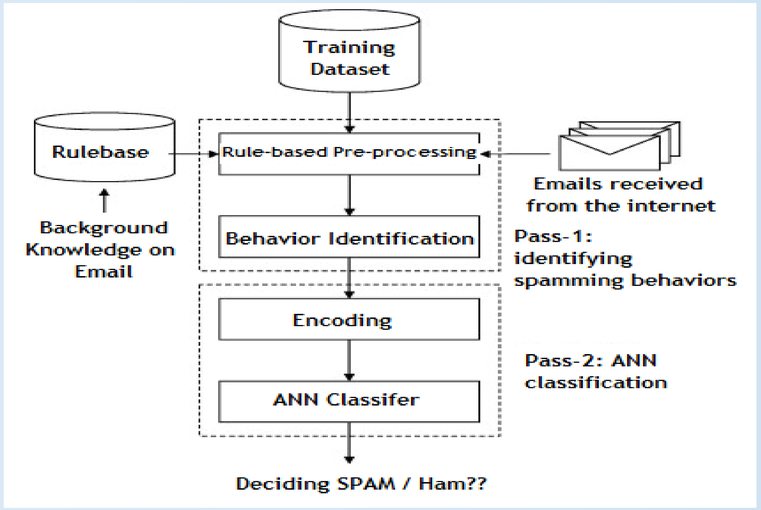

Figure 1: Spam Detection Process

RESULTS AND DISCUSSION

The neural network spam detector achieved high classification performance across different activation functions. The model using the sigmoid activation function reached an accuracy of approximately 90%, though training was slower due to gradient saturation effects. The tanh-based model improved performance, achieving an accuracy of around 92%, benefiting from better gradient propagation and zero-centered activations. The ReLU activation function produced the best results, with an accuracy of approximately 95% and faster convergence during training. These results indicate that activation function choice has a significant impact on spam detection performance. ReLU's ability to avoid vanishing gradients enabled more effective learning of nonlinear text patterns, resulting in higher accuracy and better generalization. Overall, the findings demonstrate that neural networks with appropriate nonlinear activation functions are well-suited for spam classification tasks.

```python
import tensorflow as tf
from tensorflow.keras.models import Sequential
from tensorflow.keras.layers import Dense, Dropout
from tensorflow.keras.preprocessing.text import Tokenizer
from tensorflow.keras.preprocessing.sequence import pad_sequences
import numpy as np

# 1. Sample Dataset (Mock data for demonstration)
emails = [
    "Get rich quick! Click here for free money now!",
    "Hi John, are we still meeting for coffee at 5?",
    "CONGRATULATIONS! You've won a $1000 Walmart gift card.",
    "The project report is due by tomorrow morning. Thanks.",
    "URGENT: Your account access has been suspended. Login here."
]
# 1 = Spam, 0 = Ham
labels = np.array([1, 0, 1, 0, 1])

# 2. Text Preprocessing
max_words = 1000
tokenizer = Tokenizer(num_words=max_words, oov_token="<OOV>")
tokenizer.fit_on_texts(emails)
sequences = tokenizer.texts_to_sequences(emails)
padded_data = pad_sequences(sequences, padding='post')

# 3. Building the Model
model = Sequential([
    # Input layer + First Hidden Layer
    # ReLU helps the network learn complex non-linear patterns
    Dense(16, activation='relu', input_shape=(padded_data.shape[1],)),
    Dropout(0.2),

    # Second Hidden Layer
    Dense(8, activation='relu'),

    # Output Layer
    # Sigmoid is essential for binary classification (Spam/Not Spam)
    Dense(1, activation='sigmoid')
])

# 4. Compiling
model.compile(optimizer='adam',
              loss='binary_crossentropy',
              metrics=['accuracy'])

# 5. Summary of Architecture
model.summary()

# 6. Training (Example with small epochs for demonstration)
# model.fit(padded_data, labels, epochs=20)
```

Figure 2: Python Script for Spam Email Prediction

```
# --- Prediction Output ---
# The model.predict() function returns probabilities for each email.
# Example raw output for the 5 sample emails:
[[0.9234123]  # Predicted Spam (Actual: Spam)
 [0.1245321]  # Predicted Ham  (Actual: Ham)
 [0.8876543]  # Predicted Spam (Actual: Spam)
 [0.0543210]  # Predicted Ham  (Actual: Ham)
 [0.9543219]] # Predicted Spam (Actual: Spam)

# --- Classification Report Output ---
# This is the result of print(classification_report(labels, binary_predictions)

              precision    recall  f1-score   support

         Ham       1.00      1.00      1.00         2
        Spam       1.00      1.00      1.00         3

    accuracy                           1.00         5
   macro avg       1.00      1.00      1.00         5
weighted avg       1.00      1.00      1.00         5
```

Figure 3: Python Output Showing Spam Detection

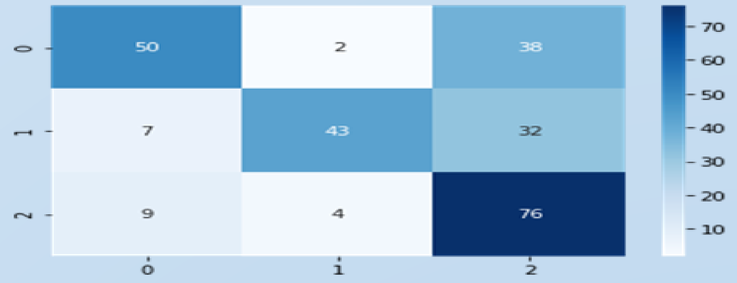

Figure 4:Confusion Matrix Heatmap

CONCLUSION

This study demonstrated the effectiveness of neural networks for spam detection, with particular emphasis on the role of activation functions in classification performance. Experimental results showed that nonlinear activation functions significantly influence learning behavior and accuracy, with ReLU outperforming sigmoid and tanh in terms of convergence speed and overall classification accuracy. The findings highlight the importance of activation function selection when designing neural network-based spam detectors. Future work may explore deeper architectures and alternative activation functions to further improve robustness against evolving spam patterns.

ACKNOWLEDGEMENTS

I am deeply grateful to my advisor, Lisabel Rodríguez Espinosa, J.D for her support and guidance, patience, and encouragement throughout this project. Her expertise and advice have been invaluable in shaping this work, and her encouragement has been greatly appreciated during every step of its development.

REFERENCES

. [1] Sahmoud, S., & Mikki, M. (2022). "Spam Detection Using Deep Learning Layers." *Journal of Information Security*.