Instituto Tecnológico y de Estudios Superiores de Monterrey

SMS Spam Detection System: Protecting Seniors from Phishing Attacks

Team 66 | Members:

Ana Karen Moreno Acevedo	A01773018 |IRS
David Flores Becerril	A01368391 |ITC
Diego Flores Becerril	A01769546 |ITC
Matias Piedra Pichardo	A01772503 |IRS
Santiago Benítez Pérez	A01782813 |ITC

26 oct 2024

Executive Summary:

This report details the development of an SMS spam detection system specifically designed to protect senior citizens from phishing attempts. The system combines a Naive Bayes machine learning model with Google's Gemini AI technology to not only detect spam messages but also provide user-friendly explanations of why certain messages are flagged as suspicious.

Introduction:

SMS phishing (smishing) has become increasingly prevalent, with senior citizens being particularly vulnerable to such attacks. Our solution addresses this challenge by implementing an intelligent filtering system that can:

- Detect potentially fraudulent SMS messages
- Provide clear, elderly-friendly explanations of why messages are flagged
- Forward legitimate messages while blocking suspicious ones

System Architecture:

Core Components:

1. **Machine Learning Model**
    - Implementation: Multinomial Naive Bayes classifier
    - Dataset: Set of SMS tagged messages that have been collected for SMS Spam research. It contains one set of SMS messages in English of 5,574 messages, tagged according to ham (legitimate) or spam.
    - Text Processing: NLTK library for advanced natural language processing
    - Feature Extraction: TF-IDF vectorization for text analysis
2. **API Service:**
    - Framework: Flask RESTful API
    - Entry Point: Single endpoint `/predict` for message classification
    - Response Format: JSON with prediction results and preprocessed message details
3. **External Services Integration:**
    - Google Gemini 1.5 Flash: Generates human-friendly explanations
    - Twilio: Handles SMS message delivery and routing

Technical Implementation:

Text Processing Pipeline:

1. Text Cleaning

- Special character removal
- Case normalization
- Whitespace standardization

2. Natural Language Processing

- Tokenization
- Stopword removal
- Word lemmatization

Model Training Process:

1. Data Preprocessing

- Corpus creation from cleaned text
- TF-IDF vectorization
- Label encoding for target variables

2. Model Performance

- Training Accuracy: 97%
- Test Accuracy: 96%
- 10 K-Fold Cross-validation implemented for finest reliability

API Workflow:

1. Request Processing:

```
POST /predict
{
   "message": "Sample message text",
   "phone_number": "+1234567890"
}
```

2. Response Flow:

For Spam Messages:

- Generates explanation using Gemini AI
- Sends warning SMS with explanation

For Legitimate Messages:

- Forwards original message to recipient

Security and Environmental Configuration:

- Environment variables for sensitive credentials
- Secure API key management
- Twilio authentication token protection

Dependencies:

- Flask: 3.0.3
- joblib: 1.4.2
- numpy: 1.26.4
- nltk: 3.8.1
- google-generativeai: 0.3.2
- twilio: 8.13.0
- python-dotenv: 1.0.1

Future Enhancements:

1. Model retraining pipeline for emerging threats
2. Multi-language support
3. Enhanced explanation generation
4. User feedback integration

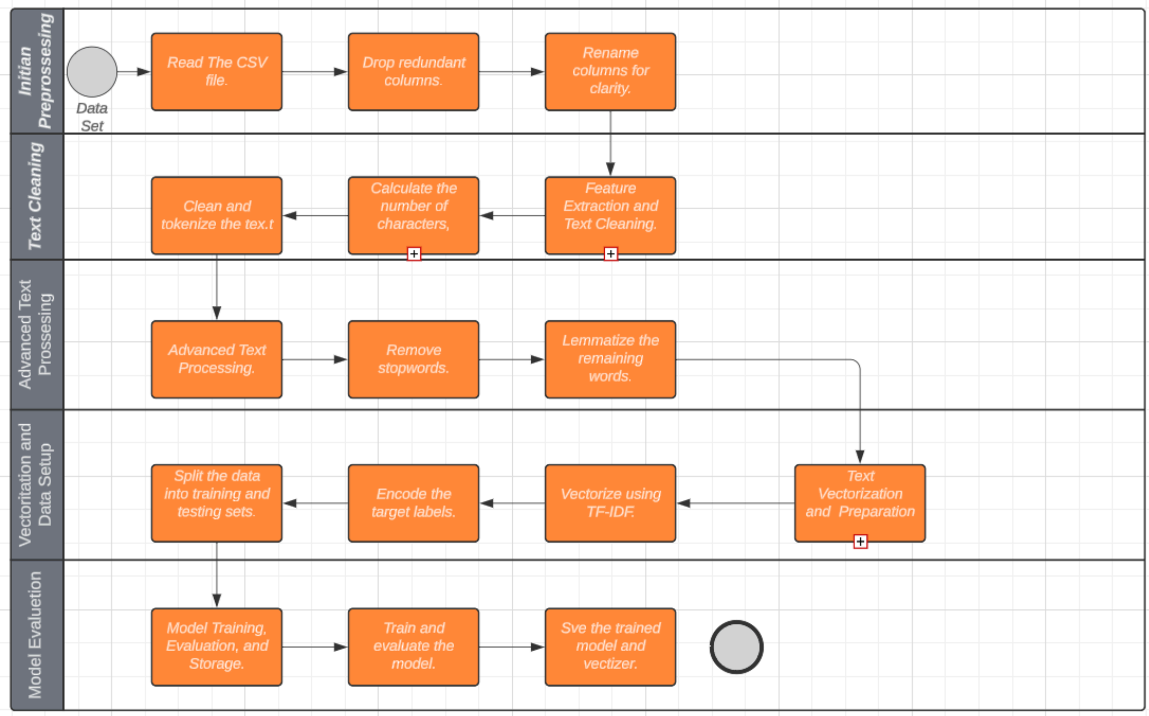
Initian Preprossesing
Data Set
Read The CSV file.
Drop redundant columns.
Rename columns for clarity.
Text Cleaning
Feature Extraction and Text Cleaning.
Calculate the number of characters,
Clean and tokenize the tex.t
Advanced Text Prossesing
Advanced Text Processing.
Remove stopwords.
Lemmatize the remaining words.
Vectoritation and Data Setup
Text Vectorization and Preparation
Vectorize using TF-IDF.
Encode the target labels.
Split the data into training and testing sets.
Model Evalueation
Model Training, Evaluation, and Storage.
Train and evaluate the model.
Sve the trained model and vectizer.

Fig 1. Architecture Diagram

Conclusion:

This system provides a robust solution for protecting senior citizens from SMS-based phishing attacks. By combining machine learning with user-friendly explanations, we create a protective barrier that not only blocks suspicious messages but also educates users about potential threats.