Spam Detector Based on ML

Project for The Bridge Data Science Bootcamp

Fernando Manzano
July 2024

The Spam Problem

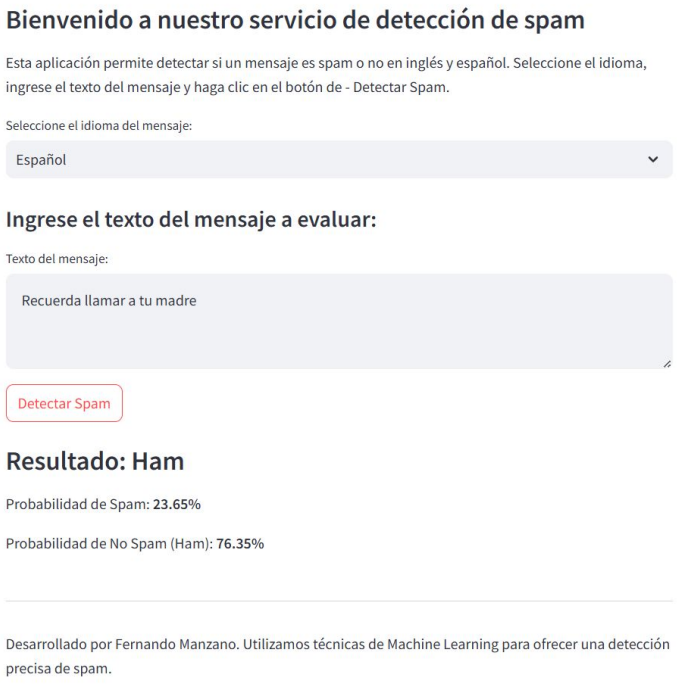
1. **Average Spam Percentage in Global Email Traffic (2021)**
   In 2021, the average percentage of spam in global email traffic was 45.56%.
   Source: Kaspersky
2. **Spain: Global Leader in Spam Reception**
   During the second quarter of 2020, Spain was the country that received the most spam attacks, accounting for 9.3% of all such threats worldwide.
   Source: Europa Press
3. **Acceptable Spam Complaint Rate**
   The industry-standard acceptable spam complaint rate is below 0.1%, which equates to one complaint per 1,000 messages sent.
   Source: ActiveCampaign

Project Description

Objective: Classify texts as spam or ham (legitimate) in both English and Spanish.

Results:

- Development of trained model(s) for spam detection in Spanish and English based on natural language texts, including the processes required for their implementation.
- **Application:**
  Developed using Streamlit to provide an easy-to-use web interface.



Data Used

Spanish:

Source: Hugging Face
https://huggingface.co/datasets/softecapps/spam_ham_spanish/tree/main

mensaje string	tipo string
Descubre como perder peso rapidamente	spam
Necesitas ayuda con tu tarea	ham
Gana dinero desde casa sin esfuerzo	spam
Reclama tu herencia de un pariente lejano	spam
Mejora tu rendimiento sexual con este producto	spam
Por favor responde a esta encuesta	ham

- **Description:** Over 1,000 text messages in Spanish labeled as spam or ham in `train.csv` and `test.csv`.
- **Structure:**
  - `Message`: Text of the message.
  - `Label`: Indicates spam or ham.

English:

Source: Kaggle
https://www.kaggle.com/datasets/venky73/spam-mails-dataset

- **Description:** 4,993 unique text messages in English labeled as spam or ham in `spam_ham_dataset.csv`.
- **Structure:**
  - `text`: Text of the message.
  - `label`: Indicates spam or ham.
  - `label_num`: Indicates spam = 1 or ham = 0.

#	label Labels of Emails which can be either Spam or Ham	text Emails data	label_num if spam it's 1, or else it's 0
0 5170	ham 71% spam 29%	4993 unique values	0 1
605	ham	Subject: enron methanol ; meter # : 988291 this is a follow up to the note i gave you on monday , 4...	0
2349	ham	Subject: hpl nom for january 9 , 2001 (see attached file : hplnol 09 . xls) - hplnol 09 . xls	0
3624	ham	Subject: neon retreat ho ho ho ,	0

Data Preprocessing

- Removal of duplicates and repeated values.
- Removal of stop words using NLTK stopwords.
- Removal of non-alphanumeric characters.
- Conversion to lowercase.
- Removal of extra spaces and prefixes such as "Subject:"

```
nltk.download('stopwords')
stopwords_sp = stopwords.words('spanish')

# Agregar palabras adicionales que no aporten significado en este contexto específico
stopwords_sp.extend(['este', 'nuestro', 'con', 'para', 'esta'])  # Podemos ajustar esta

# Función para limpiar el texto y filtrar stopwords
def limpiar_texto_con_stopwords(texto):
    texto = re.sub(r'\W', ' ', texto) # Eliminar caracteres no alfanuméricos
    texto = texto.lower() # Convertir a minúsculas
    texto = re.sub(r'\s+', ' ', texto) # Eliminar espacios extra
    palabras = texto.split()
    palabras_filtradas = [palabra for palabra in palabras if palabra not in stopwords_s
    return ' '.join(palabras_filtradas)

# Aplicar limpieza con filtro de stopwords a los mensajes
test_df['mensaje_limpio_stopwords'] = test_df['mensaje'].apply(limpiar_texto_con_stopwo
```

Models Used

- **TfidfVectorizer:**
  a. Converts texts into a TF-IDF feature matrix.
  b. Assigns weight to each word based on its frequency in the document and its inverse frequency in the document set.

- **Evaluated Models:**
  a. **Spanish:** Naive Bayes, Random Forest, and SVM. **Selected:** Naive Bayes.
  Naive Bayes is suitable for text processing and spam detection due to its simplicity, computational efficiency, ability to handle large volumes of data, and effective use of conditional independence between words to calculate probabilities.
  b. **English:** Naive Bayes and SVM. **Selected:** SVM optimized with GridSearchCV.
  SVM is notable for text processing and spam detection due to its ability to handle high-dimensional spaces, such as text data, find an optimal hyperplane that maximizes class separation, and its robustness against noisy data.

Resultados y Evaluación

	Naive Bayes (Spanish)	**SVM (English)** Optimized Parameters: {'C': 10, 'class_weight': None, 'kernel': 'rbf'}`
Precision	0.89	0.98
Recall	0.89	0.98
F1 - Score	0.89	0.98
Confusion Matrix	[[94 15] [8 92]]	[[719 13] [7 260]]

User-Centric Streamlit App

Detección de Spam en Mensajes

Bienvenido a nuestro servicio de detección de spam

Esta aplicación permite detectar si un mensaje es spam o no en inglés y español. Seleccione el idioma, ingrese el texto del mensaje y haga clic en el botón de - Detectar Spam.

Seleccione el idioma del mensaje:

Español

Ingrese el texto del mensaje a evaluar:

Texto del mensaje:

Quiero conocerte

Detectar Spam

Resultado: Spam

Probabilidad de Spam: **64.75%**

Probabilidad de No Spam (Ham): **35.25%**

Bienvenido a nuestro servicio de detección de spam

Esta aplicación permite detectar si un mensaje es spam o no en inglés y español. Seleccione el idioma, ingrese el texto del mensaje y haga clic en el botón de - Detectar Spam.

Seleccione el idioma del mensaje:

Inglés

Ingrese el texto del mensaje a evaluar:

Texto del mensaje:

Amazon prime days

Detectar Spam

Resultado: Ham

Probabilidad de Spam: **10.91%**

Probabilidad de No Spam (Ham): **89.09%**

Desarrollado por Fernando Manzano. Utilizamos técnicas de Machine Learning para ofrecer una detección precisa de spam.

Lessons Learned

Importance of Organization from the Start: Ensuring a well-structured approach from the beginning significantly enhances project efficiency and effectiveness.
Diverse Solutions for the Same Problem: Different methods can address the same issue in various ways, highlighting the importance of exploring multiple approaches.
Potential of ML and DL Libraries: There is a wealth of capabilities and tools within machine learning (ML) and deep learning (DL) libraries that offer extensive opportunities for innovation and improvement.

Future Improvements

Include App Functionality Explanation: Enhance the application by adding a detailed explanation of its workings to improve user understanding and interaction.
Apply to New Cases: Extend the approach to address additional scenarios such as phishing and fraudulent SMS.
Create a Spanish Spam Dataset: Develop a dedicated spam dataset in Spanish to facilitate further research and improvements in this language.
Improve Models for Error Reduction: Refine models to address and reduce incorrect predictions, ensuring more accurate spam detection.