Innovative Strategies for SMS Spam Detection

Yu-Chih (Wisdom) Chen

5 May 2024

Executive Summary

- The project successfully developed a predictive model using the SMS Spam Collection dataset to accurately classify SMS messages as spam or ham.
- The Pooled Bi-Directional GRU model demonstrated the highest accuracy, followed by the improved CNN model, with LSTM showing the least accuracy among the three.
- The methodology included preprocessing, feature engineering with tokenization and p adding, and leveraging pre-trained embeddings for model training.

Methodology

- A combination of Convolutional Neural Networks (CNNs), Long Short-Term Memory (LSTM) networks, and Pooled Bi-Directional Gated Recurrent Units (GRUs) were employed to address the classification task.
- Hyperparameter tuning, regularization, and dropout adjustments were applied to optimize model performance and prevent overfitting.

Source Data

- The SMS Spam Collection dataset consists of 190K English SMS messages, each labeled as either spam or ham, sourced from various online platforms for research purposes.
- The dataset's composition reflects a real-world distribution of spam and ham messages, pr oviding a solid foundation for model training and evaluation.
- Download the data directly from https://www.kaggle.com/datasets/meruvulikith/190k-spam-ham-email-dataset-for-classification/data

Problem Statement

Objective

Leverage machine learning to develop a predictive model capable of accurately classifying SMS messages as either spam (unwanted messages) or ham (legitimate messages)

Challenge

It lies in creating a model that can effectively differentiate between these two categories with high precision and recall, using the SMS Spam Collection dataset, which comprises a set of SMS tagged messages specifically collected for SMS Spam research.

Assumptions Data & Model

01. Assumption of Data

- It accurately reflects the real-world distribution of spam and ham messages, providing a solid foundation for training and evaluating the model.
- It has undergone thorough preprocessing to ensure that the text data is clean and free from noise, which could otherwise adversely affect the model's learning process.
- It split into training and testing sets in a manner that prevents data leakage and allows for the evaluation of the model's generalization capabilities.
- The labels in the dataset (spam or ham) are correctly assigned, ensuring that the model's learning is based on accurate ground truth data.

02. Hypotheses of Model

- Advanced neural network architectures, such as LSTM and Bi-Directional GRU, are hypothesized to perform well on this text classification task due to their ability to capture sequential information and context within messages
- The Bi-Directional GRU model, in particular, is expected to outperform other models because it processes data in both directions, potentially capturing more complex patterns and dependencies in the text data.
- The performance of the models can be accurately assessed using metrics such as AUC, AP, accuracy, precision, recall, and F1-score, which provide a comprehensive view of the models' strengths and weaknesses in classifying spam and ham messages.

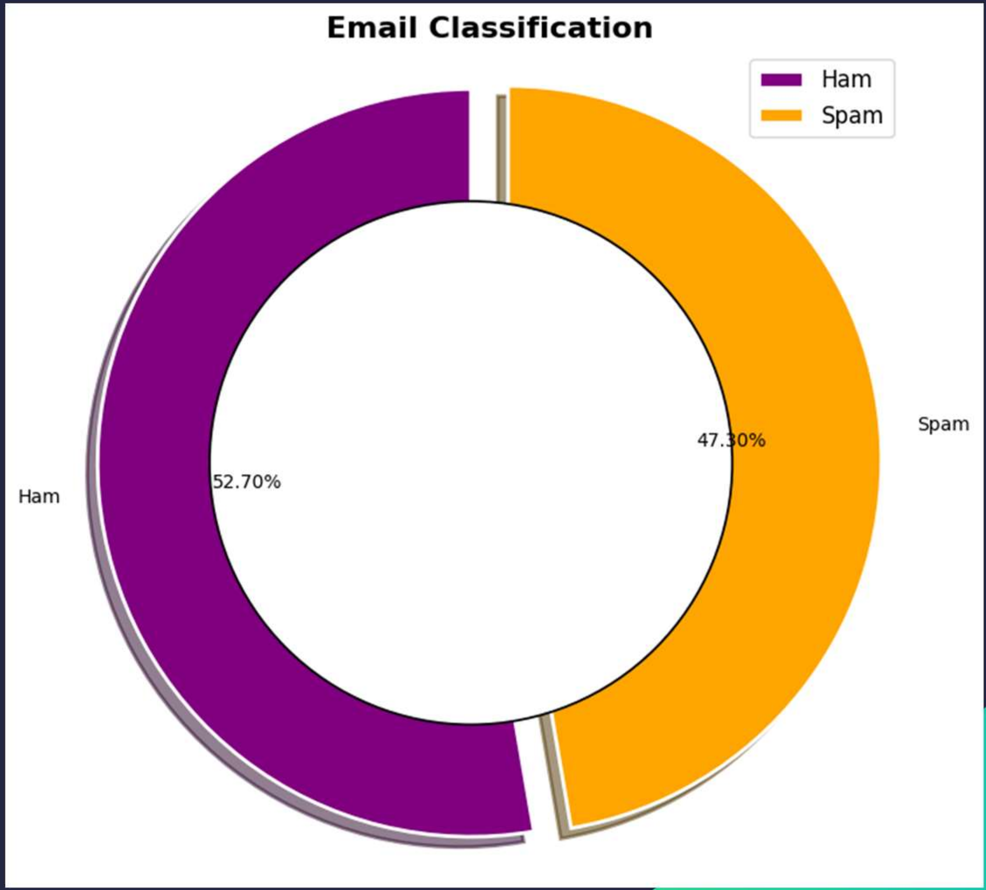
01

Exploratory Data Analysis

Spam vs. Ham

A Comparative View of Ham and Spam Emails

It is relatively balanced with 52.70% of emails classified as "Ham" and 47.30% as "Spam," which is conducive for training a machine learning model without a significant class imbalance



Overall Email Distribution Characteristics Insight

Skewness and Outlier Impact

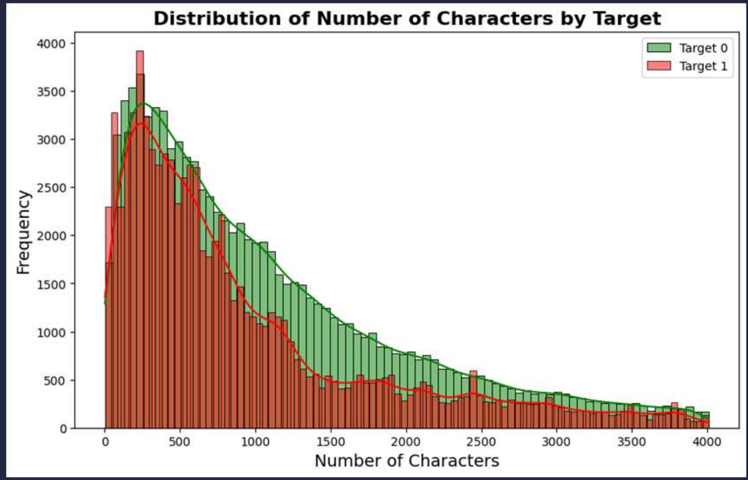
The distribution of the number of characters, words, and sentences in both "Ham" and "Spam" emails is right-skewed, indicating that most emails are shorter with a few outliers having significantly more content

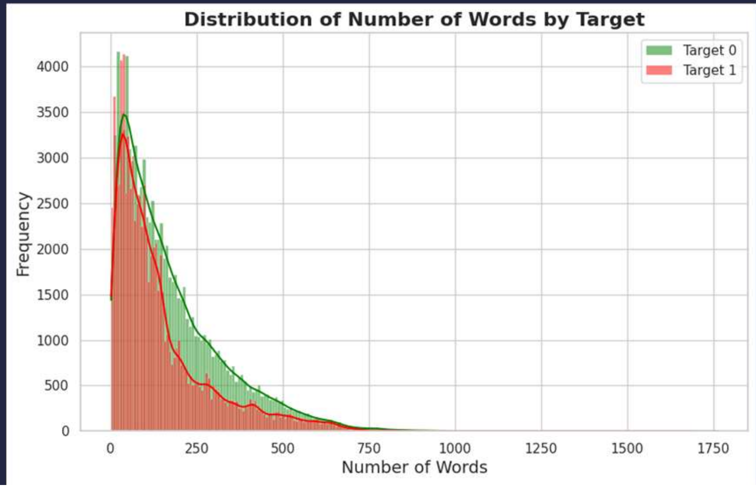
	Number of Character	Number of Words	Number of Sentence
Mean	1.839160e+03	2.805164e+02	3.690508
Median	8.120000e+02	1.290000e+02	1.000000
Maximum	1.151031e+07	1.585483e+06	3093.0000

Email Content Characteristics Insight I

Comparative Length Analysis

After removing outliers (emails with character counts beyond the upper bound defined by the interquartile range), the distributions become more normalized, which may improve the performance of the machine learning models by reducing the impact of extreme values





Email Content Characteristics Insight II

Comparative Length Analysis

"Ham" emails tend to be longer in terms of characters, words, and sentences compared to "Spam" emails, with averages of 1,087 characters, 175 words, and 2.84 sentences for "Ham" and 925 characters, 145 words, and 2.98 sentences for "Spam"

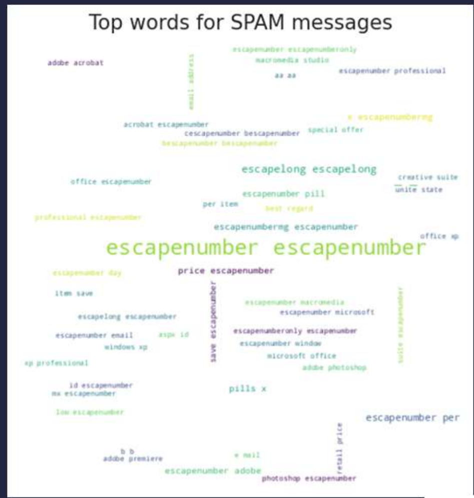
Ham			
	Number of Character	Number of Words	Number of Sentence
Mean	1,087	175	2.84

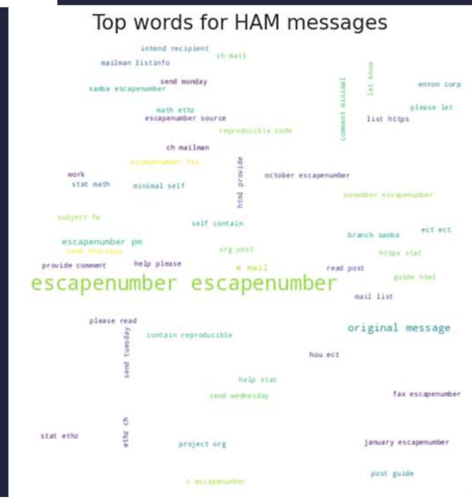
Spam			
	Number of Character	Number of Words	Number of Sentence
Mean	925	145	2.98

Comparative Analysis of SPAM and HAM Messages

Visualizing Key Terms to Differentiate Unwanted vs. Legitimate Emails

- **SPAM** messages word cloud highlights frequent terms like "escapenumber" and "professional," suggesting a focus on misleading business-related content
- Common words in SPAM also include "special offer" and "creative suite," indicative of promotional content aimed at deception
- **HAM** messages word cloud shows a prevalence of words such as "escapenumber," "please," and "help," reflecting the more personal and direct nature of legitimate communication
- Terms like "original message" and "fax escapenumber" in the HAM word cloud suggest routine, business-related correspondence





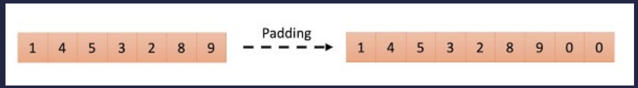
02

Feature Engineering & Transformation

Preparing Text Data for Predictive Modeling I

Optimizing Text Data for Deep Learning I

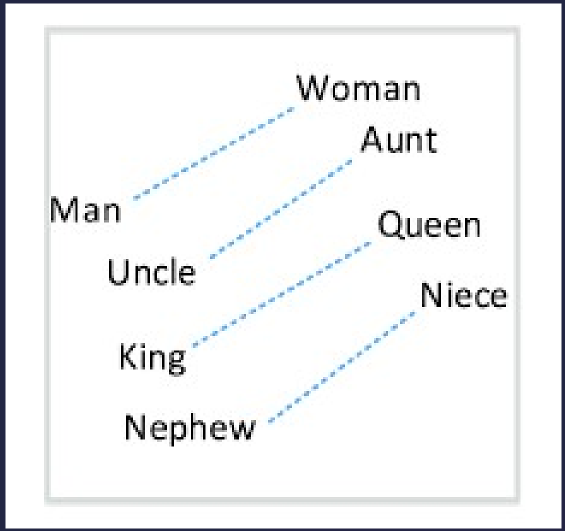
1. Tokenization Process
   - Implemented the **Keras Tokenizer** to convert raw email text into sequences of integers, enabling the model to interpret and process natural language data.
   - Restricted the tokenizer's vocabulary to the 20,000 most common words to balance the complexity and performance of the model.
2. Sequence Padding
   - Standardized the length of sequences using **pad_sequences** to a fixed size of 200 tokens, ensuring uniform input dimensions for the neural network.
   - Restricted the tokenizer's vocabulary to the **20,000 most common words** to balance the complexity and performance of the model.



Preparing Text Data for Predictive Modeling I I

Optimizing Text Data for Deep Learning II

3. Embedding Matrix Creation
   - Leveraged pre-trained GloVe word embeddings to provide the model with rich, pre-learned word representations, enhancing the model's ability to capture linguistic nuances.
   - Created an embedding matrix that associates each word in the tokenizer's vocabulary with a **100-dimensional GloVe vector**.
4. Utilization of Pre-trained Embeddings
   - Integrated GloVe embeddings into the model to benefit from existing knowledge about word relationships and semantics, which is particularly useful for the email classification task.



03

Model Deployment

Strategic Model Selection for SMS Spam Detection

Harnessing Deep Learning Architectures for Enhanced Classification

Convolutional Neural Networks

- It excel in identifying local and position-invariant patterns, which is beneficial for detecting specific keywords and phrases indicative of spam

Long Short-Term Memory

- It designed to recognize long-term dependencies and contextual relationships in sequential data, such as the order of words in SMS messages

Pooled Bi-Directional GRU

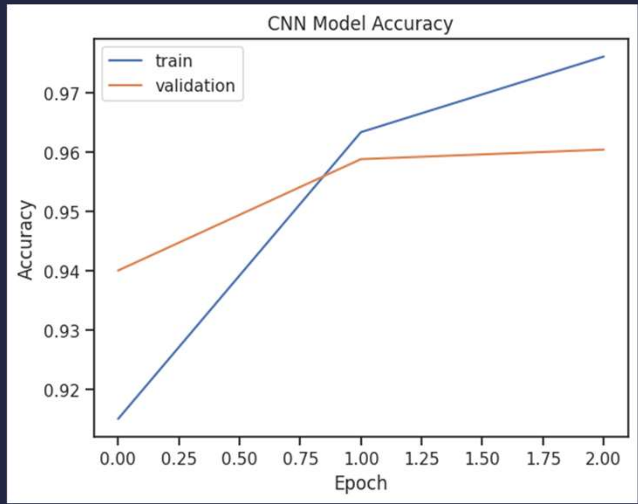
- Bi-Directional GRUs process data in both forward and reverse directions, capturing patterns that may be missed when only considering one direction

04

Model Evaluation

Model Performance and Generalization Analysis I

Evaluating CNN Model on SMS Spam Detection



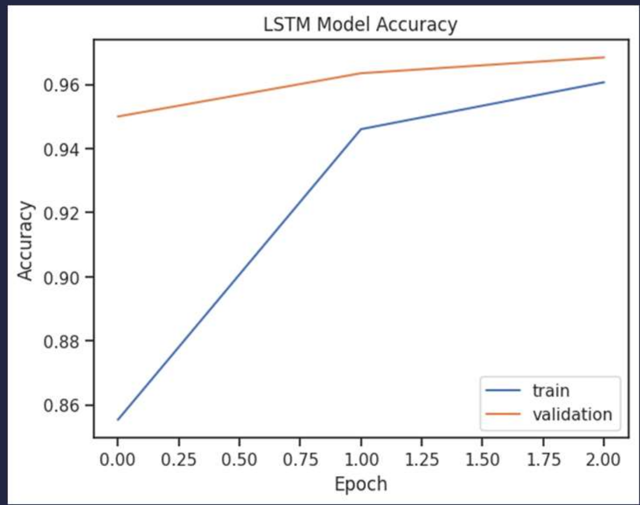
- The CNN model showed high training accuracy but lower testing accuracy, suggesting potential overfitting

Convolutional Neural Networks	
Training Error	0.262
Testing Error	1.067

```
Classification Report for train data - CNN:
              precision    recall  f1-score   support

           0       0.98      0.93      0.95     61828
           1       0.93      0.98      0.95     56316

    accuracy                           0.95    118144
   macro avg       0.95      0.96      0.95    118144
weighted avg       0.96      0.95      0.95    118144


Classification Report for test data - CNN:
              precision    recall  f1-score   support

           0       0.98      0.91      0.94     30323
           1       0.91      0.98      0.94     27868

    accuracy                           0.94     58191
   macro avg       0.94      0.94      0.94     58191
weighted avg       0.95      0.94      0.94     58191
```

Model Performance and Generalization Analysis II

Evaluating LSTM Model on SMS Spam Detection



- LSTM model displayed excellent performance with high accuracy and low errors on both training and test sets, indicating effective learning without significant overfitting

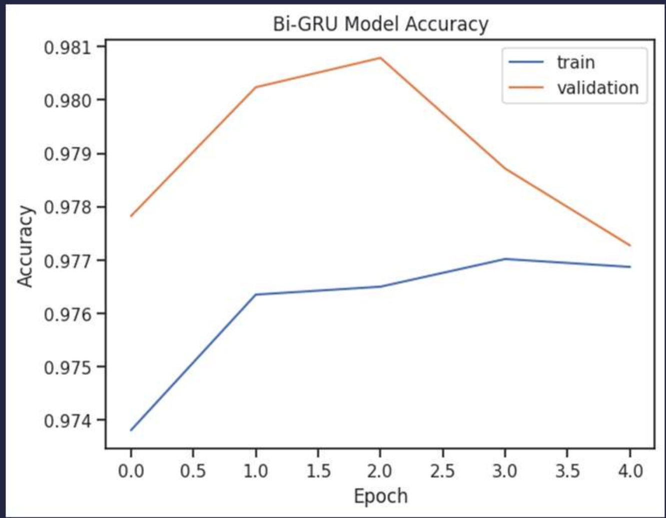
```
Classification Report for train data - LSTM:
              precision    recall  f1-score   support

           0       0.97      0.94      0.95     61828
           1       0.94      0.96      0.95     56316

    accuracy                           0.95    118144
   macro avg       0.95      0.95      0.95    118144
weighted avg       0.95      0.95      0.95    118144


Classification Report for test data - LSTM:
              precision    recall  f1-score   support

           0       0.97      0.94      0.95     30323
           1       0.93      0.96      0.95     27868

    accuracy                           0.95     58191
   macro avg       0.95      0.95      0.95     58191
weighted avg       0.95      0.95      0.95     58191
```

Long Short-Term Memory	
Training Error	0.105
Testing Error	0.135

Model Performance and Generalization Analysis III

Evaluating Bi-GRU Model on SMS Spam Detection

- Bi-GRU model outperformed other models with the highest accuracy and lowest errors, showing exceptional generalization capabilities



Pooled Bi-Directional GRU	
Training Error	0.059
Testing Error	0.050

```
Classification Report for train data - Bi-GRU:
              precision    recall  f1-score   support

           0       0.99      0.99      0.99     61828
           1       0.99      0.99      0.99     56316

    accuracy                           0.99    118144
   macro avg       0.99      0.99      0.99    118144
weighted avg       0.99      0.99      0.99    118144


Classification Report for test data - Bi-GRU:
              precision    recall  f1-score   support

           0       0.98      0.99      0.98     30323
           1       0.98      0.98      0.98     27868

    accuracy                           0.98     58191
   macro avg       0.98      0.98      0.98     58191
weighted avg       0.98      0.98      0.98     58191
```

05

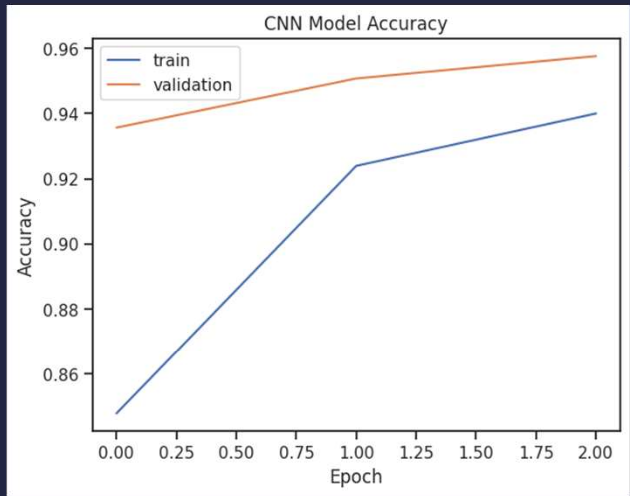
Model Improvement

Enhancements in CNN Model for SMS Spam Detection

Techniques for Boosting Performance and Generalization

1. Regularization Techniques
   - L2 regularization was introduced in the convolutional layer of the improved CNN model to combat overfitting by penalizing large weights
   - It helps the model generalize better to unseen data by encouraging simpler models that perform well on the validation set.
2. Dropout Rate Adjustment
   - An increased dropout rate was applied in the improved CNN model to prevent overfitting by randomly dropping units during training, which forces the model to learn more robust features

Convolutional Neural Networks	
Training Error	0.249
Testing Error	0.386




```
Classification Report for train data - CNN:
              precision    recall  f1-score   support

           0       0.95      0.94      0.94     61828
           1       0.93      0.95      0.94     56316

    accuracy                           0.94    118144
   macro avg       0.94      0.94      0.94    118144
weighted avg       0.94      0.94      0.94    118144


Classification Report for test data - CNN:
              precision    recall  f1-score   support

           0       0.94      0.93      0.94     30323
           1       0.93      0.94      0.93     27868

    accuracy                           0.94     58191
   macro avg       0.94      0.94      0.94     58191
weighted avg       0.94      0.94      0.94     58191
```

06

Model Result

Deep Learning Model Efficacy in SMS Spam Detection

text_lemmatized	predicted_cnn	predicted_lstm	predicted_bigru
saravana kumar write yitzle write read one lis...	0	0	0
jackie talk darren deal reference hpl deal dyn...	0	0	0
obeisance guest commission obstruct bookshelve...	0	1	1
value recipient yearas hottest accessory coach...	1	1	1
enron energy service middle market east taylor...	0	0	0
http describewomen hk miss unique escapelong p...	1	1	1

Comparative Analysis and Methodological Insights

1. Model Accuracy Results
   - The improved CNN model achieved a high validation accuracy of 0.966, indicating a successful enhancement over the original model.
   - The **Pooled Bi-Directional GRU model exhibited the highest validation accuracy** of 0.978, demonstrating its robustness and effectiveness in processing sequential data for spam detection.
2. Methodological Learnings
   - The application of pre-trained embeddings and hyperparameter tuning through Bayesian Optimization contributed to the improved performance of the CNN model
   - The superior performance of the **Bi-Directional GRU** model underscores the value of bidirectional processing and gated mechanisms in handling text classification challenges.

Model Test Accuracy Comparison	
CNNs	0.966
LSTMs	0.833
Bi-GRU	0.978

Future Step

- Explore the integration of additional linguistic features and more sophisticated natural langu age processing techniques to further enhance model accuracy.
- Investigating the impact of class imbalance and developing strategies to mitigate its effects will be a priority.
- Continuous model evaluation with real-world data will be conducted to ensure the model's r obustness and adaptability over time.
- Efforts will be made to reduce the model's complexity without compromising performance to facilitate deployment in resource-constrained environments.

Thanks !