SPAM

Stop Ph**n Annoying Me!

Enhancing Email Spam Detection: A Comparative Analysis of Traditional ML Algorithms and RoBERTa Language Model"

By: Rishabhraj Jalota
DS-UA 301 Final Project

Executive Overview

Problem Statement (Goal): Aim to develop an ML-driven system capable of detecting spam emails, ensuring enhanced security and better user experience.

Solution Approach:

- Utilized state-of-the-art LLM, RoBERTa, and benchmarked its performance against traditional ML models like SVM, NB, etc.
- Implemented using the transformers library on Google Colab, and conducted experiments on four distinct datasets

Value/Benefit: Efficient and Accurate Spam Detection:

- Achieved a F1 score of up to 1.0 using RoBERTa on the YouTube dataset, showcasing the potential of LLMs in spam detection.
- Scalability: LLMs can be adapted for larger datasets or other text classification tasks, promising adaptability for future challenges.
- Time-saving: With automated filtering, users can trust the system to minimize manual sorting of spam, freeing up time and enhancing email experience

Background Information

Spam-T5: Benchmarking Large Language Models for Few-Shot Email Spam Detection (2023). Retrieved 13 August 2023, from https://arxiv.org/pdf/2304.01238.pdf

jpmorganchase. (2023). GitHub - jpmorganchase/llm-email-spam-detection: LLM for Email Spam Detection. Retrieved 13 August 2023, from https://github.com/jpmorganchase/llm-email-spam-detection

Ahmed, N., Amin, R., Aldabbas, H., Koundal, D., Alouffi, B., & Shah, T. (2022). Machine Learning Techniques for Spam Detection in Email and IoT Platforms: Analysis and Research Challenges. *Security And Communication Networks*, *2022*, 1-19. doi: 10.1155/2022/1862888

https://www.hindawi.com/journals/scn/2022/1862888/

Technical challenges

Model Complexity vs. Performance: SPAM-T5, a highly powerful model, posed computational challenges during training, requiring a shift to RoBERTa,

Balancing Multiple Models: Ensuring comparable performance across various models from classical machine learning to cutting-edge language models.

Feature Extraction: Leveraging transformer-based models for automatic feature extraction, bypassing manual feature engineering but introducing its own challenges.

Evaluation Metrics: Ensuring that metrics like F1 score, precision, and recall are considered, given the nature of the imbalance in spam datasets.

Integration with J.P. Morgan's Framework: Adapting the given J.P. Morgan infrastructure to work seamlessly with my datasets and models.

Approach

Diverse Data Sources:

- Aggregating and combining datasets from different YouTube videos introduced variability.

Text Data Complexity:

- Handling diverse languages, slangs, and emojis in YouTube comments.
- Preprocessing requirements for such varied data were extensive.

Model Training Challenges:

- RoBERTa, given its size and complexity, introduced significant runtime issues.
- Limited to DistilRoBERTa due to constraints in training time and computational resources.

Integration with J.P. Morgan Framework

- Adapting MY solution to be compatible with the specific directory structure and conventions of the J.P. Morgan framework.

. LLM Model Selection:

- Given runtime constraints, couldn't explore other potentially beneficial LLMs like BERT, XLNet, SPAM-T5.

Performance Evaluation:

- Ensuring fair and comprehensive evaluation across both baseline models and advanced LLMs.

Implementation & Experimentation Details

Data Processing:

- Combined YouTube datasets from five different videos.
- Preprocessing: Tokenization, removing stop words, and handling emojis/slangs.

Platform & Frameworks:

- Google Colab for computation.
- Utilized transformers library for LLMs (RoBERTa).
- Scikit-learn for baseline models and evaluation metrics.

Model Training:

- Baseline Models: Naive Bayes, Logistic Regression, KNN, SVM, XGBoost, LightGBM.
- LLM: Only lRoBERTa due to computational constraints.

Model Evaluation:

- Split data into training, validation, and test sets.
- Metrics: F1-score, Precision, Recall, and Accuracy.
- RoBERTa fine-tuned using multiple epochs to optimize performance.

Integration:

- Adapted data and model workflows to fit within the J.P. Morgan framework directory structure and conventions.

Evaluation

When your email puts legitimate emails in spam
Don't you ever not tell me things I wanna know!

Baseline Model

Baseline Model Interpretation:

Naive Bayes (NB):Achieved an F1 of 0.858 and an accuracy of 0.847.This means that NB was fairly successful in balancing precision and recall, but it wasn't the top-performing model.

Logistic Regression (LR):With an F1 of 0.891 and accuracy of 0.893, it's evident that LR was able to classify spam comments with a higher rate of success than NB. It suggests that the dataset might be linearly separable to some extent.

K-Nearest Neighbors (KNN):The relatively lower F1 score of 0.746 and accuracy of 0.730 indicate that the dataset might be too complex for simple distance-based classifications. Some comments might be close to each other in the feature space but belong to different classes.

Support Vector Machine (SVM): Achieving an F1 of 0.903 and accuracy of 0.903, SVM was one of the top-performing models. This indicates that the model could find a hyperplane that effectively separates the spam and non-spam comments.

XGBoost: With an F1 of 0.875 and accuracy of 0.878, XGBoost performed commendably. The gradient boosting algorithm might be capturing complex relationships in the data.

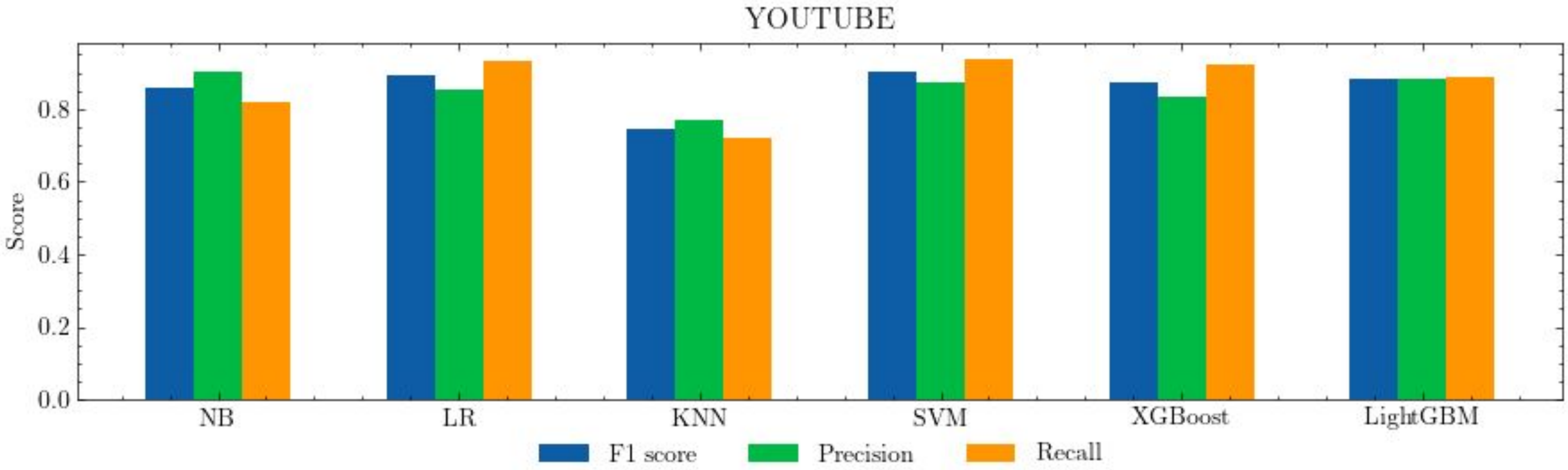
LightGBM: Achieved an F1 of 0.886 and accuracy of 0.883. Being another gradient boosting model, it performed similarly to XGBoost, suggesting that ensemble methods are effective for this dataset.

RoBERTa LLM

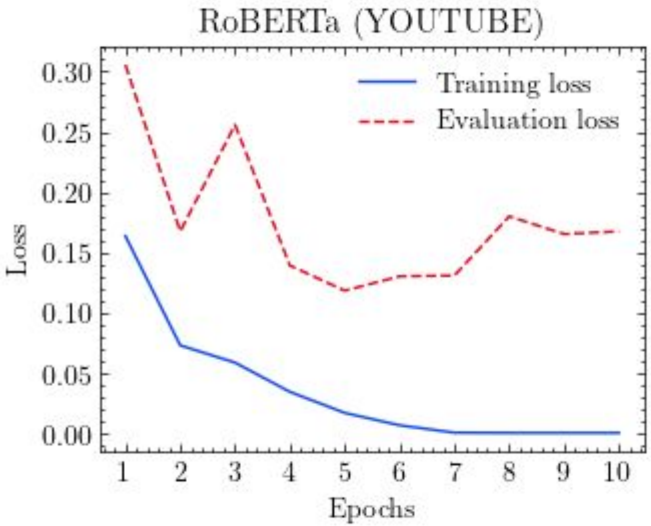
LLM RoBERTa Interpretation:

- Early Epochs (e.g., Epoch 1):
    - The F1 score of 0.972 in the first epoch itself is remarkable. This suggests that the pre-trained knowledge in the RoBERTa model is highly relevant and useful for the spam detection task on YouTube comments.
- Middle Epochs (e.g., Epoch 5):
    - The F1 score improves to 0.996, and the validation loss reduces dramatically. This indicates that as the model trains, it's becoming more adept at recognizing the nuances of spam in the YouTube comments dataset.
- Later Epochs (e.g., Epoch 10):
- Achieving a perfect F1 score of 1.000 shows that RoBERTa has almost perfectly learned to classify spam and non-spam comments from the training dataset. However, one must be cautious about potential overfitting, even though the validation loss is minimal

Some visuals Baseline



Some Visuals: LLM



Conclusion

I will believe in you
CHIBIRD
even when you have trouble believing in yourself.