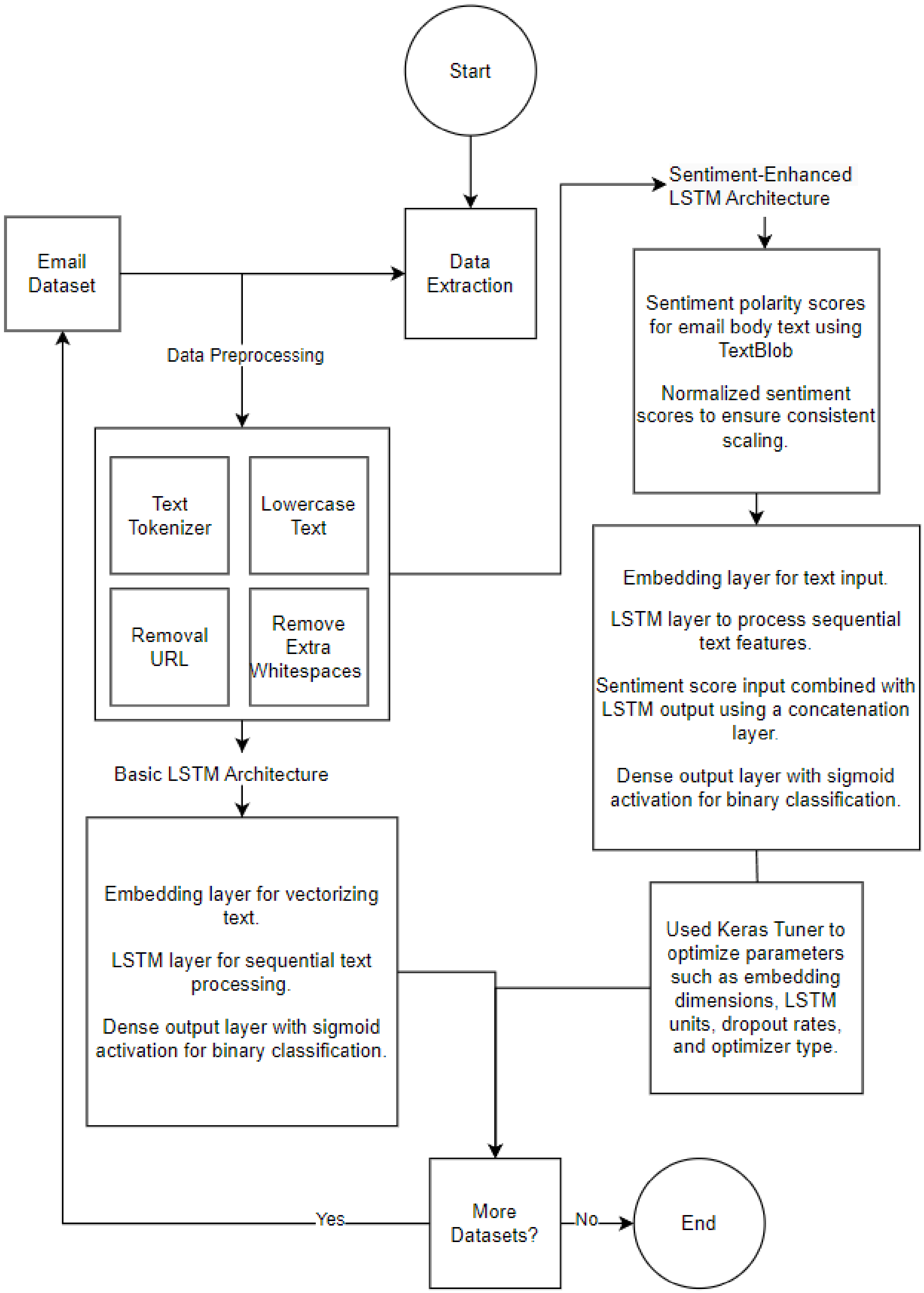
Text Classification for Cybersecurity:

Identifying Phishing Emails and Malicious Content


Start
Email Dataset
Data Extraction
Data Preprocessing
Sentiment-Enhanced LSTM Architecture
Sentiment polarity scores for email body text using TextBlob
Normalized sentiment scores to ensure consistent scaling.
Text Tokenizer
Lowercase Text
Removal URL
Remove Extra Whitespaces
Embedding layer for text input.
LSTM layer to process sequential text features.
Sentiment score input combined with LSTM output using a concatenation layer.
Dense output layer with sigmoid activation for binary classification.
Basic LSTM Architecture
Embedding layer for vectorizing text.
LSTM layer for sequential text processing.
Dense output layer with sigmoid activation for binary classification.
Used Keras Tuner to optimize parameters such as embedding dimensions, LSTM units, dropout rates, and optimizer type.
Yes
More Datasets?
No
End

By: Kerim Sever

Abstract:

Data Classification in NLP has significantly changed in the last few years but as there has also been a lot of cyber security threats as technology advances. Phishing emails have become more popular since attackers have been implementing AI generated phishing emails. Phishing Emails seem harmless as users can determine if it's fake or not but now with AI creating the emails it is not as clear. The goal of this project is to focus on phishing emails by using a traditional LSTM text classification model and comparing those results to a novel LSTM model that integrates sentiment analysis scores and randomized search for its parameters. The dataset that is used for this study is from a corporation with the users of Ling, Enron, and CEAS (comprehensive dataset) and collectively contains over 100 thousand plus emails. Having 3 different datasets allows for a robust evaluation from different sources to test the model comparisons when detecting phishing classification. The process first involves a preprocessing step to the data to create a clean dataset from all 3 sources to work off of in the deep learning models. The main idea of sentiment scoring comes from the logic that emails with a lower sentiment score will most likely be a phishing related email and emails with a higher sentiment score will be personal emails from coworkers or expected emails. The goal of this paper is to highlight a hybrid model that combines NLP techniques with advanced sentiment analysis against a traditional LSTM model to demonstrate the effectiveness of this novel approach.

Introduction:

Long Short-Term Memory (LSTM) models are promising when it comes to capturing patterns in sequential data in text classification. Because of this LSTM would be ideal when it comes to phishing detection. The research paper, *Analysis and Prevention of AI-based Phishing Email Attacks* by Eze and Shamir (2024), talk about how AI-driven phishing emails are becoming more advanced and get around traditional detection and classification methods since they rely on text recognition and keyword matching. This creates a problem of phishing emails getting through safety systems and forces companies to use more advanced models to effectively block these attacks. Ige, Tosin, Christopher Kiekintveld, and Aritran Piplai (2024), propose a multi-layered pipeline to improve phishing detection. Their suggestions to other companies is that models like LSTM, Logistic Regression, and other text classification models are required to properly detect these phishing attacks. They suggest a multi-layered framework that combines a deep learning model to improve phishing detection. While a multi-layered deep learning LSTM model has shown promise, there is room for innovation. The novelty of this paper proposes a sentiment analysis scoring alongside a deep learning LSTM model to better detect phishing attacks. Sentiment scores that come from the body of emails can give additional insights from these emails and can provide a layer of context to allow the model to better classify between legitimate and phishing emails.

This approach builds on the foundational work of NLP text classification and introduces a new novel idea of sentimental analysis as a feature which has not been explored upon in text classification for phishing email detection. The idea of sentiment scoring allows the model to better understand the tone of the message and to gather the intention behind the email. This provides an extra feature for the LSTM model to work with. This novel idea has the goal to improve phishing detection accuracy and tackle the challenges of AI generated phishing attacks.

Methodology:

The goal of this project is to test the effectiveness of a basic LSTM model for phishing email detection in text classification and investigate the improvement by including sentiment analysis scoring and hyperparameter tuning.

Basic LSTM Model:

The first model is a simple recurrent Neural Network, specifically as LSTM model, and its main purpose is to capture sequential patterns in emails to determine phishing or normal emails. The model processes each word in the email body and keeps track of memory from previous words to create patterns and relationships between each word. LSTM is a good way to detect sequences in emails and is a great method for text classification however the basic LSTM model relies solely on the text data and no other input or feature. It performs well on only the text data however the theory is that by adding an additional feature to capture an emotional tone of the email the results could be better. The inclusion of an emotional tone can help differentiate phishing emails and legitimate emails in the text classification portion of the model.

LSTM with Sentiment Analysis (Novel Approach):

The idea to overcome the basic LSTM model is to enhance the model with a sentimental analysis as an additional feature in the model. The traditional model processes emails sequentially but potentially misses the extra hints within the AI generated emails of emotional tone. In phishing emails the tone is an important feature that can be used to help determine if an email is a scam or not. Phishing emails try to trigger an emotional tone like fear, urgency, and excitement to manipulate the user to make a hasty decision.

The novelty of this approach of sentiment analysis is a preprocessing step to the email text. Using sentiment polarity scores, like positive, negative, or neutral, get extracted from the email body using a sentiment analysis model, and these scores are then normalized. The normalized sentiment scores are combined with the LSTM’s textual embeddings as an additional feature during the classification process. This use of sentiment analysis allows the model to factor in the emotional tone of the email, adding a layer of context that can improve the model’s ability to distinguish between legitimate and phishing emails. Ferrara, E. (2024), says that phishing emails try to evoke strong emotional reactions to manipulate the users actions, making sentiment analysis a valuable feature in detecting these phishing attacks.

The idea is that adding a sentimental feature to the LSTM with sentiment analysis can capture both sequential and emotional aspects of the text to get better results in text classification. Both models will measure accuracy, computational efficiency, and robustness in text classification against phishing emails. The hypothesis is that the LSTM model with sentimental analysis will perform better since it has both sequential and emotional features from phishing emails.

Related Work:

Eze, C. S., & Shamir, L. (2024) talk about how AI generated phishing emails bypass traditional security. Their research paper focuses on deep learning models but does not mention the use of sentiment analysis as a feature in their deep learning models making this approach novel.

Agarwal, A., & Kumar, R. (2024) mention a variety of machine learning models for phishing detection such as SVM, Logistic Regression, and other models but also did not consider using sentiment analysis.

Ige, Tosin, Christopher Kiekintveld, and Aritran Piplai (2024) mention a multi-layer deep learning model but do not consider the use of sentiment analysis. They have tried a multi-layer RNN to gain good accuracy but the use of sentiment analysis as a feature is a new approach that may result in better accuracy.

Experiments:

For this experiment the source is from a kaggle competition that includes CEAS_08 with 39,154 distinct rows, Enron with 29,767 distinct rows, and Ling with 2,859 distinct rows. Each dataset has a label to indicate if it is a scam email or legitimate email. The fist part is to create a baseline Recurrent neural network, LSTM, which is similar to Ige, Tosin, Christopher Kiekintveld, and Aritran Piplai (2024) methodology. Their architecture suggests a multi-layer model and in this experiment a 3 layer LSTM model is used.

The first layer is the embedding layer that consists of a max word count of 20,000, a sequence length of 50, and an embedding dimension of 100. This layer converts input tokens into dense vector representations. The second layer is the LSTM layer that processes sequential data and captures temporal dependencies. This step processes 100 units and uses tanh activation function and sigmoid for the recurrent activation. The dropout rate is 0.2 to help prevent overfitting. The final layer is trained using categorical cross entropy loss function with an Adam optimizer over 5 epochs and a batch size of 32.

The novel approach in this paper to further improve the performance of the LSTM model by including a sentiment analysis as a feature. To create the sentiment scoring the approach used is a polarity score that outputs scores in a range from negative to positive. The scores are normalized and combined with the text embeddings and allow the LSTM model to determine the emotional tone of each email. The goal is to determine emails with a negative tone which indicate that the email has a fearful, urgent, or excited tone. This model should be able to detect these emotions and help identify the phishing emails versus normal emails better.

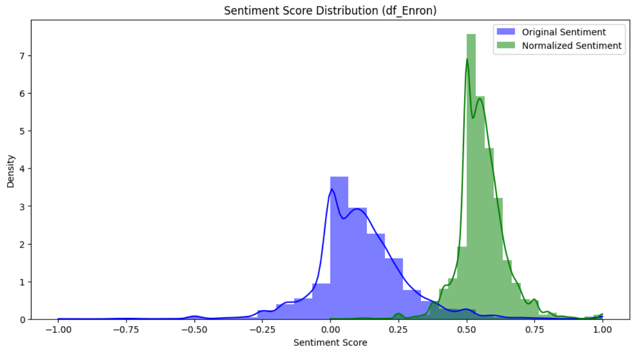
Another enhancement to the model is a RandomizedSearchCV to fine-tune the hyperparameters of the LSTM model. The parameters that are tuned are the number of LSTM units, activation functions, and recurrent dropout rates. To test the sentimental scores we include scaling factors on the sentiment feature to change the influence on the classification if an email is negative or positive. Embedding sizes(50, 100, 200), batch sizes (16, 32, 64), and epochs(5, 10, 15) are also tested in the RandomizedSearch to get the best model with the best performance. The Adam optimizer is also used with the learning rates of 0.001, 0.01, and 0.1 to find the best model combination with all the parameters. This enhancement is crucial as it finds the best parameters to improve results.

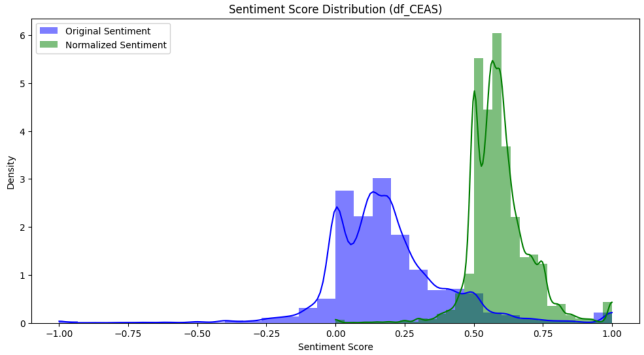
Results and Analysis:

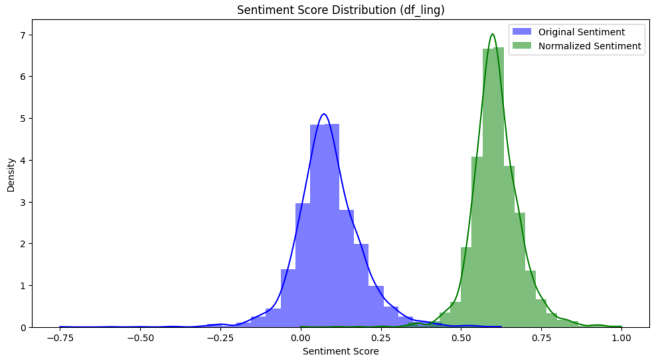
Now that both models are completed to compare the two results we look for the evaluation measures like accuracy, precision, recall, and F1-score.

Both LSTM models performed very well on all three datasets. The accuracy score was around 95% - 99%. The baseline model was computationally efficient at running the model in about 1 minute for each dataset and LSTM took about 2-3 minutes for each. For the Ling dataset baseline has an accuracy of 0.9732 while enhanced version has 0.9807. For the Enron dataset the accuracy was 0.9764 and the enhanced version had an accuracy of 0.9741. For the CEAS dataset the baseline had an accuracy of 0.9886 and the enhanced version had 0.9937. Overall the enhanced novelty did better in 2 out of the 3 datasets and performed equally on Enron's dataset.

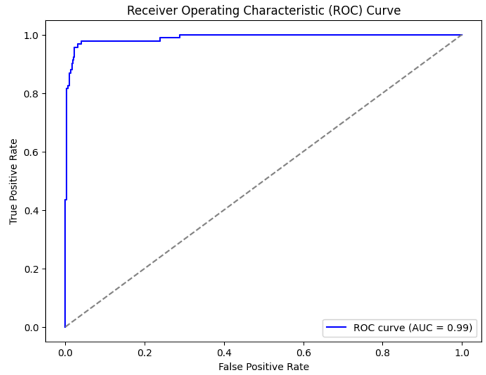
The graphs below represent original sentiment scores versus the normalized sentiment scores.







ROC Curve: 0.99



Real-World Applications:

Companies most likely already have a system that uses either Logistic Regression or keyword matching as their cybersecurity for phishing attacks. For example google most likely uses logistic regression to classify phishing emails but AI generated emails are better at bypassing the simple systems Eze, C. S., & Shamir, L. (2024). A company like Outlook can consider using a sentiment enhanced LSTM model to better classify phishing emails especially for organizational clients that use their platforms. This method would be better as a long term defense since the model can pick up these emotional cues as warnings. This experiment aims to show that a sentiment analysis scoring model can outperform the baseline model when detecting phishing attacks when dealing with real-world phishing attempts.

Conclusion:

Overall both models are a great resource to create a cyber security text classification model but the enhanced sentiment scoring LSTM model has a competitive edge in accuracy. This model gives better classification when it comes to more sophisticated AI phishing attacks. Organizations should consider using a LSTM related security against phishing attacks and smaller projects/companies should look to use the baseline model if they want a quick and efficient model for phishing detection.

References:

Agarwal, A., & Kumar, R. (2024). Curated Datasets and Feature Analysis for Phishing Email Detection with Machine Learning.

Eze, C. S., & Shamir, L. (2024). Analysis and Prevention of AI-Based Phishing Email Attacks. arxiv.org/html/2405.05435v1.

Ige, T., Kiekintveld, C., & Piplai, A. (2024). Deep Learning-Based Speech and Vision Synthesis to Improve Phishing Attack Detection through a Multi-layer Adaptive Framework. arXiv preprint arXiv:2402.17249.

Altwaijry, N., Al-Turaiki, I., Alotaibi, R., & Alakeel, F. (2024). Advancing Phishing Email Detection: A Comparative Study of Deep Learning Models. Sensors, 24(7), 2077.

Ferrara, E. (2024). GenAI Against Humanity: Nefarious Applications of Generative Artificial Intelligence and Large Language Models. Journal of Computational Social Science, pages 1–21.

Data Source: KaggleLink

Baseline: **NLP_Project.ipynb**

Novelty: **NLP_Project2.ipynb**