自然言語処理を用いたフィルタリング手法の提案

Proposal of a spam mail filtering method using natural language processing

脇谷 峡平・ネットワーク分科会・情報セキュリティ大学院大学

In recent years, the number of victims by phishing e-mails has been on the increase. According to the “10 Major Security Threats 2023 (IPA),” phishing continues to rank first in the list, and is expected to increase in the future. In addition, phishing emails are becoming more diverse with the advent of generative AI, etc. Recorded Future re- ported that ChatGPT may be used for cybercrime activities such as creating phishing messages. In this study, we propose and experiment a filtering method as an approach to deal with the increasing and diversified phishing messages. The experiments include training and filtering verification of ALBERT, BERT, Bidirectional LSTM and LSTM natural language processing on an email dataset. The results showed that BERT achieved the highest accuracy of 99.55% F1score, and the validation results showed that filtering was possible with 90% accuracy, and even sentences created by a generative AI could be filtered with 90% accuracy.

背景

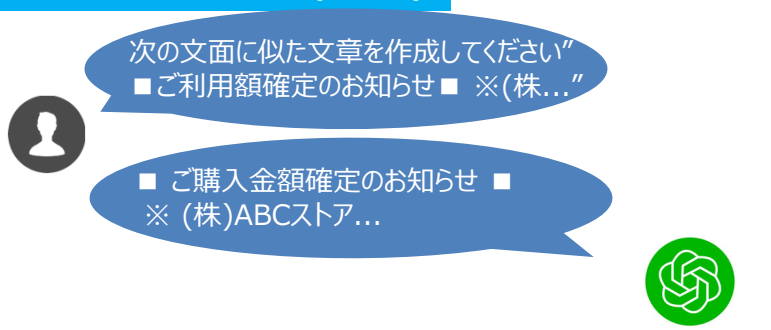
・フィッシングメールによる被害が増加傾向
情報セキュリティ10大脅威2023によるとフィッシングによる被害は昨年に引き続きランキング一位であり
→今後もフィッシングメールが増加

・フィッシングメールが多様性を帯びている傾向
Recorded Futureは“ChatGPTがフィッシングメッセージを作るサイバー犯罪行為に用いられる可能性がある”と報告
SMSフィッシングも昨年に引き続き流行
→フィッシングの種類が多い

目的

対策として4種の自然言語処理を用いたフィルタリング手法を提案と検証実験を行う

検証文の作成方法(一例)



メール文5件
SMS文5件
累計10件で応答された文章をベクトル化
フィルタリング検証を行う

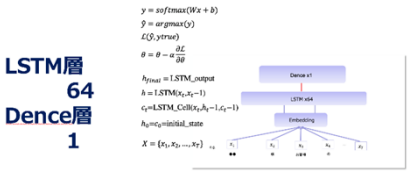
結果

検証結果

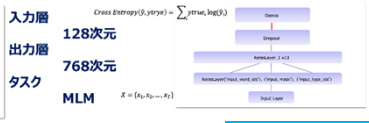
モデル一覧	判定結果
LSTM	40%
Bidirectional LSTM	50%
BERT	90%
ALBERT	80%

提案手法

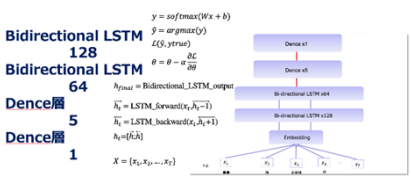
LSTM



BERT



B-LSTM



ALBERT

分類学習・閾値からフィルタリング

まとめ・課題

-まとめ-
・自然言語処理を用いてフィルタリング手法の提案を行った
・生成系AI・SMS・フィッシングメールで検証を行った
-課題-
・更なる精度向上の余地と追検証
・フィッシング対象が変更した際のデータセットの更新