E メールセキュリティサービスをすり抜ける 不審メールの傾向と対策の検討

提出年月日　　2022 年 9 月 29 日

執筆者
いとう みねゆき
伊藤 峰行
株式会社エネルギア・コミュニケーションズ
情報システム事業本部
IT インテグレーション部

原稿量

本文	8,800 字
要約	1,100 字
図表	11 枚

＜キーワード＞
セキュリティ対策、E メール、機械学習、自然言語処理、オープンソースソフトウェア(OSS)

＜要約＞

1. 背景と目的
　近年、E メールを媒介とするセキュリティの脅威が深刻化しており、フィッシングメールやマルウェア添付メール等、悪意あるメール(不審メール)による被害が大きな問題となっている。
　㈱エネルギア・コミュニケーションズでは、不審メールを検知・遮断する「E メールセキュリティサービス」を導入しており、毎月 2 万通以上を遮断しているが、攻撃者側もセキュリティ対策を回避するために様々な手法を用いるため、不審メールを遮断できない「すり抜け」が一定数発生している。
　現在、当社で発生しているすり抜けの多くはフィッシングメールである。フィッシングメールは不特定多数の宛先に繰り返し送付される傾向にあり、時間をおいて何度もすり抜けが発生するため、いち早く発見し遮断する必要がある。しかし、E メールセキュリティサービスに記録されるログは 1 日当たり数万通以上と膨大であり、目視による調査では対応が間に合わない。そのため、現在はすり抜けに気付いた受信者自身からの通報に基づいて調査・遮断を実施しているが、この方法では確実にすり抜けメールを発見できる保証が無いという問題があった。
　受信者に頼らず、すり抜けメールを発見する手法が必要となっていたため、E メールセキュリティサービスのログから不審メールを抽出する分類システムの構築を行うこととした。

2. システム概要・キーポイント
　オープンソースの機械学習エンジン「scikit-learn」を用いて、既知の不審メールと似た特徴を持つ未知の不審メールを、E メールセキュリティサービスのログから発見するシステムの構築を目指した。
　市販の E メールセキュリティサービスで検知できないすり抜けメールを発見するため、学習データには過去に当社で発見したすり抜けメールを用いて、不審メールの件名に使用される単語(約 100,000 個)を、6 種類の機械学習アルゴリズムで学習することで、分類精度の向上と見落とし(偽陰性)件数の低減に取り組んだ。
　また、計算コストの重い機械学習処理を何度も実行するため、自然言語処理によりメールの件名から重要度の低い品詞を除去することで、学習データを削減し学習効率の向上を図った。

3. 有効性評価
　実際のメールログを用いて、システムが未知の不審メールを特定する精度を確認した。システムの判定結果と人による調査結果を比較した結果、91%の精度で不審メールを特定することができた。加えて、不正解となった 9%の内、不審メールを見落とす「偽陰性」は 0 件であり、少量のサンプルデータではあるが、大量のログから不審メールを抽出するという本システムの目的に適う結果となった。

　機械学習は難易度が高いというイメージがあったが、OSS を利用することで比較的容易に実装することができた。今後は本取り組みから得た知見を活かして、さらなるセキュリティ対策の向上を進めていきたい。

目 次

1. はじめに

近年、Eメールを悪用したセキュリティリスクが大きな問題となっている。E メールを悪用する脅威を防ぐには、悪性なメール(不審メール)が受信者まで届かないよう、メールの配送経路上で内容を検査し遮断を行う「E メールセキュリティサービス」の導入が有効な対策となる。

E メールセキュリティサービスを運用する上で問題となるのが、不審メールを通過させてしまう「すり抜け」と、正規の業務メールを誤って遮断してしまう「過検知」である。「すり抜け」の防止と「過検知」の防止はトレードオフの関係にあり、不審メールの判定条件を厳しくすると過検知が増加し、緩めるとすり抜けが増加してしまう。セキュリティ対策においては、マルウェア感染のきっかけとなる不審メールのすり抜け防止が特に重要だが、業務に支障をきたす過検知の影響も考慮する必要があり、どうしても一定数不審メールのすり抜けが発生してしまう。そのため、E メールセキュリティサービスの運用においては、すり抜けた不審メール(すり抜けメール)を発見し、遮断するブラックリスト設定をいかに早く投入できるかが重要となる。

今回、すり抜けメールの早期発見を目的とし、機械学習と自然言語処理を用いて、メールの件名からすり抜けメールを特定するシステムの構築を行った。機械学習は近年、あらゆる分野で利活用が進んでおり、書籍やライブラリが多数公開されている。本稿では、オープンソースの機械学習エンジン「scikit-learn[1]」と、日本語の形態素解析エンジン「MeCab[2]」を用いた、すり抜けメールの分類に関する取り組みについて述べる。

2. すり抜けメールの傾向と対策の問題

本章では、すり抜けメールの傾向と対策の問題、システム化の検討について述べる。

2.1. すり抜けメールの傾向

現在、株式会社エネルギア・コミュニケーションズ(以下、当社)には日々大量の不審メールが送りつけられており、E メールセキュリティサービスによって、毎月約 2 万通以上の不審メールを遮断している。

当社が運用する E メールセキュリティサービスをすり抜ける不審メールの多くは、クレジットカード会員サイトや通販サイトのアカウント情報を窃取することを目的とした、フィッシングメールである。フィッシングメールは、メール本文や添付ファイルに記載された URL リンクから、正規サイトを模した偽サイトに誘導し、騙された被害者が、偽サイト上で入力する ID とパスワードを盗む攻撃である。

フィッシングメールによって誘導される偽サイトは悪意を持って作られているが、ログインの仕組み自体は正規のサイトと同じであり、コンテンツを検査しただけでは検知することが難しい。そのため、フィッシングメール対策においては、過去に発見されたフィッシングメールの情報を基に、送信元メールサーバーや、偽サイトの URL を既知の脅威情報リストと照合する「レピュテーション(評価)機能」が用いられる。しかし、この方法では新しく作られたメールサーバーや URL には対応できず、一定期間不審メールのすり抜けが発生してしまうという問題がある。

2.2. すり抜けメール対策の問題

現在当社では、すり抜けメールに気づいた受信者自身からの通報と、外部の専門機関(IPA 等)から提供される情報に基づいて、すり抜けメールの特定を行っている。しかし、受信者からの情報提供は不審メールに気づけるかどうかにかかっており、確実に発見できる保証はない。また、外部の専門機関から提供される情報は、発見されてから時間が経っているものが多く、情報の鮮度に問題があった。すり抜けメールを迅速に発見し遮断するためには、情報提供を待つという受動的な対応ではなく、より能動的に不審メールを発見する仕組みが必要となっていたため、今回 E メールセキュリティサービスのログから不審メールを特定する、分類システムの検討を行うこととした。

2.3. システム化の検討

システム化の検討にあたっては、以下の条件を満たす必要があると考えた。

(1)E メールセキュリティサービスをすり抜けた
　未知の不審メールを発見可能なこと。
(2)業務情報を含むメール情報を外部に漏洩
　しないこと。

本システムは、E メールセキュリティサービスが検知できなかった未知の不審メールを発見することを目的としており、これには機械学習が有効であると考えた。機械学習を用いたスパムメール[3]分類は、すでに多く

[1] 様々な機械学習アルゴリズムを利用することができる Python 用機械学習ライブラリ

[2] 日本語の文章を「それ以上分割できない最小単位(形態素)」に分割し、品詞や語尾の変化を識別するツール

[3] 受信者の合意なく一方的に送り付けられる迷惑メールの総称

のセキュリティサービスで実装されており、当社の使用する E メールセキュリティサービスにも実装されている。しかし、一般的な E メールセキュリティサービスが行う機械学習は、世界中で広く観測されている不審メールを学習対象としており、必ずしも当社に送付された不審メールの特徴を捉えているわけではない。

当社では、過去数年間にわたり受信者から提供されたすり抜けメールの情報を蓄積していたため、このデータを基に機械学習を行うことで、より精度の高い分類を行えるのではないかと考えた。

加えて、E メールセキュリティサービスのログには業務情報が含まれているため、既存のクラウドサービスではなく、ローカル環境で動作する独自のシステムを構築することとした。

2.4. ログの検討

次に、機械学習で扱うデータの選定を行う。E メールセキュリティサービスのログの内、送信者アドレスや送信元メールサーバーは頻繁に変更されるため、学習対象には適していない。一方、現在確認されているすり抜けメールの多くはフィッシングメールであり、件名には通販サイトやクレジットカードの名称が記載されている。また、受信者にメールを開封させるため、「至急」や「重要」など興味を引くような文言となっていることも多いため、不審メールの特徴が表れていると考え、メールの件名を使用することとした。

なお本文にも同様の特徴が表れていると思われるが、E メールセキュリティサービスのログには本文が記録されないため使用することはできなかった。

3. 不審メール分類システムの構築

本章では、機械学習と自然言語処理による不審メールの分類について記載する。なお、本稿に記載する業務メールの件名はすべて架空のものである。

3.1. 機械学習と自然言語処理の概要

先に述べた通り、現在 E メールセキュリティサービスをすり抜ける不審メールの多くは、アカウント情報の窃取を目的としたフィッシングメールであり、その件名には標的となるクレジットカード会員サイトや通販サイトに関連する単語が使用される傾向にある。本システムでは、機械学習によって過去に当社に送信された不審メールの件名を学習し、同様の特徴を持つ未知のすり抜けメールを発見することを目的とする。

機械学習アルゴリズムは、事前に正しいデータ(学習データ)を学習し、その結果に基づいて予測を行う「教師あり機械学習」と、学習データを用いない「教師なし機械学習」に分けられるが、今回の目的である不審メールの発見には、あらかじめすり抜けメールの件名を学習させる「教師あり機械学習」が適している。

3.2. 学習データの準備

教師あり機械学習では、学習データの量と多様性によって分類精度が左右されるため、いかに豊富な学習データを用意するかが重要となる。今回、学習データの作成には、(1)利用者から提供されたすり抜けメールの件名，(2)外部の専門機関から情報提供された不審メールの件名，(3)E メールセキュリティサービスが遮断した不審メールの件名の、計 13,500 件を用意した。

3.3. 前処理

機械学習では計算によって予測を行うため、日本語の文章(自然言語)であるメールの件名を単語に分解して解析する「形態素解析」と、個々の単語を数値データに変換する「ベクトル化」を行う必要がある。本節では、形態素解析とベクトル化について述べる。

3.3.1. 形態素解析

日本語の文章は、英語のように単語と単語の間に区切り文字(スペース)が含まれておらず、そのままでは文章全体が一塊とみなされてしまうため、自然言語処理において、文章を単語ごとに分割する「形態素解析」を行う必要がある。オープンソースの形態素解析エンジン「MeCab」を用いた単語分割の例を表 1 に示す。

表 1. 形態素解析の例

入力	もっと輝く，確かな明日へ	
出力	もっと	副詞,一般,*,*,*,*,もっと,モット,モット
	輝く	動詞,自立,*,*,五段・カ行イ音便,基本形,輝く,カガヤク,カガヤク
	，	記号,読点,*,*,*,*,、,、,、
	確か	名詞,形容動詞語幹,*,*,*,*,確か,タシカ,タシカ
	な	助動詞,*,*,*,特殊・ダ,体言接続,だ,ナ,ナ
	明日	名詞,副詞可能,*,*,*,*,明日,アシタ,アシタ
	へ	助詞,格助詞,一般,*,*,*,へ,へ,エ

入力された文章が単語ごとに分割して出力されていることが確認できる。続いて、使用する単語の選別を行う。

教師あり機械学習において、学習データが増えるほど計算に必要なCPUリソースが必要になるため、重要度の低いデータはあらかじめ取り除いておくことが望ましい。本システムでは、不審メールの件名から不要な単語を除去するため、件名に使用されている各品詞のバリエーションを確認することとした。

日本語は品詞によって単語の種類に大きな偏りがあり、名詞のように種類が非常に豊富な品詞と、助詞や接続詞など種類の少ない品詞がある。多くのメールで共通的に使用される単語は、不審メールの分類においては重要ではないと考え、学習データとして用意した不審メールの件名 13,500 件に対し、各品詞の出現回数と単語数を確認した。その結果を表 2 に示す。

表 2. 不審メール件名に使用される品詞の分析

品詞	出現回数	単語数	比率
名詞	**74,576**	**14,021**	**18.8%**
代名詞	9,758	29	0.3%
動詞	**11,974**	**999**	**8.3%**
形容詞	**9,786**	**160**	**1.6%**
形容動詞	9,736	1	0.0%
連体詞	9,738	19	0.2%
副詞	**9,891**	**202**	**2.0%**
接続詞	9,739	16	0.2%
感動詞	9,744	37	0.4%
助詞	19,003	69	0.4%
助動詞	11,244	62	0.6%

確認の結果、「名詞」「動詞」「形容詞」「副詞」の 4 種類以外の品詞は、出現回数に比べて単語数が少なく、同じ単語が繰り返し使用されていることが分かった。この結果に基づき、実際に不審メールの件名から 4 種類の品詞を抽出した結果を表 3 に示す。

表 3. 件名から特定の品詞を抽出した例

変更前	**【重要】あなたのアカウント情報が漏洩しました。**

項番	1	2	3	4	5	6	7
単語	**【**	**重要**	**】**	**あなた**	の	**アカウント**	**情報**
品詞	記号	**名詞**	記号	**名詞**	助詞	**名詞**	**名詞**

項番	8	9	10	11	12	13
単語	が	**漏洩**	**し**	まし	た	。
品詞	助詞	**名詞**	**動詞**	助動詞	助動詞	記号

変更後	**重要**	**あなた**	**アカウント**	**情報**	**漏洩**	**し**

元の文章と比較して単語数は 13 個から 6 個に削減されているが，「重要」「アカウント」「漏洩」などメールの内容に関わる単語は残されていることが確認できる。なお、形態素解析ツールが行う品詞の分類には一部誤りもあるが、今回はそのまま使用している。

続いて、抽出した単語同士の組み合わせについて検討する。形態素解析と品詞の選択により、必要な単語のみを洗い出したが、個々の単語をバラバラに学習しても、元の文章にあった文脈が無視されてしまい、精度の高い予測を行うことができない。そのため、自然言語処理では、隣り合う単語同士の組み合わせを切り出す「N グラム」という手法が用いられる。 N グラムでは、隣り合った N 個の単語(文字)を一纏まりの言葉として捉え、その言葉が文章中に何度出現するかを求める。不審メールの件名を用いて隣り合う単語を組み合わせた例を表 4 に示す。

表 4. 隣り合う単語を組み合わせた例

1語	重要	あなた	アカウント	情報	漏洩	し
2語の組み合わせ	重要	あなた				
		あなた	アカウント			
			アカウント	情報		
				情報	漏洩	
					漏洩	し
3語の組み合わせ	重要	あなた	アカウント			
		あなた	アカウント	情報		
			アカウント	情報	漏洩	
				情報	漏洩	し
4語の組み合わせ	重要	あなた	アカウント	情報		
		あなた	アカウント	情報	漏洩	
			アカウント	情報	漏洩	し

複数の単語を組み合わせることで「アカウント情報」や「情報漏洩」「アカウント情報漏洩」等、単語のみの場合よりも、フィッシングメールの特徴が表れている。ただし、4 語の組み合わせの場合は、3 語までの組み合わせと比べて特別な優位性は見られなかったため，本システムでは 3 語までの組み合わせを用いることとした。

形態素解析の最後に、品詞の選択による学習データの削減効果を確認する。学習データとして用意した不審メールの件名 13,500 件を形態素解析によって分割し、 隣り合う 3 語までの組み合わせを考慮する場合、学習データは 117,606 個となるが、品詞を「名詞」「動詞」「形容詞」「副詞」の 4 種類のみに限定することで学習データは 101,576 個となり、16,030 個(11.5%)の削減となった。

3.3.2. ベクトル化

2 つ目の前処理として、単語(文字)を、機械学習で処理可能な数値(ベクトル)に変換する「ベクトル化」を行う。単語のベクトル化には複数の手法があるが、今回は最も基本的な「One hot 表現(Bag Of Words)」を選択した。One hot 表現は、文章中に使用されている単語の出現頻度を数値(ベクトル)に変換する方式であり、自然言語を扱う機械学習ではメジャーな手法の

ため、情報の入手が容易で機械学習の入門に適している。

オープンソースの機械学習エンジン「scikit-learn」には、 One hot 表現によるベクトル化を行うための関数が用意されており、単語の出現頻度をカウントする「Countvectrizer」と、単語のレア度を評価する「Tfidfvectrizer」を使用することができる。Countvectrizer はシンプルに単語の出現回数のみをカウントし、Tfidfvectrizer による単語のレア度評価では、「ある文章中に何度も出てくる単語は、その文章の特徴を表している」が「多くの文章で共通して使用されている単語は重要ではない」という考え方に基づいて数値化する。Countvectrizer(出現頻度)と Tfidfvectrizer(レア度)による数値化の例を表 5 に示す。

■変換前の文章
・文章 A:広島は牡蠣と穴子と鯛が美味しい
・文章 B:広島はレモンとブドウとミカンが美味しい

表 5. 出現頻度とレア度による数値化の例

Countvectrizer(出現頻度)

単語	文章A	文章B
広島	1	1
牡蠣	1	0
穴子	1	0
鯛	1	0
レモン	0	1
ブドウ	0	1
ミカン	0	1
は	1	1
と	2	2
が	1	1
美味しい	1	1

Tfidfvectrizer(レア度)

単語	文章A	文章B
広島	0.26	0.26
牡蠣	0.37	0
穴子	0.37	0
鯛	0.37	0
レモン	0	0.37
ブドウ	0	0.37
ミカン	0	0.37
は	0.26	0.26
と	0.53	0.53
が	0.26	0.26
美味しい	0.26	0.26

Countvectrizer では、単純に 1 文中に唯一 2 回出現している「と」が最も高い「2」となっており、その他の単語はすべて「1」ないし「0」となっている。一方、Tfidfvectrizer では、片方にしか出現しない「牡蠣」や「レモン」のスコア(0.37)が、2 つの文章で共通して出現する「広島」や「美味しい」のスコア(0.26)よりも高くなっており、文章固有の単語が重視されていることがわかる。なお、この例では Tfidfvectrizer においても「と」が最も高いスコアとなっているが、これは文章数と単語数が少ないためであり、文章の数や一文の長さが増えるほど、共通的に出現する単語のスコアは低く判定される。

これら 2 種類のベクトル化方式のうち、どちらが不審メールの分類に適しているかを確認する。ニュース記事や小説など、ある程度の長さを持った文章を機械学習で分類する場合、文章中で繰り返し使用される、助詞や接続詞を除外できる Tfidfvectrizer を用いた方が、Countvectrizer よりも分類精度が高くなる傾向にある。ただし、不審メールの件名には通販サイトの名称など特定の単語が多く使われているため、Tfidfvectrizer ではこれらの単語も不要とみなされてしまう可能性がある。加えて、本システムでは、あらかじめ不要な品詞を排除しており、最も重複していると思われる「助詞」や「接続詞」はデータに含まれていない。そのため、Countvectrizer による変換でも有効な結果を得られる可能性があると考え、2 種類の変換方法を比較することとした。不審メールの件名を 2 種類の方式で数値化し、重要度が高いと判定された単語 10 個を降順に並べた結果を表 6 に示す。

表 6. 出現頻度とレア度で抽出した単語上位10 個

Countvectrizer

順位	単語	順位	単語
1	研究	6	web
2	写真	7	出口
3	知らせ	8	free
4	知ら	9	from
5	凄い	10	warning

Tfidfvectrizer

順位	単語	順位	単語
1	ポイント	6	your
2	人気	7	the
3	アカウント	8	セール
4	商品	9	する
5	am*zon	10	オン

処理結果を見ると、いずれの方式でもフィッシングメールの件名に用いられる、「am*zon」や「アカウント」「warning」といった単語が抽出されており、2 種類とも有効に機能していることが確認できる。この結果に基づき、本システムでは Countvectrizer と Tfidfvectrizer を併用することとした。

3.4. 機械学習によるメール分類

ここからは、機械学習による予測について述べる。機械学習が行う予測には「分類」と「回帰」の 2 種類がある。「分類」はあらかじめ人の手で正しく分類されたデータに基づいて、新しく与えられたデータのラベルを予測する。「回帰」は連続する数値データから、将来の数値を予測するために用いられる。今回の目的であるすり抜けメールの特定では、メール件名に対して「不審メール」か「正規メール」いずれかのラベルを付与するため、「分類」処理となる。

「scikit-learn」には分類に使用できる分類器(関数)が複数用意されており、数値データを渡すだけで簡単に予測を行うことができる。処理対象となるデータの件数と種類に応じて、どのアルゴリズムを使用すればよいかは、scikit-learn のドキュメント[4] で公開されており、ドキュメントに基づくと「学習データが 100,000 件未満」の場合は、まず「線形 SVC」を用いて効果を確認

[4] scikit-learn Choosing the right estimator https://scikit-learn.org/stable/tutorial/machine_learning_map/

し、期待する結果が得られない場合は、対象がテキストデータであれば「ナイーブベイズ」を、それ以外の場合は「K 近傍法」「カーネル SVC」「アンサンブル法(ランダムフォレスト)」のいずれかを使用すればよいとある。今回用いる学習データは約 100,000 個のテキストデータであり、「線形 SVC」か「ナイーブベイズ」が適しているということになるが、本当にその分類器が適しているかを確認するため、各分類器の分類精度を比較することとした。

評価するアルゴリズムは、上記 5 種類に、ディープラーニングで用いられる「ニューラルネットワーク(多層パーセプトロン)」を加えた 6 種類とし、公式ドキュメントで推奨されている「ナイーブベイズ」には 2 種類の分類器を用いて検証を行った。検証に用いた分類器を表 7 に示す。

表 7. 検証を行った分類器の一覧

分類器	概要
ベルヌーイナイーブベイズ	結果を1/0で表現するベルヌーイ分布で特徴量を表し、ある事象が起きた場合に条件Aが前提である確率(ナイーブベイズ)に基づいて分類する方式。
多項ナイーブベイズ	複数の結果を表現する多項分布で特徴量を表し、ある事象が起きた場合に条件Aが前提である確率(ナイーブベイズ)に基づいて分類する方式。
K近傍法	ベクトル空間内で距離の近いK個の学習データの多数決により未知データのクラスを分類する方式。
線形SVC	平面上に分布する学習データに対して、直線で各クラスの境界を求める分類方式。
カーネルSVC	直線で境界を表せないサンプルに対して，既存の特徴量から求められる異なる特徴量を追加することで，次元を増やして非線形の境界を求める分類方式。
ランダムフォレスト	決定木(条件分岐の組み合わせ)を複数作成し、決定木の多数決によりクラスを決定する分類方式。
多層パーセプトロン	入力値と重みづけにより1/0の結果を出力する演算手法(パーセプトロン)を階層的に組み合わせ、最終的な結果と正解値の誤差を元に、重みづけの調整を行って精度を向上させる方式。

なお、各分類器の動作原理を説明するには、筆者の知識が不足しているため、本稿では結果の比較のみを行っている。

分類器の評価においては、既知の不審メールの内、学習データとして使用していない件名を用いて、分類器による予測結果の正答率を確認すると共に、混同行列を用いて内訳を確認した。混同行列は、予測結果を「真陽性」「真陰性」「偽陽性」「偽陰性」の 4 項目で表現することができる。本稿では以下のように混同行列を定義し、各分類器の比較を行った。

・真陽性:「不審メール」を「不審メール」と正しく判定
・真陰性:「正規メール」を「正規メール」と正しく判定
・偽陽性:「正規メール」を「不審メール」と誤って判定
・偽陰性:「不審メール」を「正規メール」と誤って判定

本システムでは大量のメールログの中から、不審メールを抽出することを目的としており、不審メールを正規メールと誤って判定する「偽陰性」は可能な限り低く抑えることが望ましい。そのため、分類器の選定においては、精度に加えて、偽陰性件数の低さも重要であると考えた。各分類器の精度を表 8 に示す。

表 8. 各分類器の評価結果

Countvectrizer(出現頻度)を用いた分類結果

分類器	真陽性	真陰性	偽陽性	偽陰性	精度	偽陰性割合
ベルヌーイNB	916	904	50	42	**95.2%**	2.2%
多項NB	913	902	52	45	**94.9%**	2.4%
K近傍法	807	602	352	151	73.7%	7.9%
線形SVC	917	888	66	41	94.4%	2.1%
カーネルSVC	922	868	86	36	93.6%	**1.9%**
ランダムフォレスト	907	880	74	51	93.5%	2.7%
多層パーセプトロン	906	905	49	52	**94.7%**	2.7%

Tfidfvectrizer(レア度)を用いた分類結果

分類器	真陽性	真陰性	偽陽性	偽陰性	精度	偽陰性割合
ベルヌーイNB	926	865	89	32	93.7%	**1.7%**
多項NB	916	880	74	42	93.9%	2.2%
K近傍法	934	571	383	24	78.7%	1.3%
線形SVC	916	879	75	42	93.9%	2.2%
カーネルSVC	924	892	62	34	95.0%	**1.8%**
ランダムフォレスト	907	874	80	51	93.1%	2.7%
多層パーセプトロン	900	895	59	58	93.9%	3.0%

評価の結果、ほとんどの分類器で精度 90%以上となっている中、K 近傍法のみ 70%台となっており、不審メールの抽出には適していないことが分かった。また、その他の分類器は精度には大きな差が出ていないが、偽陰性の割合をみると、若干の差が出ていることが確認できた。この結果に基づき、本システムでは以下 6 種類の組み合わせによる分類を行うこととした。

(1) Countvectrizer＋ベルヌーイナイーブベイズ
(2) Countvectrizer＋多項ナイーブベイズ
(3) Countvectrizer＋カーネル SVC
(4) Countvectrizer＋多層パーセプトロン
(5) Tfidfvectrizer ＋ベルヌーイナイーブベイズ
(6) Tfidfvectrizer ＋カーネル SVC

3.5. スコアリング

最後に、各分類器から得られた判定結果をもとに、抽出されたメールがどの程度「不審」と判定されたかを表すスコアリングを行う。各分類器の判定結果は「不審メール」か「正規メール」の 2 択で出力されているが、分類器によって評価が分かれたメールについては、どの程度差がでているかを確認するため、6 種類の分類方法の分類精度を用いて重みづけを行った。

正規メールと判定された場合はスコア「0」、不審メ

ールと判定された場合は、判定した分類器の分類精度(0.94～0.95)をスコア値に加算し、複数の分類器から得られたスコアの合計値を、そのメールの最終的なスコアとする。本システムにより実際のメールログを分類した結果を表 9 に示す。

表 9. 分類システムによるスコアリング結果

No	メール件名/分類方法	(1)	(2)	(3)	(4)	(5)	(6)	合計
1	Am*zonアカウントは停止されました、個人情報を更新してください	**0.95**	**0.95**	**0.94**	**0.95**	**0.94**	**0.95**	**5.68**
2	【重要】異常ログイン通知	**0.95**	**0.95**	**0.94**	**0.95**	**0.94**	**0.95**	**5.68**
3	【重要なお知らせ】カード情報更新のお知らせ！番号：034297808	**0.95**	**0.95**	**0.94**	**0.95**	**0.94**	**0.95**	**5.68**
4	【Cloud PARK】新しいお知らせが届いています	0	**0.95**	**0.94**	**0.95**	**0.94**	0	**3.78**
5	ルール更新通知	0	0	**0.94**	0	0	0	**0.94**
6	[CLIP]XXXエナジー、日豪で次世代供給網構築／現地企業と協業へ	0	**0.95**	0	0	0	0	**0.95**
7	[2021-6] 新入社員自己紹介	0	0	0	0	0	0	0
8	[ご依頼]2021年度実績に関する情報(2022年度計画) 送付について	0	0	0	0	0	0	0
9	お問い合わせをお受けいたしました。■■市 地域安全課	0	0	0	0	0	0	0

不審メール(No1～3)には高いスコアがつけられ、正規メール(No7～9)には低いスコアがつけられていることがわかる。また、No4～6 のように、不審メールか正規メールかが曖昧な件名については、分類器による評価が分かれ、どの程度怪しいと判定されたかを数値として表現することができた。

4. システムの有効性確認

本章では、システムの有効性を確認した結果について述べる。

4.1. 性能評価

本システムが、未知の不審メールに対してどの程度有効なのかを評価する。既知の不審メール件名を用いた検証では 93%以上の精度となっていたが、未知の不審メールをどの程度発見できるのかを確認するため、無作為に抽出したEメールセキュリティサービスのログを用いて不審メールの分類を行い、人による判定結果と、本システムのスコア値を比較することとした。

人が行う判定では、件名の特徴だけでなくメールサーバーの所在地や管理組織名等を、外部のセキュリティ機関が公開する脅威情報に照らして確認することで、より詳細に判定することができる。実際に E メールセキュリティサービスのログから無作為に抽出したメール件名 100 件を、システムと人がそれぞれ判定する検証を行った結果、正答数 91 件、偽陽性件数 9 件、偽陰性件数 0 件となった。本システムにおいては判定精度に加え、偽陰性を低く抑えることが重要になるが、いずれも実用的な数値を出すことができたと考える。

4.2. 分類精度の向上

最後に，本システムの分類精度について考える。既知の不審メールを用いた検証では、94%～95%の精度を出せていたが、未知の不審メールに対する判定では 91%(91 件)と、わずかに低くなった。これは実際のメール件名には、学習データに含まれていない単語が多数含まれており、分類器が評価できない未知の単語に影響を受けたものと考えられる。教師あり機械学習においては、予測対象となるデータに対して、じゅうぶんな量と多様性を備えた学習データを用意することが重要であり、これが不足していた場合に分類精度が低下してしまう。今回、学習データとして不審メールの件名約 13,500 件を用意したが、当社で実際に送受信されるメールを分析するには、学習データが不足していたと考えられる。分類精度の向上には、学習データの拡充が有効な対策となるため、今後の運用において、受信者から提供されたすり抜けメールの件名を学習データに追加していく必要があることが分かった。

5. おわりに

機械学習と自然言語処理を用いて、E メールの件名から不審メールを発見する取り組みにより、これまで受信者や外部機関からの情報提供を待つしかなかったすり抜けメールの発見と遮断を、より能動的且つ迅速に行うための筋道を立てることができたと考える。

今回構築したシステムは、機械学習としては基礎的な内容であり、まだまだ分からないことの方が多いが、対象を限定し有効な学習データを用いれば、オープンソースの機械学習エンジンでも実用的なシステムを構築できることが分かった。

今後は機械学習アルゴリズムに対する理解を深め、さらなる精度の向上に取り組んでいくと共に、本システムを実際の業務で活用することで、より一層のセキュリティレベル向上を図っていきたい。

本稿が同様の課題を抱える組織の一助になれば幸いである。

参考文献

[1]Python ではじめる機械学習 － O'Reilly Japan、2017 年 5 月 初版発行
[2]Scikit-learn Machine Learning in Python 参考 URL：https://scikit-learn.org/stable/, 2022.9.15
[3]MeCab: Yet Another Part-of-Speech and Morphological Analyzer 参考 URL：https://taku910.github.io/mecab/, 2022.9.15
[4]IPA 情報セキュリティ 10 大脅威 2021 参考 URL：https://www.ipa.go.jp/security/vuln/10threats2021.html, 2022.9.15

ワークショップ、セッション、および資料は、IBM またはセッション発表者によって準備され、それぞれ独自の見解を反映したものです。それらは情報提供の目的のみで提供されており、いかなる参加者に対しても法律的またはその他の指導や助言を意図したものではなく、またそのような結果を生むものでもありません。本論文に含まれている情報については、完全性と正確性を期するよう努力しましたが、「現状のまま」提供され、明示または暗示にかかわらずいかなる保証も伴わないものとします。本論文またはその他の資料の使用によって、あるいはその他の関連によって、いかなる損害が生じた場合も、IBM またはセッション発表者は責任を負わないものとします。 本論文に含まれている内容は、IBM またはそのサプライヤーやライセンス交付者からいかなる保証または表明を引きだすことを意図したものでも、IBM ソフトウェアの使用を規定する適用ライセンス契約の条項を変更することを意図したものでもなく、またそのような結果を生むものでもありません。

本論文で IBM 製品、プログラム、またはサービスに言及していても、IBM が営業活動を行っているすべての国でそれらが使用可能であることを暗示するものではありません。本論文で言及している製品リリース日付や製品機能は、市場機会またはその他の要因に基づいて IBM 独自の決定権をもっていつでも変更できるものとし、いかなる方法においても将来の製品または機能が使用可能になると確約することを意図したものではありません。本論文に含まれている内容は、参加者が開始する活動によって特定の販売、売上高の向上、またはその他の結果が生じると述べる、または暗示することを意図したものでも、またそのような結果を生むものでもありません。 パフォーマンスは、管理された環境において標準的な IBM ベンチマークを使用した測定と予測に基づいています。ユーザーが経験する実際のスループットやパフォーマンスは、ユーザーのジョブ・ストリームにおけるマルチプログラミングの量、入出力構成、ストレージ構成、および処理されるワークロードなどの考慮事項を含む、数多くの要因に応じて変化します。したがって、個々のユーザーがここで述べられているものと同様の結果を得られると確約するものではありません。

記述されているすべてのお客様事例は、それらのお客様がどのように IBM 製品を使用したか、またそれらのお客様が達成した結果の実例として示されたものです。実際の環境コストおよびパフォーマンス特性は、お客様ごとに異なる場合があります。