F-050

深層学習を用いた電子メールの自動分類
Automatic Classification of Emails Using Deep Learning

陳 春祥† 下馬場 もえ†
Chun-Xiang Chen Moe Shitababa

1. はじめに

電子メールは、インターネットともに普及したインターネット上のサービスの 1 つであり、今日まで広く使われている。しかし、電子メールの仕様（SMTP）のシンプルさ、利用のしやすさゆえに、メール受信者が望まない広告を目的とした内容や不正プログラムが添付された、メール（いわゆる迷惑メール）が増え、大きな社会問題となっている。電子メールの歴史は迷惑メールとの戦いの歴史でもあると言っても過言ではない[1][5]。

近年、ネットコミュニケーションツールとして SNS（例えば LINE、WeChat、Meta など）やクラウドサービスが高度発達してきた。電子メールによる通信が相対的に減っているように見えるが、ビジネスのシーンでは、電子メールが正式な通知や報告、依頼などに使われ、公式文書としての役割を果たしている。また、SNS やオンラインショッピングのアカウントや多要素認証として電子メールアドレスが使われることも多い。このように、電子メールが依然として主要なコミュニケーション手段となっている。そのため、迷惑メールの捉え方を再考し、的確な対策の方法と対策ツールの開発が求められる。

従来の手法では電子メールを迷惑メール(spam)とそうでないメール(ham)に分類して対策を講じるのが一般的である。しかし、様々な SNS の普及並びにクラウドサービスとオンライン取引の多様化により付随した電子メールはこの 2 択 1 の分類手法では電子メールの内容を捉えきれないと考えられる。

本稿では、迷惑メールであるか否かのみの判別に留まらず、深層学習を利用して電子メールを複数のカテゴリへの分類を試み、分類の手法および解析の結果について報告する。本稿の構成は以下である。2 節で迷惑メール対策の現状について概説する。3 節では深層学習で利用するデータセットの前処理及び学習モデルについて述べる。4 節ではメールを複数カテゴリへの分類精度を示す。また、従来の SpamAssassin の判定結果を比較し、迷惑メールの判定精度を考察する。最後に 5 節でまとめと今後の課題を述べる。

2. 迷惑メール対策技術の現状と課題

2.1 送信者認証による対策

従来の SMTP（Simple Mail Transfer Protocol）では、送信者の認証及び検証する仕組みを設けず、メールを送信する非常にシンプルな仕様である。そのため、送信元の詐称や望ましくない内容のメールであっても送信できてしまう[1]。

こうした問題を解決するため、POP before SMTP が提案された[2]。これはメールを送信する前に POP を利用して送信者を認証し、承認された正規の送信者だったら送信を許可するという仕組みである。

† 県立広島大学/Prefectural University of Hiroshima

また、SMTP を拡張して送信者認証を行う SMTP-AUTH が提案された[3]。これはメールクライアントからメールサーバに対してメール送信を依頼する際に送信者の認証を受けさせ、承認された正規ユーザのみ配信を受け付ける仕組みである。

2.2 サーバ側での対策

送信者認証では迷惑メール送信を不可能にする一方、正規ユーザの送信にもコストがかかってしまう。更に正規ユーザでありながら、認証を通過できないとき、サーバ側の対応やサポートが求められるため、サーバ運用にもコストが高くなる。従って送信者認証を求めず、従来の仕様である SMTP でサーバ側での迷惑メール対策を施すといった形態の運用が最も多い。

メールサーバサイドの対策としては、(1)SMTP セッションでメールクライアントのふるまいによる対策。例えばブラックリストやグレイリスなど[4]がある。(2)送信ドメイン認証による対策。例えば、SPF、DKIM、DMARC などがある[7][8]。(3)メールコンテンツ解析による対策。例えばベイジアン方式のフィルタリング[6]、コンテンツ解析の SpamAssassin が多く活用されている[9]。

しかし、迷惑メール対策とその対策を潜り抜けて迷惑メールを送信する手口はいたちごっこである。特に利用者の心理や活動につけ込んであたかもあったかのような内容のメールを送りつけられると、迷惑メールであるかどうかの自動判別が難しい。ドイツの Statistics 社の発表によると 2023 年に世界で送信されたメールの 45.6%が迷惑メールであった。総務省の発表によると日本国内大手 ISP が受信したメールの約 4 割が迷惑メールである。近年 SNS やクラウドサービス、ネット商などの取引が高度発達し、利用者が一度登録し利用したことのある業者からの広告も含むサービス案内メールは、迷惑メールであるかどうかは、個人の嗜好や価値観に大きく依存する。

そこで、本研究ではメールを迷惑メール(spam)かそうでないか(正常メール ham)に分類するという従前の考え方と異なり、メールの内容とユーザの嗜好に基づき、多層パーセプトロン（MLP）を用いた深層学習でメールを複数のカテゴリへ自動分類を試みる。

3. 機械学習用データセットの作成

機械学習用のオープン電子メールデータセットは日本語文字コードに対応していない[11]ため、本研究では、著者らが日頃蓄積した電子メールを複数カテゴリに分けて学習と判定用データセットとして利用することにした。

3.1 複数カテゴリの考え方

前述のように迷惑メールの判定は個人の嗜好や価値観に大きく依存するため、主観的であるが、著者らの蓄積したメールを以下の職業柄の関係で 4 カテゴリで分類することにした。

(1) spam（迷惑メール）：架空請求、ウィルスメール、フィッシング詐欺メール、標的型攻撃メールなど、ユーザの価値観によらない典型的な迷惑メールをこのカテゴリに分類する。図 1 は詐欺メールの例である。「会員ログイン」の先はアマゾンと関係のないサイトへのリンクになっている。

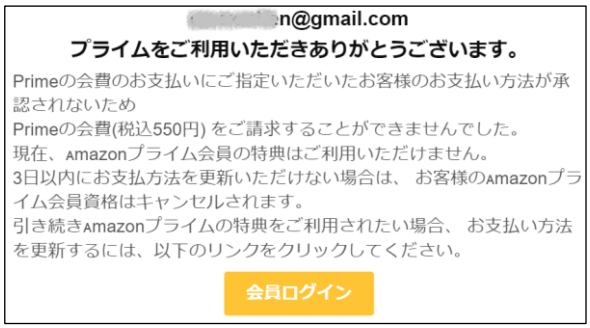

図 1 詐欺メールの例

(2) cm（コマーシャルメール）：ユーザは何らかの形で利用したことがあるような業者やネットサービスサイトからの宣伝メール、サービスの案内メールをこのカテゴリに分類する。

(3) puh（職場関係のメール）：著者らが所属する大学関係の業務上のメールをこのカテゴリに分類する。

(4) cfp（論文募集関係のメール）：職業柄より外部から論文の募集や学会への参加案内のメールをこのカテゴリとする。

3.2 データセットの前処理

本節では、著者らが蓄積したメールを機械学習用のデータセットとしての前処理について述べる。

3.2.1 メールのデコード

蓄積したメールは Maildir 形式で生のまま保管している。中身は基本的に MIME マルチパートになっている。以下の手順に従ってメールをデコードする。

(1) メールの“Content-Type”、“Content-Transfer-Encoding”に応じて、mha-decode と mpack などのツールを用いて MIME 形式のメールをデコードし、各マルチパートの内容に対応したファイルに分離する。

(2) text/html 形式のパートは html のタグを取り除き、テキストファイルに変換する。

(3) 分離されたテキストファイル（text/plain）を UTF-8 コードに変換する。

(4) 今回の機械学習による分類は、元のメールに添付されたバイナリファイル（画像ファイルや pdf ファイルなど）を考慮しないため、テキストファイル以外のバイナリファイルを除外する。

3.2.2 コーパスファイルの準備

デコードしたメールを各カテゴリに対応したフォルダ（spam、puh、cm、cfp）に仕分ける。フォルダ名はカテゴリのラベルになる。そして、学習用のコーパスファイルに変換して 1 つ csv ファイルにまとめる。コーパスファイルの書式は 1 行に 1 件のメールとカテゴリ（学習する際のラベル）で構成される。ただし csv ファイルの区切りと区別するため、メール内容部分は'メール文'で括っている。表 1 にそれぞれのカテゴリの 1 つのサンプルを示す。

表 1 各カテゴリのメールサンプル

'ご利用店名（売場名）：****ガス ガスリヨウキンご利用金額：6、***円初回年月：2023 年 4 月▼ご利用に相違ないか念のためご確認ください▼ご請求明細はこちら：<https://alpus.ebayjo.com>',spam
'教務委員からのお知らせです。情報処理基礎の授業の際にも案内しましたが，経営情報学科 4 年生の卒業論文発表会が以下の日程で開催されますので，再度案内します。日時 ： 2 月 7 日(木) 9:00〜15:30 、8 日(金) 9:30〜12:10 会場 ： 1331 講義室・・・・・・なお，卒業論文発表会の会場は出入り自由ですので，興味のある発表のみ参加することも可能です。卒業論文発表会プログラム http://www.pu-hiroshima.ac.jp/site/management/mis20190131.html',puh
'NanoCom 2024: 11th ACM International Conference on Nanoscale Computing and Communication……ACM NanoCom 2024 is endorsed by theTechnical Committee on Molecular, Biological, and Multi-Scale Communications of the IEEE Communications Society',cfp
'広島県公立大学法人 ご担当者様, Amazon ビジネスでは、急に必要になった 1 回きりの少額の購買だけでなく、部署や会社全体で使うような物品をまとめ買いしたり、金額が大きくなるものは相見積もりを取りたいといったニーズにも対応しています。また、スマートフォンで仕事での購買をスムーズに行えるよう、Amazon ビジネス専用のアプリを用意しています。・・・・・・Amazon ビジネスのご利用を開始するにあたり、以下のようなメールをお送りしてまいりました。Amazon ビジネスによって、お客様が業務を円滑に遂行できることを願っております。ぜひご活用ください。',cm

今回の研究で用いた学習用データと判別用データの数は以下（表 2）の通りである。

表 2 各カテゴリのデータセット

カテゴリ	学習用データ(通)	判別用データ(通)
cfp	2,932	120
puh	9,235	25
spam	8,751	66
cm	2,712	32

4. 分類評価実験

4.1 評価方法

本研究では、シンプルな分類器（sklearn の MLPClassifier()分類器）を用いてメールを複数カテゴリへの分類を行う[12]。また、日本語（UTF-8 コード）の形態素解析は MeCab を、ワードのベクトル化は CountVectorizer()を使用した。因みに HTML 形式のメールは、HTML 関連のタグを取り除かれ、改行コードを削除してその他の特殊記号はそのまま保持して、テキストに変換される。

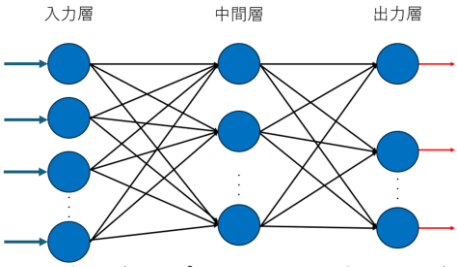

図 2 多層パーセプトロンによる分類イメージ

図 2 に多層パーセプトロンによる分類イメージを示す。以下、中間層及び各層のニューロン数を調整して解析の結果を示す。

4.2 実例解析の結果

多層パーセプトロンで機械学習を行い、判別用データから結果を考察していく。図 3、図 4 の縦軸はニューロン数、横軸は各カテゴリの判定率を表している。まず、中間層が一層の場合でニューロン数を変化していき、次に中間層を複数の場合でニューロン数を変化させた。

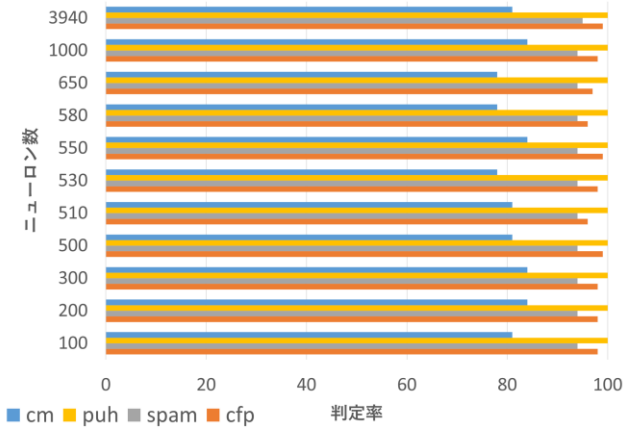

図 3 中間層が 1 層の場合

中間層が一層の場合（図 3）、一番精度が高いニューロン数は 550 であった。次に、中間層が複数層の場合（図 4）、ニューロン数は最初に一番大きい値から、後から減らしていくため、最初の一層は 550 をはじめとし、層を増やしていった。

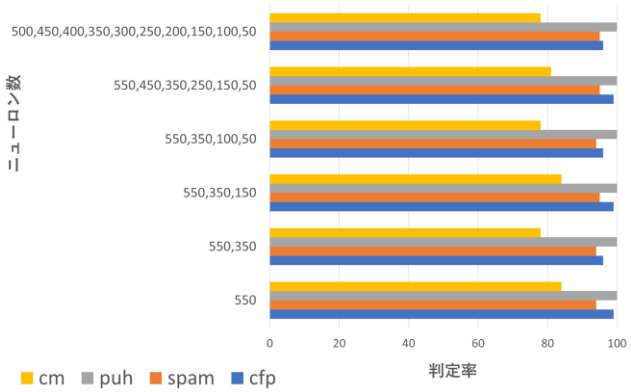

図 4 中間層が複数の場合

その結果、中間層が三層でニューロン数が（550、350、150）の時が、一番判定率が高いことが分かった。そして、この結果からどのように各カテゴリに分類されたか、以下の図 5 にまとめた。

多層パーセプトロン（MLP）を用いて機械学習を行った結果、どのカテゴリに分類されたか考察していきたいと考える。広告（cm）は 16%が迷惑メール（spam）に分類されており、このカテゴリが一番誤ったカテゴリに分類されていることが分かった。本研究では、ユーザが自ら登録して利用したことがある業者やネットサービスサイトから案内などのメールを迷惑メールと区別して広告というカテゴリにしている。従って、広告（cm）メールと迷惑メール（spam）の本文の内容が類似していることから、一割強の広告メールを迷惑メールに判定された。反対に、職場（puh）は 100%で職場（puh）のカテゴリに分類されていた。そのため、専門用語や受信されるメールの内容に特徴があると、分類する精度が高くなると考えられる。

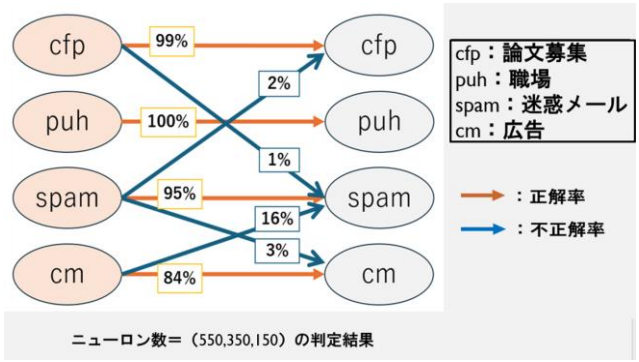

図 5 各カテゴリへの分類結果

また、筆者ら宛てのメールは、研究室内のメールサーバにて最終的に受け付けることになっている。当メールサーバでは、SpamAssassin を利用してサーバサイドで迷惑メールの判定を行っている。SpamAssassin では、受信メールを迷惑メール(spam)か、そうでないメール(ham)の 2 種類へ分類するようになっている。SpamAssassin の標準的な設定と定期的に学習、required_score=5 で運用している。

SpamAssassin の迷惑メール判定メカニズムと多層パーセプトロンによる分類の動作原理において違う面があるが、迷惑メール(spam)の数は、多層パーセプトロンの機械学習で用いた判別用データと同じものである。迷惑メール(spam)を多層パーセプトロンと SpamAssassin との比較を表 3 に示す。

表 3 各カテゴリの判定率

カテゴリ	MLPClassifier の判定率	SpamAssassin の判定率
cfp	99%	-
puh	100%	-
spam	95%	84%
cm	84%	-

表 3 の迷惑メール(spam)の判定を見ると、多層パーセプトロンの判定率（適合率ともいう）は 95%、SpamAssassin の判定率は 84%であった。SpamAssassin と多層パーセプトロン（MLP）を用いた機械学習による迷惑メールの判別の

精度を比較すると、多層パーセプトロン（MLP）を用いた機械学習による判別の方が精度が高いことが分かる。cfp のカテゴリにおいて 1%程度を迷惑メールに分類されている。これは一部の出版業者からの論文募集メールでは、掲載料を目的としているから[10]、そのようなメールは迷惑メールとして分類されたと考えられる。

5. まとめと今後の課題

今回は、電子メールを効率よく処理するため、多層パーセプトロンを利用して複数カテゴリへの分類を試みた。ニューラルネットワークの中間層及びニューロン数の変化により分類の精度に与える影響を調べた。著者ら宛てに届いたメールを 4 カテゴリに分類して正解率（適合率）を解析した。分類の結果（図 5）より、優先度の高いカテゴリ puh においては 100%の正解率を達成した。カテゴリ cm においては 16%を迷惑メールに分類された。これは迷惑メールとコマーシャルメールは多くの共通性があるから考えられる。個人の価値観によっては、この 16%をそのまま迷惑メールであると認定してもよいと思われる。

また、SpamAssassin と多層パーセプトロン（MLP）を用いた機械学習による迷惑メールの判別の精度を比較すると、多層パーセプトロン（MLP）を用いた機械学習による判別の方が精度が高くなっている。これは、多層パーセプトロン（MLP）を用いた機械学習の方が、実際に受信されたメールを元に機械学習を行うため、密に個人の価値観で機械学習を行うことができると考えるが、SpamAssassin でも定期的に学習を行っており、更なる調査が必要であろう。

今回は、メールを複数カテゴリへの分類と、分類の正解率に焦点を絞って行った。結果から本手法の有用性が確認できたと言えろうが、今回の手法では、メール本文のテキスト情報しか利用しなかった。今後の課題としては、①メールのテキストデータだけでなく、添付ファイルや挿入されたイメージなどのバイナリデータの利用；②SMTP セッションの情報と配送過程のメールヘッダー情報の利用；③メールサーバと連携してユーザのメールボックスへの自動仕分けの実装；等が考えられる。

参考文献

[1] 林 治尚：“迷惑メールはなぜ届くのか”、電気学会誌、128 巻 4 号 pp. 215-218（2008）。

[2] 山井 成良、岡山 聖彦ら、“大規模組織における POP before SMTP に基づく管理の容易な電子メールシステム運用方法”、情報処理学会論文誌、Vol.46、No.4、pp.1041-1049（2005）。

[3] RFC 4954, https://datatracker.ietf.org/doc/html/rfc4954

[4] 陳 春祥、佐々木宣介、田中 とし次朗："SMTP セッションフィルタとグレイリス トを併用した迷惑メール対策"、情報処理学会論文誌、第 47 巻、第 4 号、pp.1000-1009、2006 年 4 月。

[5] 松村 拓哉、車古 正樹、井町 智彦、“spam メール対策システムの現状”、学術情報処理研究（JACN）、Vol.10、pp.85-89、2006 年 9 月。

[6] 山口 博之、角 朝香、杉 井学、松野 浩嗣、“ベイジアン方式と機械学習の併用によるスパムメールフィルタリング”、情報処理学会論文誌、Vol.54、No.2、pp.1002-1011、2013 年 2 月。

[7] 櫻庭 秀次、依田 みなみ、清 雄一、田原 康之、大須賀 昭彦、“送信ドメイン認証を用いた送信者レピュテーション構築手法の提案”、 情報処理学会論文誌、Vol.62 No.5、pp.1173-1183、May 2021。

[8] 北倉 奈菜、陳 春祥、“送信ドメイン認証を用いた迷惑メール対策に 関する検討”、2021 年度(第 72 回) 電気・情報関連学会中国支部連合大会、2021 年 10 月。

[9] SpamAssassin, https://spamassassin.apache.org/

[10] Kozak, M., Iefremova, O. and Hartley, J. : “Spamming in scholarly publishing: A case study”, Journal of the Association for Information Science and Technology(2015). doi: 10.1002/asi.23521

[11] 「SMS Spam Collection Dataset」 https://www.kaggle.com/uciml/sms-spam-collection-dataset

[12] Scikit-learn, https://scikit-learn.org/stable/index.html