7ZC-01

日本語 SMS スパムフィルタリング手法の検討

今井京志郎† 金子直史† 鷲見和彦†
青山学院大学 理工学部 情報テクノロジー学科†

1. はじめに

現在，電子メールや Short Message Service（SMS）を用いた広告の配布や，認証方法などがスタンダードになっている．様々な産業の DX が進む中で，こうした仕組みは今後広がっていくだろう．

この中で近年，こうしたメールや SMS を利用したフィッシング詐欺が横行し，増加傾向にある．特に SMS を利用したフィッシング詐欺は“スミッシング”と呼ばれ，こうした詐欺被害 2018 年以降増加傾向にある[1]．

被害を防ぐためにはスパムフィルタリングのシステムが不可欠だが，今のスパムは非常に巧妙で，現行のフィルタリングシステムでは容易に突破されてしまう．特に SMS はメールと比較して，短文であることや，「件名」「送り主」などの情報が少ないことから，フィルタリング精度が落ちることがわかっている[2]．

加えて言語間でも精度に差があり，英語と日本語のスパムフィルタリング精度を比較すると，日本語の方が精度が下がることが分かっている[3]．

これらのことより，日本語 SMS スパムフィルタリングは，新規性・有用性がある．そこで本研究では，日本語の中でも SMS スパムのテキストベースなフィルタリングにおいて，従来手法の精度を超える，精度向上を目標とする．

2. 提案手法

下図 1 は一般的な，テキスト前処理後の文章を，機械学習による識別器で，スパムか否かを判別する枠組みを示している．本研究では，特に前処理部分の改良による性能向上を目指す．

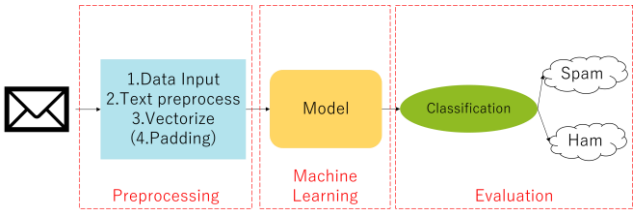

図 1 テキスト分類の流れ

A Study of Japanese SMS Spam Filtering Method
† Kyoshiro Imai, Naoshi Kaneko, and Kazuhiko Sumi
Department of Integrated Information Technology,
Aoyama Gakuin University

2.1. Data Input（入力データ）

入力データには，スパムとハムが一定数存在する．まず訓練用データだが，日本語データセットは入手が難しく，スパムデータに日本語の“メール”スパムニュースを Web スクレイピングによって取得したものを使用し，ハムデータには英語 SMS データセット内のハムを和訳することによって生成した擬似日本語データを用いた．データの構成は下表 1 の通りである．

スパムデータにメールスパムを使用している点において，本研究の対象としている SMS との齟齬が生まれるのではないかという懸念があったが，結果としてテストデータである本物の SMS スパムにおいて，精度をだすことに成功した．

表 1 データセット概要

データ種類	スパム数	ハム数	スパム比率	スパム取得数	ハム取得方法
訓練用	2542	4825	34.51%	メールスパムニュースの Web スクレイピング	英語 SMS データセットの和訳
テスト用	59	46	56.19%	SMS スパムニュースや実際に届いたメッセージ	SMS を利用した広告や認証コードや日常会話

2.2. Text Preprocess（テキスト前処理)

テキストに対する前処理を紹介する．考案した手法は，“指定した品詞のみにする処理（品詞処理)”，“テストデータを訓練データに近づける処理（類似語探索)”，“分類において重要となる単語の重みづけ（重要語処理)”の 3 つである．

品詞処理

テキスト内の品詞を指定したもののみにすることで，英語と違った日本語特有な冗長な情報を削除できると共に，スパムフィルタリングタスクにおいて重要とされる情報の取得を促す．処理前後における TF-IDF 値を比較すると，処理後の方が値が大きく，同じ単語でも重要度が高くなった．

類似語探索

学習済みの分散表現を用いることによって，ある単語と類似している単語を取得することができる．それを用いてテストデータの中身を訓練用データに近づけることによるドメインギャップの削減が可能であり，これが実際有効である結果も出た．

重要語処理

直近のスパムニュースを Web スクレイピングによって取得し，そこに含まれる“名詞”を“重要語”とみなし TF-IDF 値に重みを付けることによって，直近で流行しているスパムをさらに重点的に検出できるといったものである．これはテストデータに最近届いたスパムが含まれていることもあり，精度向上が見られた．

具体的には，元々算出した TF-IDF 値に，独自で算出した重みを加算することによって重要語の扱いをする．重みの算出は，基準とする重みを設定し，そこに前述の類似語探索の類似度を乗算することによって行う．基準とする重みはいくつか試したうち，0.6 が適切であった．

2.3 Evaluation(評価)

評価指標は，Accuracy(正解率)，Recall(再現率)，Precision(適合率)，F1 スコア，ROC-AUC スコアを用いる．

総合スコアとしては ROC-AUC スコアを用いるが，それ以外にも Recall なども考慮した考察を行う．Recall は，実際に“スパム”である中でどのくらい“スパム”として検出できているかという指標であり，本研究ではこちらの方がトレードオフの関係にある Precision より重要であると考えている．

本論文では ROC-AUC スコアと Recall のグラフのみを紹介する．

3. 評価実験

各モデルを実装し，評価実験を行った．各前処理のみの場合と，処理無しの場合，すべての前処理を実装した場合で性能比較を行う．

ROC-AUC スコアをまとめた結果は以下の通りである（下図 2）．

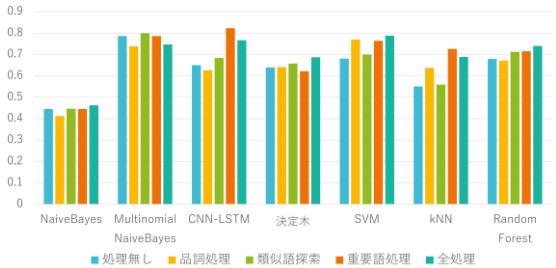

図 2　性能比較（ROC-AUC スコア）

また，下図は Recall でまとめた場合のグラフである（下図 3）．

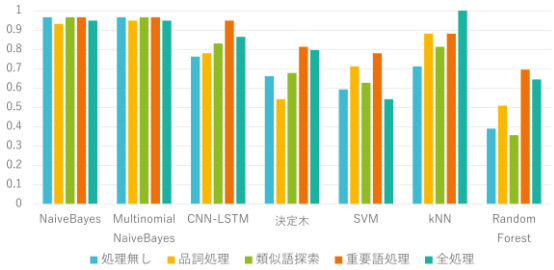

図 3　性能比較（Recall）

4. 考察と結論

前節の結果を考察すると，総合的な性能（ROC-AUC スコア）でいうならば，上図 2 の通りで，“CNN-LSTM”モデルに“重要語処理”を施した手法が最適であることが分かった．これは現行システムとしてのナイーブベイズ系の精度を超えることに成功している点で非常に良い結果と言える．

また Recall（図 3）を見ると，“kNN”が最高値を実現しており，スパムを分類する能力は一番高いことも分かる．

まとめると，次の方策が有効であることを実証した．

1. 学習用データセットとして，一部に和訳データセットを用いても有効であること．
2. 前処理部分に 3 つの工夫を施すことが有効であること．
   (1) 指定した品詞のみにする処理（品詞処理）
   (2) テストデータを訓練データに近づける処理（類似語探索）
   (3) 分類において重要となる単語の重みづけ処理（重要語処理）

これは識別機の種類を問わず，スパムフィルタリングタスクにおいて，汎用的に有効であることが実験により確かめられた．

参考文献

[1] フィッシング対策協議会：“報告書類 月次報告書”．https://www.antiphishing.jp/report/monthly/
[2] 金明哲,村上征勝：“ランダムフォレスト法による文章の書き手の同定”,統計数理,55, 2, pp.255-268 (2007).
[3] 藤森夏輝：“Spam メールの判別に適した機械学習”,高知工科大学 情報学群学士学位論文(2013).
[4] Q. M. A. Abdallah Ghourabi, Mahmood A. Mahmood: “A hybrid cnn-lstm model for smsspam detection in arabic and english messages”, Future Internet 2020,12, 9, p. 156.