

Word2Vec と階層型ニューラルネットワークを活用した パワハラメールの自動判定

Automatic Recognition of Power Harassment Mail using Word2Vec and Multi Layered Neural Network

平山 健一郎¹ 豊谷 純¹ 大前 佑斗¹ 村田 大治²

Kenichiro Hirayama¹, Jun Toyotani¹, Yuto Omae¹, Daiji Murata²

¹日本大学生産工学部 ²テクバン株式会社

¹College of Industrial Technology, Nihon University, ²Techvan Co., Ltd.

Due to the spread of power harassment, the increase in retirees has become a social problem. Therefore, in this research, we developed an automatic judgment system for power harassment mail using Word2Vec and a neural network with the aim of preventing power harassment damage. As a result, it was found that some emails can be judged correctly. The results of this research will contribute to the prevention of power harassment damage on e-mail and SNS, and will contribute to the creation of a comfortable working society.

1. はじめに

近年、パワーハラスメント（以降、パワハラ）が社会問題となっている。厚生労働省が都道府県労働局等に設置した総合労働相談コーナーに寄せられる「いじめ・嫌がらせ」に関する相談は年々増加しており（図1参照）、その中でもパワハラに関する相談が最多となっている。このような状況を鑑み我が国は2019年5月に、改正労働施策総合推進法（通称：パワハラ防止法）を成立させた。大企業では2020年6月、中小企業では2022年4月から施行される予定である[2]。パワハラ防止法により、事業主は職場におけるパワーハラスメント防止のために、雇用管理上必要な措置を講じることが義務となった。これにより、企業はパワハラを早期に発見することが必要となる。

ここで、いくつか考えられるパワハラの手段の一つとして、メールによる精神的な攻撃が考えられる。したがって、メールからパワハラを検出することができれば、パワハラを早期に発見することが可能であると考えられる。メールは自然言語により記述されるため、自然言語処理の各種技術を組み合わせれば、パワハラメールの検出が可能である。

自然言語処理を利用した研究は、我が国で多数の実績がある。例えば、北村ら[3]は日本語を母国語としない人に向けて、日本語の難易度を調整するシステムを開発している。戸次ら[4]は助詞などの機能語を自動検出するシステムを構築している。村越[5]は学校教育現場における教師が作成した授業改善報告書のテキストデータを解析する方法について言及し

ている。

その他にも多数存在する自然言語処理の研究を踏襲し、メールテキストからパワハラか否かを検出する手法の開発が取り組まれている（例えば、FRONTEOが2019年にこのシステムを開発した[6]）。ただし、自然言語処理によるパワハラメールの検出については、未だ歴史が浅く、発展途上の状況と言って良い。また、日本語の解析という特性上、他の研究と比べ、取り組む研究者や技術者が相対的に少ないとも言える。

以上の背景を鑑み本研究では、自然言語処理と機械学習を組み合わせることで、パワハラメールを自動検出する手法を開発する。本稿では、本手法における技術的要素及びテキストデータの処理過程について言及する。

都道府県労働局等への相談件数

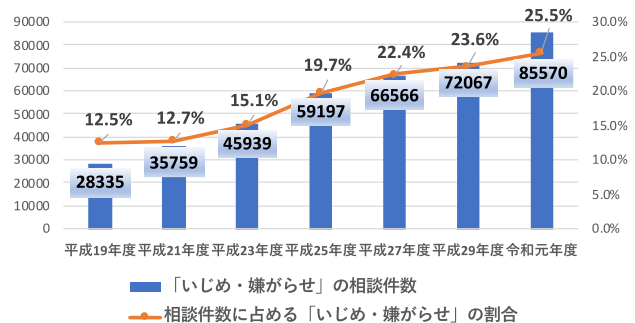


図1. 「いじめ・嫌がらせ」相談件数
(厚生労働省[1]の公開データより著者らが作成)

例：今日はいい天気です。

今日	は	いい	天気	です	。
・名詞 ・副詞可能	・助詞 ・係助詞	・形容詞 ・自立	・名詞 ・一般	・助動詞	・記号 ・句点

図 2. 形態素解析の一例

2. 要素技術と提案手法

本研究では、形態素解析、Word2Vec、ニューラルネットワーク（以降、NN）を組み合わせることでパワハラメールの自動検出する手法を開発している。本章ではこれらの概要について言及する。

2.1 要素技術：形態素解析と Word2Vec

形態素解析とは、与えられた自然言語を言葉の最小単位である形態素に分割し、それぞれの形態素の品詞情報を明らかにする技術である（図 2）。本研究では、Python 3.0 のインポートファイルである形態素解析ライブラリ MeCab と、辞書の一つである ChaSen を使用する。これを利用することで、「東京駅から電車で上野駅まで行きます」のような和文テキストを入力すると「東京/駅/から/電車/で/上野/駅/まで/行き/ます」のように、形態素に分解することができる。また各形態素には、名詞や助詞などの品詞情報も付与されている。日本語は英語などとは異なり、各品詞間がスペースで区切られることがないため、入力されたテキストのままでは、解析しにくい状態と言えるが、形態素解析を実施することで、この問題が解消されるという利点がある。

続いて、Word2Vec について言及する。これは、Mikolov et al.[7]らによって提案された、自然言語処理技術の一つである。構築された Word2Vec を利用することで、入力された単語と指定した単語との類似度を測定することができる。事前に与えられた膨大なテキスト内で、各形態素同士が同時に発生した程度に関する情報を参考とし、類似度測定を実現している。この技術を活用することで、数字として表現することができなかった形態素を、実数値のベクトルとして表現することが可能になる。

2.2 要素技術：ニューラルネットワーク

最後の要素技術として、階層型 NN を紹介する。これは、回帰あるいは分類問題を解くことのできる教師あり学習モデルの 1 つである。すなわち、入力と正答のペアで構成される教師データを用意することで、特定の入力に対して妥当な出力を推定する機能を構成することができる。入力層、中間層、出力層と呼ばれる三つのレイヤーを有している。入力層に入力されたベクトルが、ウェイトパラメータ、バ

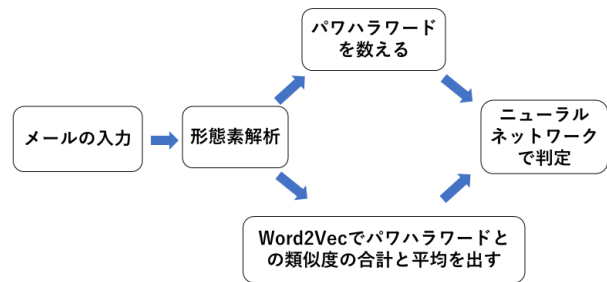


図 3. 提案手法の概要

イアスパラメータにより変換され、それが活性化関数に代入されることで、別の次元となった新たなベクトルが生成される。このベクトルを中間層と呼び、同様の演算により出力層に伝播する。出力層として算出された値が推定結果を意味する。

なお、NN の入力の実数値によるベクトルしか採用することができない。これは、自然言語であるテキストをそのまま入力することができないことを意味する。このため、テキストから何らかの状態を自動判定するために NN を利用する際は、Word2Vec などの技術を利用して、テキストを実数値に変換することが要求される。このように、入力されたテキストデータが、何らかの意味を持つ数値情報として変換されたものを、特徴量と呼ぶ。すなわち、特徴量が NN の入力層に与えられるベクトルを意味する。

2.3 提案手法：自動判定モデル

続いて、これまで紹介した要素技術を用いて考案したパワハラメールの自動判定モデルについて言及する（図 3）。提案手法においては、多数のメールテキストを利用して、事前に Word2Vec を学習しておくことが必要となる。これにより、パワハラメールで使用されやすい形態素同士の類似を高いと出力するような Word2Vec を構築する（例えば、「このバカ、給料下げるぞ。休日も出勤してやれ」という文があった場合、「バカ」、「給料」、「休日」間の類似度が高くなる）。

続いて、与えられたメールテキストを形態素解析により、形態素に分解する。その後、事前に登録しておいたパワハラワード（PW）の登場頻度をカウントする。PW とは、「バカ」「アホ」「死」などのように、パワハラメールで使用されやすい形態素のことである。すなわち、直接的にパワハラに関連する用語が使用されている場合、それを個数として記録する意味を持つ。

上記処理により、パワハラで直接使用される用語が登場したときに、パワハラだと判定されやすくなる NN を構築することが可能だと考えられるが、PW

表 1. 特徴量ベクトル (PW が 2 つの場合)

次元数	特徴量
1 次元目	PW1 の使用数
2 次元目	PW2 の使用数
3 次元目	PW1 との類似度の平均値
4 次元目	PW2 との類似度の平均値
5 次元目	PW1 との類似度の合計値
6 次元目	PW2 との類似度の合計値

PWn: n 番目のパワハラワード

が直接使用されないパワハラメールの存在も、容易に想像することができる。このため、すべての形態素について、事前に構築した Word2Vec に代入し、PW との類似度を算出する。すなわち、ここで算出される類似度は、パワハラメールで登場しやすい形態素であれば高くなり、普通のメールのみで登場しやすい形態素であれば低くなることが予想される。すべての形態素について得られた類似度について、その平均値および合計値を算出する。この情報を利用することで、PW が直接利用されていなくても、それと類似度の高い用語が使用されていれば、パワハラだと判定できる NN が構築できると考えられる。類似度の平均値と合計値の 2 つを採用している理由は、様々なメールの長さに対応するためである。

以上、一連の議論を踏まえた上で構成した特徴量ベクトルは表 1 の通りである。1, 2 次元目は PW の使用数, 3, 4 次元目は PW との平均値, 5, 6 次元目は合計値である。なおこれは、PW の数が 2 個（例えば、「バカ」「アホ」など）の場合に限る次元数である。PW は 3 個以上登録しても良い。1 つの PW あたり 3 つの特徴量が生成されるので、採用する PW の数の 3 倍が、NN の入力層に代入される特徴量ベクトルの次元数となる。

この特徴量ベクトルとパワハラか否かの正答値がペアとなったデータが教師データとなる。これを利用して階層型 NN を学習させることで、パワハラメールの自動判定モデルが構築される。

3. 実験

3.1 目的と概要

提案手法の信頼性を検証することを目的に、実験を実施した。このため、はじめに教師データを構築した。Word2Vec の教師データとして、厚生労働省のホームページに掲載されていた過去のパワハラの判例から抜粋した PW が含まれた文章、および、それに類似する文章を複数作成し、それらをランダムに結合させ、500 文のメールテキストを生成した。NN の教師データとして、上記 500 文の中から 20 文をラ

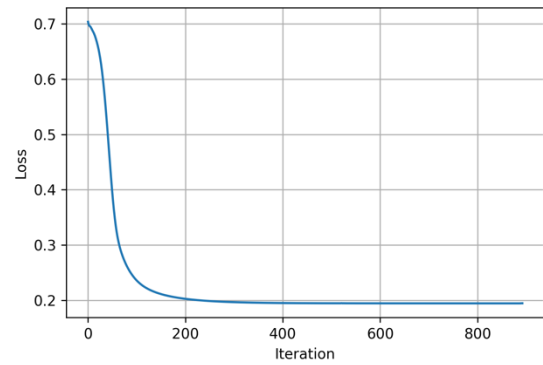


図 4. NN の学習過程

ンダムに選択したものをパワハラメール、著者らが作成した 20 文を普通のメール、合計 40 文のメールテキストを生成した。パワハラメールには 1、普通のメールには 0 とラベリングしている。

2.3 節で述べた PW については、「バカ」と「アホ」の 2 つを採用した。すなわち、特徴量ベクトルの次元数はその 3 倍である 6 次元となる。NN としては、入力層、中間層 1、中間層 2、出力層の 4 層構造とした。入力層は特徴量となる 6 次元、中間層 1 と 2 は 10 次元、出力層は 1 次元となる。出力層は最小 0 最大 1 を取る実数値であり、0.5 以上ならパワハラ、0.5 未満なら普通のメールと判定される。最適化手法は Adam、活性化関数はシグモイド関数、学習回数は 1000 回、学習係数は 0.01 をそれぞれ採用した。PW の類似度測定に利用する Word2Vec は、Skip-gram 法を採用した。登場回数が 1 回未満の単語は除去し、前後 5 単語までを対象とするよう設定した。学習回数は 3000 回、学習係数は 0.1 を指定した。

3.2 結果と考察

3.1 節の条件で構築した NN の学習過程を図 4 に示す。横軸は学習回数 (Iteration)、縦軸の Loss は NN のコスト関数であり、低くなるほど学習が進んでいることを意味する。図 4 を確認すると、学習回数が 100 回までで Loss が急激に減少し、200 回以降はほぼ収束していることが確認できる。学習回数が 1000 回に至る前、900 回付近で Loss が改善されなくなり、学習が打ち切られている。Loss が低いことから、教師データとして設定したメールについて、パワハラかどうか正しく判定できるようになったことが確認できる。

構築したモデルを利用して、いくつかのメールテキストにおいてパワハラか否かを自動判定した結果を表 2 に示す。ID 列の A は、「PW が直接使用されている短いパワハラメール」、B は「PW が直接使用されている長いパワハラメール」、C は「PW が使用

表 2. 提案手法による自動判定結果

ID	メール本文	字数	実測	推定	出力
A1	バカ。使えない。おまえ。おまえの存在が目障りだ。何時間かかてるんだ、俺だったら 10 分で出来るわ。	49	パワハラ	パワハラ	87.2 %
A2	いつでもおまえを飛ばすことができる。有休取れるわけないだろ。アホ。どれだけ迷惑かければ気が済むんだ。	50	パワハラ	パワハラ	99.5 %
A3	面接のときの発言はうそだったのか？だまされたわ。バカ。アホ。クソ。金返せ。もう契約更新しないから。	49	パワハラ	パワハラ	98.8 %
A4	意欲がない。アホ、そんなこともできないのかバカ。甘えるな。何時間かかてるんだ、俺だったら 10 分で出来るわ。	54	パワハラ	パワハラ	99.0 %
A5	おばさん。消えろ。反省してるのか。Fラン大学のバカ。すごい優秀な新人が来るぞ、おまえやばいな。	47	パワハラ	パワハラ	99.9 %
B1	契約とれるまで帰ってくるなバカ。殺す。すごい優秀な新人が来るぞ、おまえやばいな。いつでもおまえを飛ばすことができる。アホ。高卒。上司に逆らうのか？。おっさん、ちゃんと考えたのか？ 社会人として常識だ。無礼なやつだな。	108	パワハラ	パワハラ	99.9 %
B2	ふざけるな。バカ。ぶさいく。おまえより成長が遅いやつは見たことがない。給与を返せ。おまえの子供がかわいそう。くさいわアホ。みんなやってるぞ、おまえ以外できてるぞ。アホ。死ぬ気でやれ。	92	パワハラ	パワハラ	99.9 %
B3	ゆとり。上司に逆らうのか？。おまえのせいで、みんなイライラしてるぞ。死ぬ気でやれアホ。休みの日、勉強したことを報告しろ。これ、常識だから。ゆとり。上司に逆らうのか？そんなこともできないのかバカ。	97	パワハラ	パワハラ	99.9 %
B4	消えろバカ。有休取れるわけないだろ、アホか？休めと思うな。おまえ何回目だ、何回教えたらわかるんだ。やる気あるのか？のろま。金返せ。そんなこともできないのか。甘えるなバカ。消え失せろアホ。	94	パワハラ	パワハラ	100.0 %
B5	まじめにやってる？どれだけ迷惑かければ気が済むんだ。休みの日、勉強したことを報告しろ。アホ、そんなこともできないのかバカ。おまえ以外みんな出来てるぞ。チビ。中卒バカ。こんな成績で正社員なれないぞ。	98	パワハラ	パワハラ	99.9 %
C1	そんなこともできないのか。甘えるな。何時間かかてるんだ。すごい優秀な新人が来るぞ、おまえやばいな。	50	パワハラ	パワハラ	98.7 %
C2	反省してるのか。どれだけ迷惑かければ気が済むんだ。面接のときの発言はうそだったのか？消えろ。	46	パワハラ	パワハラ	69.1 %
C3	ふざけるな。有休取れるわけないだろ、おまえやばいな。ゆとり。これ、常識だから。	39	パワハラ	パワハラ	98.9 %
C4	いつでもおまえを飛ばすことができる。消えろ。休みの日、勉強したことを報告しろ。	39	パワハラ	パワハラ	87.2 %
C5	意欲がない。おまえより成長が遅いやつは見たことがない。おまえ何回目だ、何回教えたらわかるんだ。	47	パワハラ	普通	39.4 %
D1	金返せ。もう契約更新しないから。まじめにやってる？どれだけ迷惑かければ気が済むんだ。おまえの子供がかわいそう。やる気あるのか？おまえ以外みんな出来てるぞ。中卒、チビ。ふざけるな。	90	パワハラ	パワハラ	99.8 %
D2	おまえ何回目だ、何回教えたらわかるんだ？そんなこともできないのか。甘えるな。休みの日、勉強したことを報告しろ。いつでもおまえを飛ばすことができる。これ、常識だから。甘えるな。くさいわ	91	パワハラ	パワハラ	99.4 %
D3	何時間かかてるんだ、俺だったら 10 分で出来るわ。おまえより成長が遅いやつは見たことがない。おまえ何回目だ、何回教えたらわかるんだ。まじめにやってる？くさいわ、本当くさいわ。	87	パワハラ	普通	34.7 %
D4	クソ。金返せ。もう契約更新しないから。有休取れるわけないだろ、ゆとり。おまえ何回目だ、何回教えたらわかるんだ。上司に逆らうのか？こんな成績で正社員なれないぞ。	79	パワハラ	パワハラ	99.6 %
D5	使えないおまえ。面接のときの発言はうそだったのか？だまされたわ。おまえの存在が目障りだ。すごい優秀な新人が来るぞ、おまえやばいな。甘えるな、消え失せろ中卒。	79	パワハラ	パワハラ	90.5 %
E1	この度は、お忙しい中、貴重な時間を割いていただき誠にありがとうございます。何卒よろしくお願い致します。	51	普通	普通	1.7 %
E2	この度詳しくお話を伺いすることができ、大変参考になりました。本当にありがとうございました。	46	普通	普通	8.4 %
E3	ご理解感謝いたします。とても助かります。それでは表題の件、ご確認のほどよろしくお願いいたします。	48	普通	普通	9.4 %
E4	その後の進捗は、いかがでしょうか？期限の方が今週金曜日と迫っているため、本日 15 時までにご送付をお願いできますでしょうか。	61	普通	普通	12.9 %
E5	また、本メールと行き違いでご連絡をいただいておりますら申し訳ありません。	37	普通	普通	2.5 %
F1	さて、この度お客さまからご返却いただきましたクレーム品の調査結果を下記の通りご報告させていただきます。当サンプルが必要となる会議が今週金曜日と迫っているため、本日 15 時までにご送付をお願いできますでしょうか。	104	普通	普通	10.6 %
F2	この度は、こちらの不手際で混乱を招いてしまいましたこと、心よりお詫び申し上げます。なお、正しい内容のメールを「打ち合わせについて」という件名で再送させていただきます。それでは、よろしくお願い致します。	101	普通	普通	1.4 %
F3	表題の件ですが、引き続き発生経緯の調査をすすめ、今後発生頻度を踏まえて対策を検討していきます。この度は、こちらの不手際で混乱を招いてしまいましたこと、心よりお詫び申し上げます。	88	普通	普通	1.0 %
F4	先日はお時間の無い中ご対応いただきありがとうございます。早速ですが、ご提案いただいた件につきまして、提案資料を作成していただきたくお願いいたします。それでは、よろしくお願い致します。	92	普通	普通	1.7 %
F5	お世話になっております。その後の進捗は、いかがでしょうか。ご不明な点等ございましたらご連絡ください。ご参考までに過去の資料を添付させて頂きました。すみませんが、ご確認のほどよろしくお願いいたします。	99	普通	普通	20.2 %

されていない短いパワハラメール」、D は「PW が使用されていない長いパワハラメール」、E は「パワハラではない短いメール」、F は「パワハラではない長いメール」をそれぞれ意味する。すなわち、ID 列 A, B, C, D をパワハラと判定すれば正答であり、E, F を普通と判定すれば正答となる。また、A, B は PW が直接利用されているため NN にとって難易度の低い問題、C, D は PW が使用されていないパワハラメールであるため難易度の高い問題になると考えられる。

出力列は NN の実数値による出力値であり、0 を 0%, 1 を 100% となるように表記している。

表 2 における A, B を見るといづれも出力値が高いことがわかる。すなわち、PW である「バカ」と「アホ」のいずれかが直接使用されている場合、メールテキストの長さに依存せず、パワハラだと検出できることがわかる。これは、表 1 に示すように、PW の登場回数の特徴量としているためだと考えられる。

表2におけるC, DはPWが直接利用されていない場合である。この条件であっても、多くのパワーハラスメントを正しくパワーハラと判定することに成功している。これは、Word2Vecにより得られるPWとの類似度を特徴量としたことが理由であると考えられる。一方、C5とD3について、誤判定が認められた。該当のテキストには「成長」や「出来る」といった前向きなワードも含まれていることが原因かもしれない。この改善策として、Word2VecとNNの双方の教師データに、前向きなワードを含むパワーハラスメントを追加する必要があると考えられる。

表2におけるE, Fは普通のメールである。これらはいずれも適切に判定されていた。これは、テキスト内にパワーハラワードが直接使われていないことに加え、PWとの類似度の平均値と合計値が低いことが原因だと考えられる。

4. おわりに

本研究では、パワーハラスメントという重要な社会問題の1つを解決するために、メールにおけるパワーハラを自動判定するシステムを、形態素解析、Word2Vec, NNを用いて開発した。表2に示すように多様なメールにおいて適切な判定をできることが確認された。ただし、教師データや精度検証用のデータは少数であるため、今後はより多くのメールテキストを用意することが必要である。また、PWとして「バカ」と「アホ」の2つのみを採用したが、提案手法はその他にも様々なワードを採用できるので、今後は、PWの種別を増やすとともに、それが検出精度に与える影響を検証していく予定である。

参考文献

- [1] 厚生労働省：データで見るハラスメント, <https://www.no-harassment.mhlw.go.jp/foundation/statistics/>, 2020.10.30 閲覧。
- [2] 厚生労働省：職場におけるハラスメントの防止のために, https://www.mhlw.go.jp/stf/seisakunitsuite/bunya/koyou_roudou/koyoukintou/seisaku06/index.html, 2020.10.30 閲覧。
- [3] 北村達也, 住田真一, 孝橋一希, 文難易度の調整を支援するシステム「かぶとエディタ」, 日本語教育方法研究会誌, vol.22, no.1, pp.14-15, 2015.
- [4] 戸次徳久, 仁科喜久子, 機能語の抽出ツールの開発, 日本語教育方法研究会誌, vol.9, no.2, pp.10-11, 2002.
- [5] 村越亮治, 統計解析ソフト R によるテキストマイニングの学校現場での応用可能性-授業改善のためのアクション・リサーチ報告書の分析を例に, 神奈川県立国際言語文化アカデミア紀要, vol.2, pp.165-175, 2013.
- [6] PR TIMES, FRONTEO, AI を活用した電子メール監査システム「Email Auditor 19」にハラスメント情報を学習させたナレッジベースを追加, <https://prtimes.jp/main/html/rd/p/000000267.000006776.html>, 2019, 2020.

11.12 閲覧。

- [7] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, J. Dean: Distributed Representations of Words and Phrases and Their Compositionality, In NIPS, 2014.