Twitter におけるアイコン画像と攻撃ツイートの関連性

田中 智大 清 雄一† 田原 康之‡ 大須賀 昭彦‡

†電気通信大学情報理工学域 〒182-8585 東京都調布市調布ケ丘１ー５－１ ‡

E-mail: †t1910758@gl.cc.uec.ac.jp, seiuny@uec.ac.jp, tahara@uec.ac.jp, ohsuga@uec.ac.jp

あらまし ＳＮＳは多くの人が利用するので, 誹謗中傷を行う攻撃的な人も存在する. そのような状況で, 特定のアイコンユーザは攻撃的であるといった意見がネット上で散見される. しかし, 実際にそのような内容の文献を目にしたことはない. そこで, 本研究ではアイコン別に収集したユーザのツイートの攻撃性を日本語用 Sentence BERT を用いて算出し, アイコンとツイートの攻撃性の関連性の調査を行う. また, アイコンから攻撃性を算出するシステムを考案する. その結果, 人, イラスト, 電車アイコンに有意差があり, システムの正答率は 8%であった.

キーワード アイコン，攻撃性，Twitter，Sentence BERT

1. はじめに

近年, ＳＮＳは我々の生活になじみ深いツールになった. 特にコロナ禍以降は，ＳＮＳで会話をする機会が増えている.

しかし ,ＳＮＳには負の側面も存在する .その１つが匿名による誹謗中傷である. そのような状況で, 特定のアイコンが攻撃的であるといった意見がネット上で散見される. しかし, 実際にそのような内容の研究は著者の知る限り無い.

そこで, 本研究ではアイコンと攻撃性の関連性の調査を行う. また, ツイートからの攻撃性の算出には多くの計算時間が必要となる. そこで, アイコンを入力とした攻撃性判定システムを考案, 作成し, Twitter の利用をより快適にできないかと考えた.

2 関連研究

2.1 Sentence-BERT[4]

Sentence-BERT は, シャム/Triplet ネットワークを用いて, 意味のある分の埋め込みを導出する BERT ネットワークの改良版である .これにより ,これまでの BERT が適用できなかった新しいタスクに BERT を適用することができる.

2.2 CLIP[6]

CLIP は OpenAI の事前学習画像分類モデルであり,
主な特徴は次の 3 つである.

①カテゴリを利用者側で自由に設定できる自然言語教師型画像分類モデル

②巨大な自然言語教師データ「WebImage Text」の利用

③多様なタスクに対するゼロショット転移で転用可能

多様なタスクに対し, ゼロショット転移で優れた精度を出した.

2.3 アイコン画像に注目した **Twitter** 研究の提案**[1]** 冨永らはアイコンとのツイート数, フォロー・フォロワー数の関連を調べる研究を行った. アイコンのカテゴリは 13 種類であり, 動物, たまご, 自画像, 顔隠し, 文字, ロゴ, オブジェ, オタク, 本人一人, 本人複数, 景色 , 他人, キャラクタであり , このカテゴリについて, 一致度と網羅性の評価が行われている.

表 1 :[2]冨永 登夢, 土方 嘉徳, Twitter 上のアイコン画像とユーザ行動の関係の調査と分析(2016)273 表 1 より引用表 1 : アイコン画像のカテゴリとその説明[2]

カテゴリ	定義
本人一人 (On)	ユーザ本人の顔写真
自画像 (Sp)	ユーザの顔がイラストで描かれたもの
顔隠し (Hf)	顔がはっきりと確認できない写真
本人複数 (As)	友人や家族と一緒に写っているもの
他人 (Dp)	芸能人やスポーツ選手などの有名人の写真
文字 (Le)	文字のみで構成される画像
ロゴ (Lo)	会社や学校，その他組織のロゴ
オタク (Ot)	アニメ／漫画調で書かれた美少女キャラクタ
キャラクタ (Ch)	くまモンやふなっしー，ドラえもんなどのキャラクタの画像
動物 (An)	犬や猫，鳥などの動物が写った写真や画像
オブジェ (Ob)	好きな車や趣味で使う道具などの物体
景色 (Sc)	風景の写真や画像
たまご (Eg)	ユーザが Twitter に新規登録した時にデフォルトで設定されている画像

3 提案手法

アイコンごとの攻撃性算出は次の手法を提案する.

(1)定義した 13 種類のアイコンについて, それぞれ 100 人のユーザを収集

(2)各ユーザから最大 1000 ツイートを収集

(3)誹謗中傷をクエリとし日本語用 Sentence-BERT[5]にてクエリとのコサイン距離を算出

(4)アイコンごとにクエリとのコサイン距離の近い上位 200 ツイートの 1 ツイート当たりのクエリとのコサイン距離の平均を求める

この手法のフローチャートを次に示す.

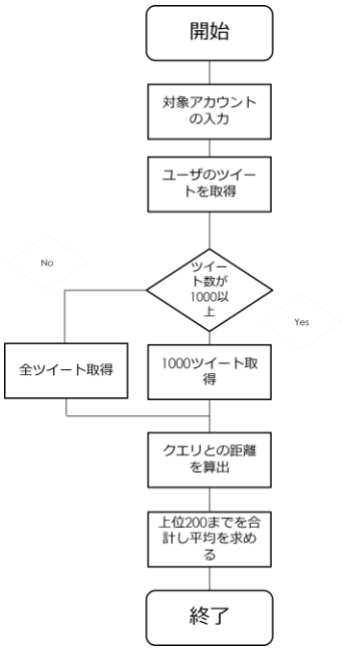

図 1:攻撃性算出のフローチャート

攻撃性判定システムは次の手法を提案する.

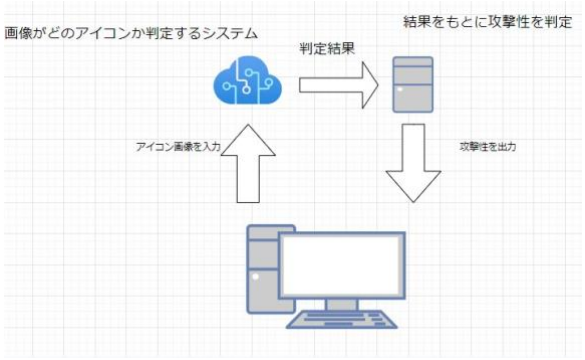

図 2:攻撃性判定システムのシステム構成図

アイコンを入力, 攻撃性を出力とするシステムを考案する.

このシステムは入力された画像がどのカテゴリにあてはまるかを OpenAI の CLIP[6]を用いて判定する. その後, 各カテゴリとの一致度とアイコンごとの攻撃性算出の結果を用いて攻撃性を判定する.

4 実験（ツイート分析）

アイコンカテゴリは参考文献[1][2]で用いられたカテゴリを一部改変した, 人 , オブジェ, 動物, 初期アイコン, 飲食物, 女キャラ, 男キャラ, その他キャラ, 文字, 景色, 電車, その他乗り物, イラストの 13 カテゴリである.

Twitter のアイコンのみから本人であるか判断するのは困難である. よって人物に関するカテゴリを一つにまとめた.

ロゴアイコンは, 冨永らの研究[1][2]から企業のアカウントが多いことが明らかとなっている. よって, ロゴアイコンを削除した.

自画像も区別が困難なためイラストとした.

キャラクタとオタクに関しては, 細分化し男キャラ, 女キャラ, その他キャラの 3 カテゴリに分けた.

新たに電車アイコンとその他の乗り物アイコンを追加した. これは, 電車アイコンに関するうわさが散見されたためである.

食べ物アイコンを追加したのは , アイコン収集の際に多くの食べ物アイコンのユーザを確認したためである.

ツイートの収集は Twitter API を用い, 収集したツイート数は 1 ユーザにつき最大 1000 ツイートである.

ツイートの分析は日本語用 Sentence-BERT[5]を用いて行う. クエリは”誹謗中傷”であり, 1000 ツイートのクエリとのコサイン距離が小さい上位 200 ツイートを合計し, 平均を求める.

また, 収集するツイートから URL やユーザ id, 不要記号 , “ 誹謗中傷 ” を削除した . “ 誹謗中傷 ” は , Sentence-BERT によりクエリとのコサイン距離を求める際に , “誹謗中傷はよくない”等のツイートとコサイン距離が近くなってしまうのを防ぐために削除した.

下に各アイコンの上位 100 から 600 ツイートまでの各コサイン距離の合計を示す.

表 2 : 各カテゴリのクエリとのコサイン距離が近いツイートのコサイン距離の合計

	人	オブジェ	動物	初期アイコン	飲食	女キャラ	男キャラ	その他キャラ	文字	景色	電車	その他乗り物	イラスト
100	3227	3246	3324	3283	3320	3318	3259	3265	3148	3289	3484	3381	3306
200	6194	6408	6419	666	6456	6434	6098	6334	6349	6227	6738	6690	6533
300	8426	9149	8978	9997	8986	8977	8317	8840	9311	8882	9595	9109	9105
400	10746	11890	11465	13424	11547	11533	10551	11197	12270	11468	12364	11618	11743
500	13101	14455	13814	16872	13983	13902	12671	13377	15250	13961	14694	14100	14323
600	15227	16751	15973	20303	16155	16091	14554	15262	18176	16220	16518	16310	16689

縦軸がツイートの合計数, 横軸がアイコンの分類カテゴリである.

この表を見ると, どのカテゴリも 200 ツイートまでのコサイン距離の合計が 100 ツイートまでの距の合計の倍近くになっている. しかし, 300 ツイートまでの合計は, 過半数のカテゴリが 3 倍よりも小さい値になっている. これは 200 以上 300 未満のツイート数であるユーザが多いためだと考えられる.

よって, ユーザの偏りをなくすために 200 ツイートとした. また, 100 ツイートでないのは, 1 人の攻撃的なユーザから受ける影響を少なくするためである.

5 結果（ツイート分析）

表 3:13 カテゴリの 1 ツイート当たりのクエリとのコサイン距離の平均

人	オブジェ	動物	初期アイコン	飲食	女キャラ	男キャラ	その他キャラ	文字	景色	電車	その他乗り物	イラスト	ベース
0.357	0.366	0.366	0.362	0.367	0.360	0.362	0.365	0.363	0.367	0.385	0.373	0.357	0.365

ベースはアイコンを限定せず無作為に収集したユーザである. 結果, 人アイコン , イラストアイコンがベースよりクエリとのコサイン距離が小さい結果となり, 電車アイコン, その他乗り物アイコンがベースよりクエリとのコサイン距離が大きい結果となった.

ここで, この差が有意であるかを確かめるため片側 t 検定を行った.

表 4:人, イラストとベースの片側 t 検定の結果

	人	ベース		イラスト	ベース
平均	0.357019	0.36577	平均	0.357431	0.36577
分散	0.00107	0.001486	分散	0.00099	0.001486
観測数	100	100	観測数	100	100
プールされた分散	0.001278		プールされた分散	0.001238	
仮説平均との差異	0		仮説平均との差異	0	
自由度	198		自由度	198	
t	-1.73115		t	-1.67591	
P(T<=t) 片側	0.042492		P(T<=t) 片側	0.047667	
t 境界値 片側	1.652586		t 境界値 片側	1.652586	

表 4:電車, 乗り物とベースの片側 t 検定の結果

	乗り物	ベース		電車	ベース
平均	0.373058	0.36577	平均	0.385215	0.36577
分散	0.000986	0.001486	分散	0.001072	0.001486
観測数	100	100	観測数	100	100
プールされた分散	0.001236		プールされた分散	0.001279	
仮説平均との差異	0		仮説平均との差異	0	
自由度	198		自由度	198	
t	1.465792		t	3.844706	
P(T<=t) 片側	0.072146		P(T<=t) 片側	8.13E-05	
t 境界値 片側	1.652586		t 境界値 片側	1.652586	

片側 t 検定の結果, 人, イラスト, 電車アイコンに有意差が見られた. よって, 人, イラストアイコンの攻撃性が高く, 電車アイコンの攻撃性が低い結果となった.

6 実験（攻撃性判定システム）

システム構成は, 入力したアイコン画像を CLIP[6]で分類し, それをもとに攻撃性を判定するというものである. システムで用いるカテゴリは攻撃性算出時の 13 カテゴリから初期アイコンを除いた 12 カテゴリを用いる. これは初期アイコンが一様であり, ユーザごとにアイコンの差が生じないからである.

α はカテゴリと一致する割合, β はそのカテゴリの 1 ツイート当たりのクエリとのコサイン距離の平均, $\alpha \times \beta$ の総和を sum とする. この操作を α の合計が 0.99 以上になるまで行う. その後, 残りの 1%を補うため sum に式 (1)を加える.

$$(1-\sum\alpha)\times\beta max \qquad (1)$$

$\sum\alpha$はカテゴリとの一致度の合計, β max は最も一致するカテゴリの 1 ツイート当たりのクエリとのコサイン距離の平均である.

7 結果（攻撃性判定システム）

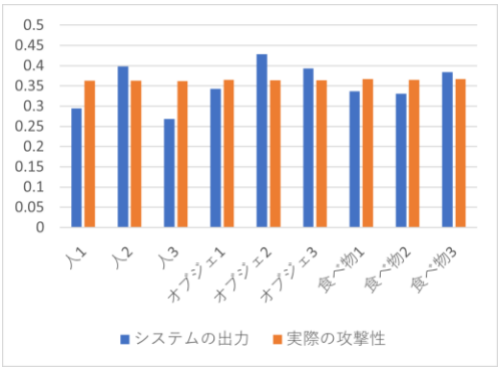

図 3: システムの出力と実際の攻撃性の比較

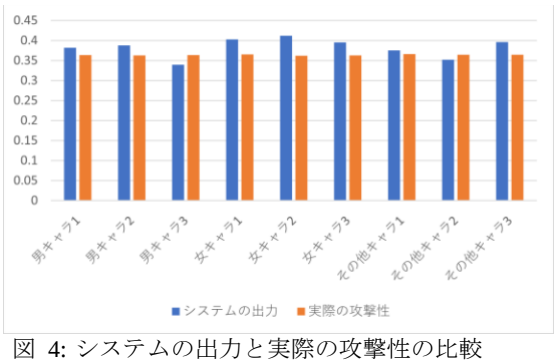

図 4: システムの出力と実際の攻撃性の比較

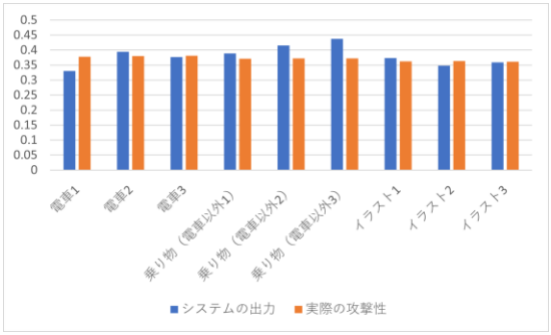

図 5: システムの出力と実際の攻撃性の比較

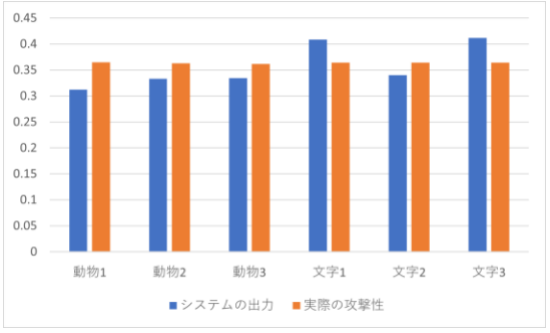

図 6: システムの出力と実際の攻撃性の比較

システムの出力と実際の結果を図 3 から図 6 で比較する. 結果, システムの出力と実際の攻撃性の差が 0.01 未満であったものは全体の約 8%であった.

8 考察

ツイート分析の結果, 片側 t 検定により有意差が見られたのは人アイコン, イラストアイコン, 電車アイコンであった.

人アイコンユーザのツイートを確認すると, アイコンをセクシーな女性にし, 特定のサイトへ誘導しようとしているアカウントを複数確認した. よって, 人アイコンはそのようなアカウントが誘導する過程で, 一般ユーザのツイートより過激な内容を投稿しているので攻撃性が高いと考えられる.

また. イラストアイコンのユーザは絵を描くユーザを複数確認した. それらのユーザが自分の描いた絵を自虐, または無断転載などへの怒りのツイートを行っているのを確認した. よって, これらの要因でイラストアイコンの攻撃性が高いのだと考えられる.

一方, 電車アイコンは趣味である電車に関するツイートをしているユーザを多く確認した. 電車に関するツイートの中では攻撃的な言葉を使う頻度が少ないため攻撃性が低いと考えられる.

また, 攻撃性判定システムの出力と実際の攻撃性の差が 0.01 未満であるものは全体の約 8%であり, これはアイコンのみを入力としているからだと考えられる. アイコンのみが入力であると, 同アイコンのユーザは同じ出力になるからである. よって, アイコンと少数のツイートを入力とするシステムに改良するとより良い結果が得られると考えられる.

9 まとめ

本研究では, Twitter ユーザをアイコンにより 13 カテゴリに分類し, 日本語用 Sentence-BERT を用い, ツイートと“誹謗中傷”のコサイン距離を算出した. その結果, 有意差があるのは 3 カテゴリであり, 人アイコンユーザ, イラストアイコンユーザの攻撃性が高く, 電車アイコンユーザの攻撃性が低い結果となった.

また, その結果を用いて入力を画像 , 出力を攻撃性とするシステムを考案, 作成した. その結果, 出力と実際の攻撃性の一致率は約 8%であった.

10 展望

本研究で提案した攻撃性判定システムは, 入力がアイコンのみという性質上, 一致率は約 8%と低い値であった. よって, アイコン画像とツイートの 2 つを入力とするシステムに改良するとより精度が向上するのではないかと考える. しかし, ツイートの分析には多くの時間が必要であり, 本システムの目的はその時間を短縮することであり, 入力するツイート数を考慮する必要がある. よって, アイコン画像と共に入力するツイート数の検討を行いたいと考える. また, アイコンを収集する際 , 種類によってはあまりユーザが見つからないカテゴリもあり, データ収取に多くの時間を必要とした. よって, ユーザ数の少ないアイコンを短時間で収集する手法も検討したいと考える.

謝辞

本研究は JSPS 科研費 JP21H03496, JP22K12157 の助成を受けたものです. 本研究は, 電気通信大学人工知能先端研究センター（AIX）の 計算機を利用して実施したものです.

参 考 文 献

[1] 冨永登夢, 土方嘉徳, 西田正吾“アイコン画像に注目した Twitter 研究の提案”, The 28th Annual Conference of the Japanese Society for Artificial Intelligence, 2014

[2] 冨永登夢, 土方嘉徳, 西田正吾“Twitter 上のアイコン画像とユーザ行動の関係の調査と分析” 情報処理学会インタラクション 2016

[3] 瀬川 友香, 浅谷 公威, 坂田 一郎, ユーザーに着目した SNS 上の攻撃とそのメカニズムに関する分析(2021)

[4] Reimers, N. and Gurevych, I.: Sentence-bert: Sentence embeddings using siamese bert-networks, arXiv preprint arXiv:1908.10084 (2019)

[5] 日本語用 Sentence-BERT モデル（バージョン 2）“sonoisa/sentence-bert-base-ja-mean-tokens-v2

[6] Alec Radford Jong, Wook Kim , Chris Hallacy ,Aditya Ramesh , Gabriel Goh , Sandhini Agarwal , Girish Sastry , Amanda Askell , Pamela Mishkin , Jack Clark , Gretchen Krueger , Ilya Sutskever, Learning Transferable Visual Models From Natural Language Supervision, arXiv:2103.00020, Fri, 26 Feb (2021)

.