デマの検知に向けたモダリティの活用の検討

2025/02/21

国立研究開発法人 情報通信研究機構
サイバーセキュリティ研究所　サイバーセキュリティ研究室
藤田 彬

紹介する研究の論文著者

1. 松田 美慧
   - ✓横浜国立大学 大学院環境情報学府
   - ✓NICT サイバーセキュリティ研究所 サイバーセキュリティ研究室 研究員

2. 藤田 彬
   - ✓NICT サイバーセキュリティ研究所 サイバーセキュリティ研究室 主任研究員
   - ✓横浜国立大学 大学院先端科学高等研究院 IAS客員研究員

3. 吉岡 克成
   - ✓横浜国立大学 大学院環境情報研究院 教授
   - ✓横浜国立大学 先端科学高等研究院 情報・物理セキュリティ研究ユニット 主任研究者

背景

- ソーシャルメディアは現代社会に欠かせない重要なコミュニケーション手段
- ソーシャルメディアにおける「悪影響を及ぼす可能性のある投稿」が社会的脅威に

【悪影響を及ぼす可能性のある投稿の種類】

Disinformation （偽情報） ディープフェイク，生成AIによる画像・動画・文章など	Misinformation （誤情報） 視覚的ミスインフォメーション(Cherry-pickingなど)，誤解など	Mal-information （悪意の情報） ハラスメント，漏洩，ヘイトスピーチなど

Fake news

これらの情報及びそれが拡散されていく様 ＝ "デマ"（便宜上の呼称）

デマの対策と課題

- 対策
  - ✓ユーザ　　：リテラシー教育→内容の注意深い確認，通報
  - ✓プラットフォーマー：投稿の削除，アカウントの凍結措置
  - ✓ファクトチェッカー：事実確認の実施，結果の公開
  - ✓官公庁組織　　：不審な投稿に対する通報の受付，捜査活動，法的措置
    （アカウントの情報開示請求を可能に）
- 課題

  デマの選別→対策の判断は複雑 そして センシティブ

  対策が実行されるまでの間に多大な人的コストと時間を要する

  「デマを流す者・加担する者優位」な実態

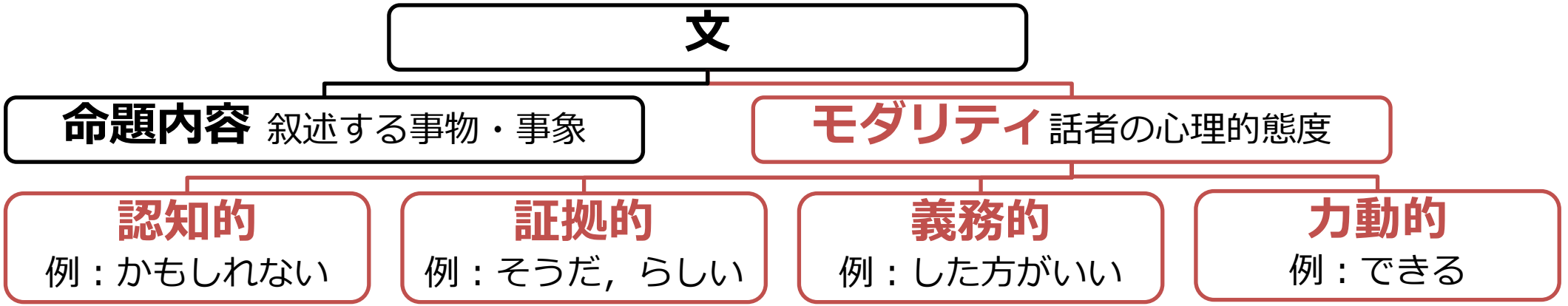
解決案：容易に検出可能かつ識別力が高い特徴量で デマである可能性が高い投稿をスクリーニング

● **投稿文の言語上の特徴「モダリティ(modality)」に着目**

✓**言語学における概念：話者の心理的態度**

例：新型コロナウィルス対策として、すぐに緑茶を飲んでください！！

新型コロナ対策として緑茶を飲む　＋　すぐに，ください，！！



Palmerらによるモダリティの分類

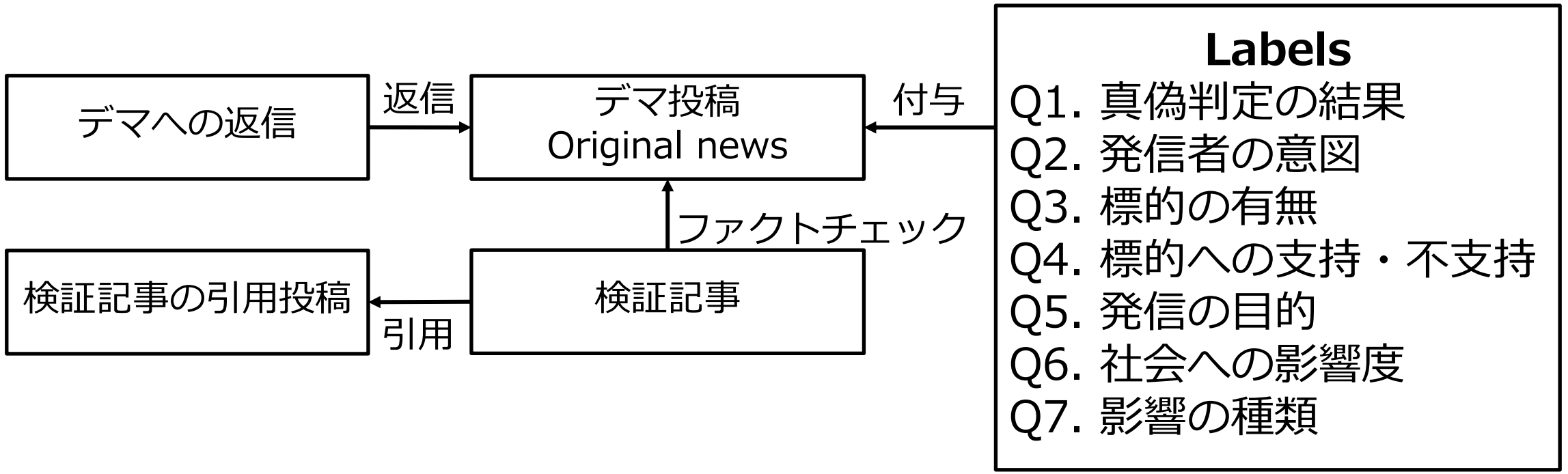
⇨ 仮説：特定のモダリティが込められた投稿は、特定の特徴をもつデマに関連することが多いのでは？

仮説が正しい場合、スクリーニングに応用可能（特定のデマ発見の効率化）

果たして、モダリティ検出はデマのスクリーニングに有効か？

- 目的：**文のモダリティがデマの検出に有用な情報であるか検証**
- データ：Japanese Fake News Detection Dataset(村山らが作成)を利用
- 手順
  1. **アノテーション**：ChatGPTを用い，モダリティラベルを付与
  2. **分析**：
     - デマの特徴(情報の真偽や発信の意図など)との有意差検定の実施
     - デマへのリアクション(返信数や引用された回数など)との相関分析の実施

Japanese Fake News Detection Dataset の構造



モダリティラベルの定義

	大分類	中分類	小分類		言語学的分類	例	定義
対事的	命題的	認知的	強	**断言**	断定・説明	のだ、わけだ、ことだ、からだ、つまり	ただ事実を述べているのではなく，揺るぎない主張・論理として強調している箇所
			弱	**推測**	推量・判断・可能性・蓋然性	だろう、にちがいない、はずだ、かもしれない、まい，でしょう、思う，きっと，たぶん	自分の意見を意見だとわかる形で強調して記述している箇所
		証拠的	強	**経験**	様態	（～し）そうだ	文脈から自分の経験を根拠として記述している箇所 ※断定的な言い方であっても，根拠としての活用であれば様態とする．
			弱	**伝聞**	伝聞	らしい、ようだ、そうだ、によると	自分は当事者ではないが，他社の経験をそのまま聞き伝えている箇所
	事象的	束縛的	強	**強制**	義務・命令・禁止	べきだ、ほうがよい、（ものだ、ことだ、）なければならない、なさい、てはいけない、こと	他者に対して行動を示唆，強要する箇所
			弱	**誘導**	許可・勧誘・依頼・誘い	てもいい、ましょう、てください、てくれ、していただけませんか，ほしい	他者に対して行動をお願いする箇所
		力動的		**意志**	意志	つもりだ，したい	ー
				能力	能力	できる	ー
対人的				**対人**	確認・強調	ね、よ，！	話口調に関する箇所
モダリティなし				**引用**			
				なし			

モダリティラベルの付与

ChatGPTへのプロンプト

- 人格の指定
- 作業の指示
- 出力形式の指定
- モダリティラベルの説明
- 例示

上記の文章について，あなたは言語学者として言語学的に正確に，モダリティのラベル付けを行ってください.
- ラベル付けは一文ずつ区切って行ってください.
- 一文に命題が二つ以上ある場合は，ラベルを複数付与して下さい.
- 出力は以下の形のcsv 形式とします.
==========
(一文),(ラベル),(確信度),(理由)
==========
(ラベル) には「断言，推測，経験，伝聞，強制，誘導，意志，能力，対人，引用，なし」のいずれかを入れてください.
特に(ラベル) には「批判，否定，危険」と回答しないでください.
(ラベル) は言語学上の分類と以下の様に対応しています.
・断言: 断定・説明・確信例）ことだ
・推測: 推量・判断・可能性・蓋然性・仮定例）かも，思う
・経験: 様態例）そうだ
・伝聞: 伝聞例）らしい，ようだ
・強制: 義務・命令・禁止例）すべき，～しろ
・誘導: 許可・勧誘・依頼・誘い例）してください
・意志: 意志・願望例）したい
・能力: 能力例）できる
・対人: 確認・強調・感嘆例）よ，ね
・引用: URL や引用した記事を示す場合例）http
・なし: 以上の定義のいずれにも当てはまらない場合
(確信度) には(ラベル) が正しいと思う度合いを0 から100 の数値で表してください.
(理由) にはラベルづけの理由を書いてください.
ラベル付けの例を示すので，参考にしてください.
＜例１＞
入力例：人民解放軍が出動したという噂が出たそうだけど、やはり本当だったよ。
回答例：
人民解放軍が出動したという噂が出たそうだけど、やはり本当だったよ。, 伝聞,90,「そうだ」は人から伝え聞いていることを示す表現であるため.
人民解放軍が出動したという噂が出たそうだけど、やはり本当だったよ。, 断言,90,「やはり本当だった」と確信する表現があるため.
人民解放軍が出動したという噂が出たそうだけど、やはり本当だったよ。, 対人,90,「よ」と主張を強調する対人的な表現があるため.
＜例２＞（以下，略）

ラベリング結果の例

NewsID	tweetID	text	ラベル	確信度	理由
5	1318466262796627968	本当にいい加減にして欲しい。	意志	90	「して欲しい」という表現は意志または要望を表しているため。
5	1318466262796627968	中国人技能実習生は一万人以上になる筈。	断言	80	「筈」と言う語は何かが期待されたり、予想される状態を強調するため断言と解釈される。
5	1318466262796627968	更に中国資本が所有する宿泊施設も多数。	断言	90	具体的な情報提供や説明と受け取れるため。
5	1318466262796627968	例えば村の人口の３割以上が中国人の占冠村などは中国のトマムリゾートがぼろ儲けしないのか？	推測	70	「ぼろ儲けしないのか？」という疑問形の表現から仮定や推測される問いかけがあるため。
5	1318466262796627968	【北海道は外国人技能実習生の水際対策の宿泊費を全額補助】	引用	95	URLが含まれているため、これは情報の引用であると判断される。
5	1318466655169622016	帰国させた方がいいですよ。	誘導	90	「～方がいいですよ」と他者に特定の行動を取ることを勧める表現であるため.
5	1318466655169622016	帰国させた方がいいですよ。	対人	80	「よ」という言葉を用いて、発言に対する相手の同意や注意を引く対人的な機能があるため.
5	1318467908662521856	信じられない対応。	断言	95	「信じられない」という強い表現が確信を示しているため.
5	1318467908662521856	というか，どんどん進めて日本を中国にしようとしていますね。	断言	90	「しようとしています」と確定的な表現を用いているため.

自動付与の性能評価

ラベル	正解率	適合率	再現率	F 値
断言	0.732	0.608	0.678	0.641
推測	0.903	0.480	0.615	0.539
経験	0.950	0.500	0.190	0.276
伝聞	0.893	0.607	0.333	0.430
強制	0.976	0.769	0.588	0.667
誘導	0.962	0.833	0.417	0.556
意志	0.979	0.636	0.583	0.609
能力	0.976	0.500	0.200	0.286
対人	0.725	0.913	0.156	0.266
引用	0.953	0.867	0.619	0.722
なし	0.853	0.154	0.091	0.114
all	0.900	0.609	0.415	0.494

分析

Original Newsに対する返信の早さが上位50までの投稿群に対して以下を分析

- **デマの特徴(情報の真偽や発信の意図)との有意差検定**
  - ✓投稿に含まれるモダリティの**出現割合**
  - ✓比較対象：Japanese Fake News Detection Dataset のLabels

- **デマへのリアクションとの相関分析**
  - ✓投稿に含まれるモダリティの**出現回数**と**出現割合**
  - ✓比較対象：リアクション数（リツイートの数，返信の数，いいねの数，引用された数，ブックマークされた数，閲覧された数）

結果：[推測]のモダリティについて有意差を確認

- 対象：返信の早さが上位50までの投稿群に含まれるモダリティの出現割合
- 結果：
  - ✓**情報の真偽**に関するラベルが「Misleading」の投稿で推測のモダリティの出現割合が特に高かった
  - ✓「False」と「Misleading」の間では有意差あり
- 考察：
  - ✓「～かも」など推量的な表現は，**事実ではないことを述べる表現と，事実を述べているが誤解を招く表現の区別に役立つ**
  - ✓確信を得ない表現が新たな推測を生みながら拡大していると考えられる

表 6: Q1: 情報の真偽と [推測] との多重比較の結果

ラベル	統計量 (p 値)
False:Inaccurate	0.875 (0.818)
False:Misleading	2.860 (0.022)
False:Unknown-Evidence	0.166 (0.999)
Inaccurate:Misleading	2.062 (0.166)
Inaccurate:Unknown-Evidence	0.373 (0.982)
Misleading:Unknown-Evidence	2.286 (0.101)

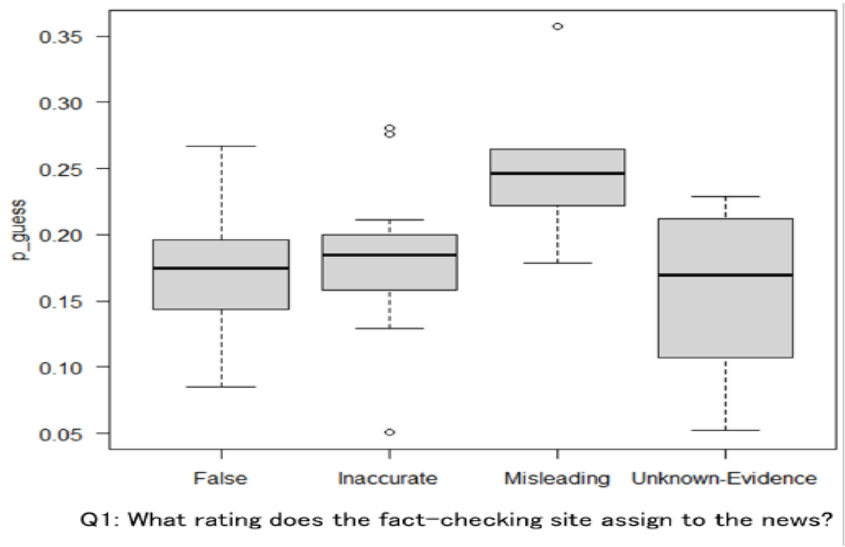


図 2: [推測] の出現割合の Q1 ラベルごとの分布

結果：[意志]のモダリティと発信の目的との有意差

- 対象：返信の早さが上位50までの投稿群に含まれるモダリティの出現割合
- 結果：
  - ✓**発信の目的**に関するラベルが「Propaganda」の投稿で意志のモダリティの出現割合が特に高かった
  - ✓「Partisan」と「Propaganda」の間では有意差あり
- 考察：
  - ✓PartisanとPropagandaの違いは，誘導の有無
  - ✓～したいなどの**発信者の願望を含む表現**は，社会など他者に対する願望を含むことから**誘導の有無を分けるのに有用**である可能性あり

表 7: Q5: 発信の目的と [意志] との多重比較の結果

ラベル	統計量 (p 値)
2.Partisan:3.Propaganda	2.466 (0.036)
2.Partisan:4.No purpose/Unknown	0.689 (0.770)
3.Propaganda:4.No purpose/Unknown	2.193 (0.072)

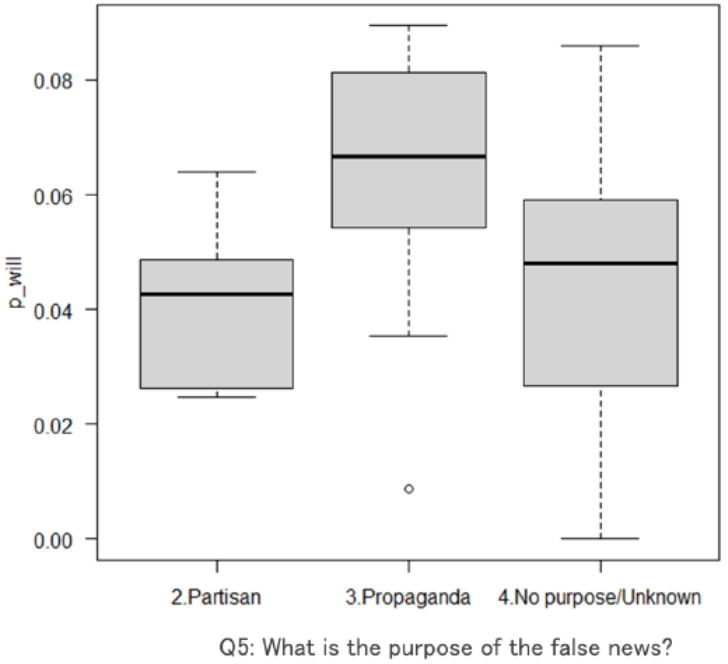


図 3: [意志] の出現割合の Q5 ラベルごとの分布

結果:リアクション数とモダリティとの相関分析

- 対象：返信の早さが上位50までの投稿群に含まれるモダリティの出現回数と出現割合
- 結果：**返信の数，いいねの数，引用された数，ブックマークされた数について相関**を確認
- 考察：
  - 「～して」など命令や禁止を呼び掛ける表現は，いいねとブックマークを得にくい
  - 他者の話を伝え聞いている表現は，引用とブックマークの数を得やすい　　など

表 8: モダリティと返信の数との相関

ラベル	First50_cnt	First50_per
断言	0.230 (0.120)	-0.156 (0.294)
推測	0.275 (0.061)	-0.079 (0.598)
経験	0.194 (0.192)	0.088 (0.558)
伝聞	0.193 (0.194)	0.017 (0.909)
強制	0.099 (0.509)	-0.087 (0.561)
誘導	0.486 (0.001)	0.356 (0.014)
意志	0.391 (0.007)	0.204 (0.169)
能力	0.230 (0.119)	0.151 (0.311)
対人	0.364 (0.012)	0.092 (0.540)
引用	-0.120 (0.421)	-0.335 (0.021)

表 9: モダリティといいねの数との相関

ラベル	First50_cnt	First50_per
断言	0.177 (0.233)	0.101 (0.501)
推測	0.223 (0.131)	-0.043 (0.772)
経験	0.267 (0.070)	0.199 (0.180)
伝聞	0.230 (0.120)	0.155 (0.297)
強制	-0.188 (0.205)	-0.305 (0.037)
誘導	0.235 (0.112)	0.161 (0.281)
意志	0.175 (0.238)	0.109 (0.467)
能力	-0.052 (0.727)	-0.077 (0.609)
対人	0.144 (0.333)	-0.025 (0.869)
引用	0.070 (0.640)	-0.076 (0.610)

表 10: モダリティと引用された数との相関

ラベル	First50_cnt	First50_per
断言	0.078 (0.602)	-0.192 (0.195)
推測	0.133 (0.372)	-0.165 (0.268)
経験	0.168 (0.260)	0.098 (0.511)
伝聞	0.340 (0.020)	0.249 (0.092)
強制	-0.110 (0.460)	-0.273 (0.063)
誘導	0.216 (0.145)	0.097 (0.516)
意志	0.181 (0.224)	0.083 (0.581)
能力	0.094 (0.528)	0.051 (0.734)
対人	0.406 (0.005)	0.220 (0.137)
引用	0.060 (0.690)	-0.103 (0.491)

表 11: モダリティとブックマークされた数との相関

ラベル	First50_cnt	First50_per
断言	-0.010 (0.949)	-0.119 (0.427)
推測	0.125 (0.403)	-0.109 (0.466)
経験	0.138 (0.353)	0.084 (0.574)
伝聞	0.397 (0.006)	0.370 (0.011)
強制	-0.234 (0.114)	-0.351 (0.016)
誘導	0.004 (0.976)	-0.079 (0.597)
意志	0.079 (0.596)	0.086 (0.567)
能力	-0.127 (0.394)	-0.140 (0.349)
対人	0.226 (0.126)	0.114 (0.446)
引用	0.274 (0.062)	0.162 (0.277)

まとめ

- デマ（Original+返信投稿）に含まれるモダリティの出現回数と出現割合がデマの特徴やリアクション数とどのように関係があるかを検証
- 投稿文の表層のみから一定程度のスクリーニングをする機能の実現に期待できる

- 検証結果：
  - ✓**発信者の確信度が低いことを表すモダリティ** → **事実の誤りと誤解を区別する**
  - ✓**発信者の願望を含むモダリティ** → **誘導の有無を区別する**
  - ✓**命令や禁止のモダリティ** → **いいね，ブックマークを得にくい**
  - ✓**他者の話の伝聞であることを表すモダリティ** →**引用によって拡散行動につながりやすい**

- 今後の課題：
  - ✓モダリティラベルの自動付与の性能の向上
  - ✓他の表層的特徴量を用いた検出モデルへの拡張
  - ✓確実にデマにはなりえない投稿文との分離性能の検証