

## 医薬品個人間取引が疑われる投稿の検出手法の検討

分担研究者 吉田 直子（金沢大学医薬保健研究域附属 AI ホスピタル・マクロシグナルダイナミクス研究開発センター）  
木村 和子（金沢大学大学院医薬保健学総合研究科）  
研究協力者 張 若愚（金沢大学大学院医薬保健総合研究科創薬科学専攻）  
山口 典枝（日本アイ・ビー・エム株式会社・グローバル・ビジネス・サービスパートナー）

### 研究要旨

【目的】医薬品等の不適正流通経路の一つとして、ソーシャルネットワーキングサービス（SNS）を介した個人間取引が問題となっている。本研究では、国民の保健衛生上の危害を未然に防止するため、取り締まりの対象となる医薬品取引の監視手法の検討に資することを目的に、クローリングプログラムを活用して、医薬品個人間取引が疑われる投稿を的確に検出する方法を立案する。

【方法】インターネット上のリンクを辿って web サイトを巡回し、web ページ上の情報を複製・保存するクローリングプログラムを使用して、「#お薬もぐもぐ」をキーワードに、日本時間で2023年1月1日0時00分から2023年2月28日23時59分までの期間にTwitterに投稿された情報を収集した（実施日：2023年3月20日）。収集された投稿内容について、テキストマイニングにより出現単語を抽出した。医薬品の個人間取引が疑われる投稿とそれ以外の投稿における出現単語の特徴を把握するため、決定木分析を行った。

【結果・考察】前年度に開発したクローリングプログラムを絵文字も抽出できるように修正し、本研究に用いた。「#お薬もぐもぐ」をキーワードに期間内に収集された投稿3,894件中、1,651件が医薬品の個人間取引が疑われる投稿であった。全3,894件の投稿から、7,421個の単語が抽出された。出現頻度として、キーワードとした「#お薬もぐもぐ」と絵文字が文字化けした特殊文字が上位2位を占め、次いで、「DM」、「お薬チャーム」、「薬」、「RT」、「#サイレース」、「もぐもぐ」、「#ブロン」などが高頻度で使用されていた。決定木分析の結果、医薬品の個人間取引が疑われるか否かを予測するにあたり、最も重要度の高い単語として「DM」が抽出され、「DM」の記載があり、「#お薬チャーム」の記載がなく、かつ「#サイレース」の記載があることが医薬品の個人間取引が疑われる投稿に使用される単語の特徴の1パターンとして示され

た。

【結論】医薬品の個人間取引が疑われる投稿を検出するための手法として、クローリングプログラムが活用でき、本研究において、クローリングプログラムを実行する際のキーワードの選定方法が考案された。

## A. 研究目的

ソーシャルネットワーキングサービス (social networking service, SNS) は、インターネットを通じて他者とコミュニケーションをとるプラットフォームであり、私たちの日常生活の中で広く活用されている。しかし、このサービスの利便性は、情報の過剰な拡散を招いている。また、その便利さゆえに、情報の氾濫を規制することは難しく、SNS は一部の不適切な取引の温床となっている。

利益や利便性を優先し、SNS 上で法に抵触する行為を行う人もいる。医薬品の無許可販売もその一つであり、業としての取引の隠れ蓑になっている可能性も指摘されている。これまでに、向精神薬などの他、未承認医薬品などが SNS 上で個人間取引されていたことが確認されている。SNS を介して取引される医薬品等は、正規流通経路を逸脱していることから、当然、適切な保管状態は保障されず、低品質医薬品や偽造医薬品の混在リスクも孕む。国民の保健衛生上の危害を未然に防止するため、個人間取引の実態を把握し、積極的に監視指導を行うことが必要である。

本研究では、取り締まりの対象となる医薬品取引の監視手法の検討に資することを目的に、クローリングプログラムを活用して、医薬品個人間取引が疑われる投稿を的確に検出する方法の開発を試みた。

## B. 研究方法

### B-1. 対象

本研究では、主に利用されている SNS のである Twitter と YouTube における投稿を対象とした。

### B-2. クローリング

クローリングとは、プログラムがインターネット上のリンクを辿って web サイトを巡回し、web ページ上の情報を複製・保存する方法であり、日々発生している情報を収集し活用すること等を目的に様々な分野で用いられている。

クローリングによる API (application programming interface) 利用について、Twitter と YouTube に対し、API 利用制限解除を申請し、研究目的での利用について承認を受けて、本研究を実施した (承認日: Twitter, 2022 年 11 月 29 日; YouTube, 2023 年 1 月 23 日)。前年度に Twitter と YouTube それぞれを対象に開発したクローリングプログラム (日本アイ・ビー・エム株式会社, 東京, 日本) を修正し、本研究に用いた。Twitter を対象に、医薬品取引が疑われる投稿に高頻度に用いられているハッシュタグ「#お薬もぐもぐ」をキーワードとして、2023 年 3 月 20 日時点で 2023 年 1 月 1 日 0 時 00 分から 2023 年 2 月 28 日 23 時 59 分 (日本標準時) までの期間に投稿され、削除されていない投稿情報を収集した。収集した投稿情報を個人間取引が疑われる否かで定義し、さら

に、個人間取引が疑われる投稿を譲受または譲渡に区別した。収集されたテキスト情報について、テキストマイニングソフトウェア MATLAB Text Analytics Toolbox (MathWorks, Inc., MA, USA) を用いて、単語を抽出した。抽出された単語について、出現頻度を集計・解析した。

### B-3. 使用される単語の特徴抽出

個人間取引が疑われる投稿において特徴的に使用されていた単語を抽出するため、データマイニングソフトウェア (SPSS Modeler version 18.3, 日本アイ・ビー・エム株式会社, 東京, 日本) を用いて決定木分析を行った。ホールドアウト検証を目的として、ランダムに全データを学習データとテストデータに区分し、モデルの汎化性能を評価した。本研究では、学習データ : テストデータ = 70 : 30 とした。分析方法として、決定木分析の 1 つで、カイ 2 乗検定や F 検定に基づいてツリーモデルを構築する chi-square automatic interaction detection (CHAID) を用いた。

## C. 結果

### C-1. クローリングプログラム修正

Twitter 上の医薬品の個人間取引が疑われる投稿において、絵文字が、当て字として、高頻度で使用されている実態が明らかになったことから、昨年度開発した Twitter と YouTube と対象としたクローリングプログラムを修正し、絵文字も抽出できるプログラムに改良した。当該改良プログラムの実行により、絵文字を抽出できることを確認した (Table1)。

### C-2. Twitter のクローリング

これまでの調査において、個人間取引が疑われる投稿には、ハッシュタグ「#お薬もぐもぐ」が高頻度で使用されていた。本研究では、「#お薬もぐもぐ」をキーワードとしてクローリングプログラムを実行し、Twitter における投稿情報を収集した。「#お薬もぐもぐ」をキーワードに抽出された投稿は、3,894 件であった。そのうち、個人間取引が疑われた投稿は 1,651 件 (42%) で、残りの 2,243 件 (58%) においては、個人間取引が疑われる記述は見当たらなかった。個人間取引が疑われた投稿は 1,677 件のうち、譲渡に係る投稿が 1,199 件で、譲受に係る投稿が 478 件であった。

### C-3. 特徴的に使用される単語の抽出

全 3,894 件の投稿について、形態素解析により当該投稿で使用されていた単語を抽出した結果、抽出された単語数は計 7,421 個であった。抽出された単語について、医薬品の個人間取引が疑われた投稿において高頻度に出現した単語上位 100 単語を Table 1 に示した。「クローリング実行時のキーワード「#お薬もぐもぐ」の出現頻度が 1,573 回で、最も高頻度となり、第 2 位は、一部の絵文字が文字化けした特殊記号「𐀀」の 726 回であった。文字化けした例を Table 3 に記した。第 3 位以下は、「DM」603 回、「RT」505 回、「#サイレース」456 回、「#リリカ」450 回、「#マイスリー」447 回などとなった。

全 3,894 件の投稿を「医薬品個人間取引疑いなし (n=2,243) : 0」と「医薬品個人間取引疑いあり (n=1,651) : 1」に定義して目的変数とし、決定木分析を行っ

た。予測変数として重要度の高い方から順に、「DM」、「お薬チャーム」、「薬」、「RT」、「#サイレース」、「もぐもぐ」、「#ブロン」が抽出された (Figure 2)。学習データ (n=2,670) から得られた決定木を Figure 3 に示した。「医薬品個人間取引疑いあり」の投稿に使用される単語の特徴の 1 パターンとして、「DM」の記載がある投稿における個人間取引率 (413/472, 87.5%) が有意に高く ( $p<0.001$ )、そのうち「#お薬チャーム」の記載がない投稿では「#サイレース」の記載頻度が高いことが示された。一方で、「医薬品個人間取引疑いなし」の投稿においては、「DM」の記載頻度が低く、「薬」がより高頻度に記載されている特徴が示された。この結果を検証した結果、学習データとテストデータ (n=1,223) における正解率は、それぞれ 72.7%、と 72.2%であった (Table 4)。予測モデルの評価指標として、AUC は約 0.83、Gini 係数は 0.66 が得られた (Table 5)。全 3,894 件に対し、学習データとテストデータの割付から決定木分析までを 3 回繰り返したところ、3 回とも同様の結果が得られた。

#### C-4. YouTube

YouTube を対象に「お薬もぐもぐ」をキーワードとしてクローリングを実行したが、医薬品の個人間取引に関わる投稿は検出されなかった。YouTube における医薬品の個人間取引事例のハンドサーチによる検索も並行して行い、手掛かりを探している。

#### D. 考察

本研究では、昨年度開発したクローリングプログラムを絵文字も抽出できるプログラムに改良し、絵文字も含めて、より多くの情報を収集できるプログラムが整備された。当該プログラムを用いて、医薬品の個人間取引に頻繁に用いられるハッシュタグ「#お薬もぐもぐ」をキーワードに、Twitter における投稿情報を収集した。本研究では、2023 年 1 月 1 日から 2023 年 2 月 28 日 までの期間を対象にクローリングを実行した結果、2 ヶ月分の情報で、医薬品の個人間取引に関わる投稿を 1,651 件と医薬品の個人間取引に関わる記述のない投稿 (n=2,243) を収集した。これらの投稿情報について、テキストマイニングにより出現単語を抽出し、決定木分析を行うことで、医薬品の個人間取引に関わる投稿において使用される単語の特徴抽出が可能であることを実証した。有意に高頻度に記載されていた「DM」や「#ハルシオン」などの単語はキーワードとして、有意に記載頻度が低かった「お薬チャーム」等の単語は、除外キーワードの候補となる。これらを組み合わせてクローリングを実行することにより、よりの確に医薬品取引に関する投稿を検出できる可能性がある。また、医薬品の個人間取引が疑われない投稿において、特徴的に使用される単語と使用されない単語も抽出されたことから、クローリング実行後のデータ処理や精査に当該ロジックを適用して、除外すべき投稿を定義することも可能かもしれない。引き続き、「#お薬もぐもぐ」や抽出された単語をキーワードにクローリングを実行し、データセットを拡充するとともに、医薬品個人間取引が疑われる投

稿の検出法の精度向上に向けて、検討を行う。

一方で、クローリングプログラムは一部を除き (Table 3)、絵文字も抽出できるよう改良されたが (Table 1, 2)、テキストマイニングの際に、絵文字のままでは認識されず、決定木分析では正しくデータに取り込まれていない可能性が示された。別途、多変量解析等を行うことにより、欠損データについて、適切に分析されるよう加工等して、解析を進める予定である。

本結果のデータ収集期間は 2 ヶ月であり、その間にプレゼント企画等で、リツイート数 (譲受者側の投稿) が一時的に増加することで、情報に偏りが生じた可能性がある。今後、バイアスの最小化を考慮して、過去と未来の情報をさらに追加してデータセットを作成し、検出ロジックの検討を行う。

## E. 結論

本研究において、医薬品の個人間取引が疑われる投稿を検出するために、クローリングプログラムを実行する際のキーワードの選定方法が考案された。今後、投稿情報をさらに収集することでデータ数を増やし、高精度に当該投稿を検出するロジックの検討を進める。

## F. 研究発表

- 1.論文発表なし
- 2.学会発表なし

## G. 参考文献

なし

Table 1. 改良クローリングプログラムにより絵文字が含まれる投稿情報を収集した事例

投稿日時	ツイッター投稿内容
2023-02-28T07:08:14.000Z	(アカウント名) : 🦷威布論 20st 🦋 600/1st 10st で 5000 詳しくはテレもしくはDM までお願い します。 #お薬もぐもぐ
2023-02-28T04:38:27.000Z	🦷威布論 20st 🦋 600/1st 10st で 5000 詳しくはテレもしくはDM までお願い します。 #お薬もぐもぐ
2023-02-27T15:32:46.000Z	#マイスリー  🦋ます。 #お薬もぐもぐ
2023-02-27T15:01:11.000Z	🐘2 9st デパ0.5 20st 🦋 DM☑ください #お薬もぐもぐ
2023-02-25T17:23:26.000Z	(アカウント名) #お薬もぐもぐ ロラゼパム(ワイパックス)1mg 🦋

2023-02-25T01:17:49.000Z	<ul style="list-style-type: none"> <li>・デパス</li> <li>・サイレース</li> <li>・フルニトラゼパム</li> <li>・マイスリー</li> <li>・ハルシオン</li> </ul> <p>🍷実績も沢山🍷  #おくすりもぐもぐ  #お薬譲ります  #お薬もぐもぐ</p>
2023-02-24T14:21:00.000Z	<p>🍷  DM 下さい 😊  ふ リマ 使えます！</p> <p>#お薬もぐもぐ  #デパス (特定 URL)</p>
2023-02-23T16:53:53.000Z	<p>ロラゼパム (ワイパックス)  レキサルティ 🍷☑</p> <p>1st900 円</p> <p>#お薬もぐもぐ</p>
2023-02-22T09:20:31.000Z	<p>#お薬もぐもぐ  サイレース  🍷</p>

Table 2. 医薬品の個人間取引が疑われる投稿において使用されていた単語とその出現頻度（上位 100 単語）

順番	単語	出現回数
1	#お薬もぐもぐ	1573
2	□	726
3	DM	603
4	RT	505
5	#サイレース	456
6	#リリカ	450
7	#マイスリー	447
8	#ハルシオン	433
9	にて	423
10	#ゾルピデム	419
11	各種	411
12	取引	406
13	実績	343
14	1 5 0 OD	335
15	#デパス	302
16	糞	250
17	多数	240
18	詐欺	234
19	コンサ	188
20	5	174
21	m	174
22	konsa	174
23	出せる	156
24	(アカウント名)	144
25	5m	133
26	(アカウント名)	132
27	ください	131
28	デ	131
29	てる	130
30	薬	125
31	お願い	120
32	ツイート	118



33	lt	115
34	マイ	112
35	下さい	109
36	ば	107
37	パス	103
38	サイ	100
39	(アカウント名)	95
40	レース	93
41	スリー	89
42	□	88
43	#おくすりもぐもぐ	84
44	余っ	84
45	2	70
46	20	69
47	れ	68
48	1	67
49	ご	66
50	けど	61
51	#お薬譲ります	61
52	可能	57
53	欲しい	56
54	師	54
55	たら	53
56	き	53
57	(アカウント名)	49
58	なあ	47
59	手押し	46
60	◆	45
61	求	43
62	いく	43
63	(アカウント名)	43
64	連絡	42
65	もぐもぐ	41
66	ー	41
67	プレゼント	41

68	リップ	41
69	シート	40
70	10	39
71	くれ	39
72	譲っ	39
73	企画	39
74	コンサータ	38
75	在庫	38
76	相場	37
77	られ	37
78	協力	36
79	ござい	35
80	譲り	35
81	錠	34
82	つけ	34
83	安く	34
84	1mg	33
85	#リタリン	33
86	ハル	32
87	行き	32
88	いい	32
89	フルニトラゼパム	32
90	対応	31
91	致し	31
92	Twitter	31
93	リツイート	31
94	気軽	30
95	探し	30
96	05	29
97	発送	29
98	くれる	29
99	んで	29
100	更新	29

全抽出単語数 n=7,421

Table 3. 特殊記号「□」に変換された絵文字の例

ツイート本文	クローリング集計表示
リタリン欲しい、、、🥰🥰🥰 #お薬もぐもぐ	リタリン欲しい、、、□□□ #お薬もぐもぐ
夜の薬食べた👉🥰👉 #お薬もぐもぐ	夜の薬食べた□□□ #お薬もぐもぐ
朝の薬食べた👉🥰👉 #お薬もぐもぐ	朝の薬食べた👉👉 #お薬もぐもぐ
#イーケプラ #デパケンR どんなもんかおしえてください🥰 #お薬もぐもぐ	#イーケプラ #デパケンR どんなもんかおしえてください□ #お薬もぐもぐ
リリカ(ブレガバリン)75mgは どうなんだろう...🤔🗨️ #お薬もぐもぐ	リリカ(ブレガバリン)75mgは どうなんだろう...□□ #お薬もぐもぐ

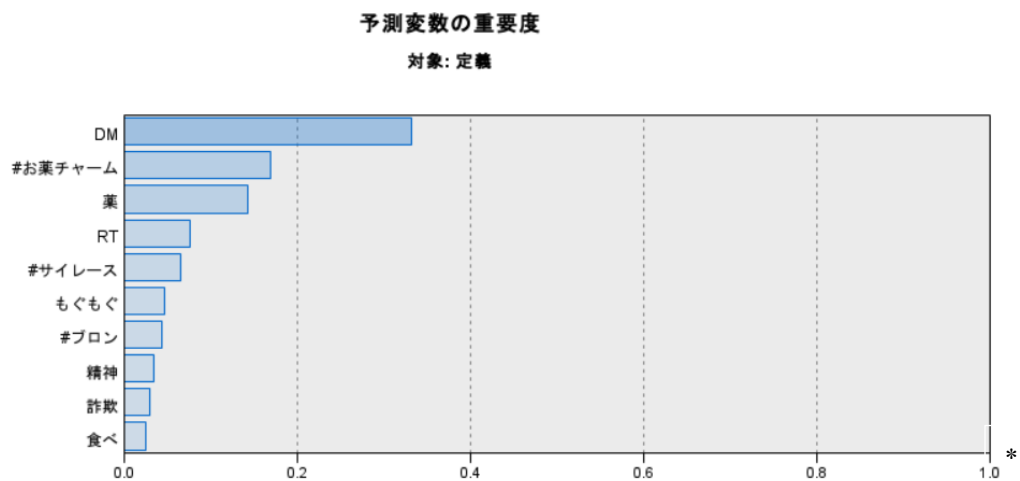
Table 4. 検証結果

	学習	精度	テスト	精度
正解	1,940	72.66%	883	72.2%
誤り	730	27.34%	340	27.8%
合計	2,670		1,223	

Table 5. 検証-評価メトリック

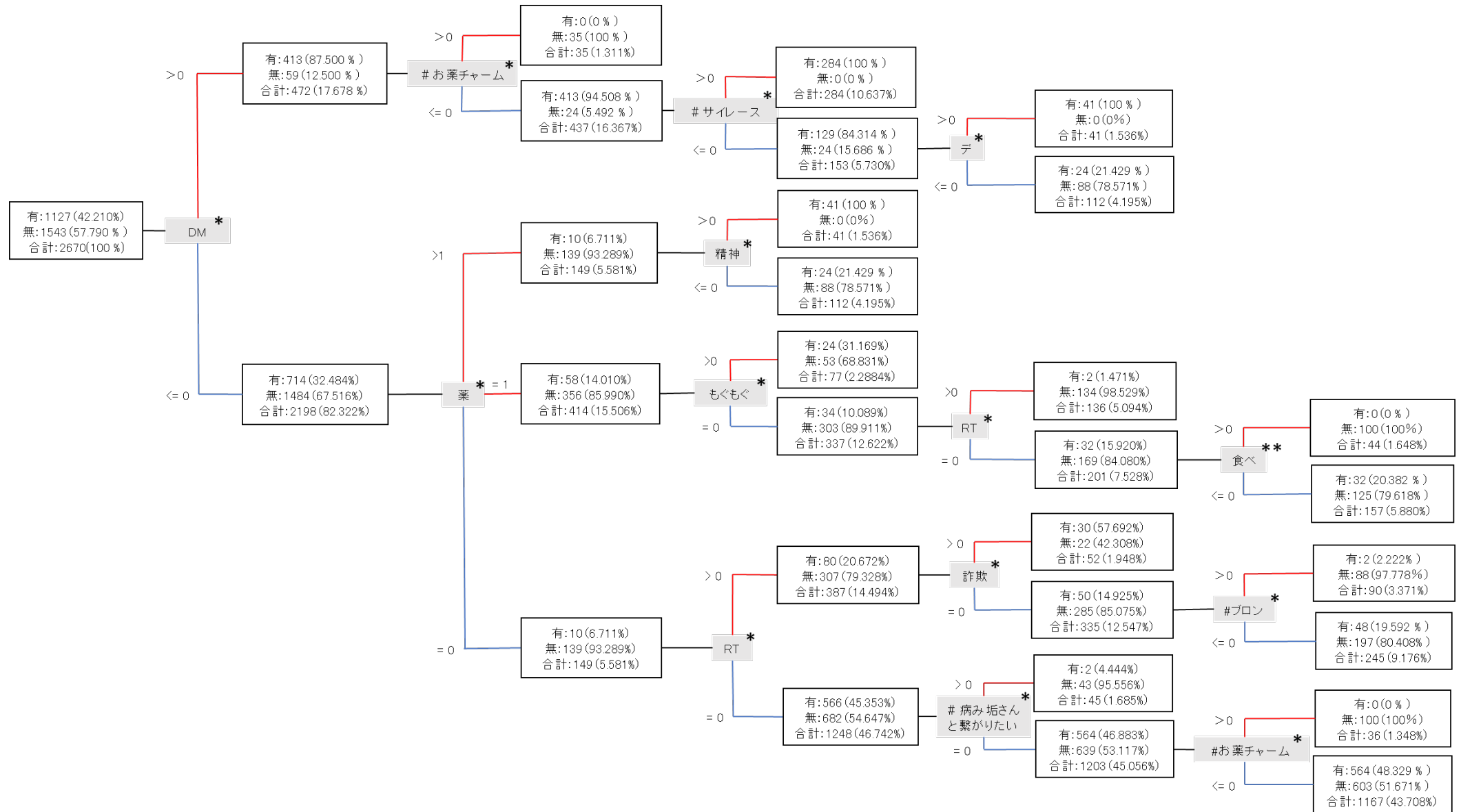
	学習		テスト	
指標	AUC	Gini	AUC	Gini
値	0.828	0.655	0.833	0.665

Figure 1. CHAID において医薬品の個人間取引疑いの有無の予測において重要度の高かった単語（上位 10 単語）



\* : 0.0-1.0 は重要度、数値は大きいほど、重要度が高い。

Figure 2. CHAID で得られた決定木



有：医薬品個人間取引疑いあり（SNS 上での個人間取引が疑われる投稿）

無：医薬品個人間取引疑いなし（SNS 上での個人間取引が疑われない投稿）

\*：  $p < 0.001$

\*\*：  $p$  値 = 0.001

