B-005

ドメイン知識強化型混合エキスパートモデルによるフェイクニュース検出手法の提案
An Approach to Identifying Fake News Using Domain Knowledge-Enhanced MoE

李 姣霖† 児玉 英一郎† 王 家宏† Bhed Bahadur Bista† 高田 豊雄†
LI JIAOLIN KODAMA Eiichiro WANG JIAHONG Bhed Bahadur Bista TAKATA Toyoo

1. はじめに

近年の IT 技術の進歩やソーシャルメディアの普及に伴い，フェイクニュースの流布がますます容易になっている．フェイクニュースとは，「センセーショナル性を持ち，広告収入や著名人・政治運動・企業などの信用失墜を目的としたオンライン上で広く共有されるように作成された偽のニュース記事」のことである[1]．特に最近は，新型コロナウイルス禍のなかで，例えば「お酒がコロナウイルスの新規感染予防になる」や「5G ネットワーク通信は新型コロナウイルスの感染経路の一つである」など，様々なフェイクニュースが拡散されていた．フェイクニュースが拡散されることで社会的不安や恐怖が煽られ，人々が不安や恐怖を感じるようになり，社会が不安定になり，個人と社会に壊滅的な影響を与える可能性がある．そのため，フェイクニュースの検出は，重要な課題になっている．

フェイクニュースの検出には，サポートベクターマシンやニューラルネットワークなどを用いた機械学習を用いたの手法，ニュースソースの信頼度評価を用いた手法，また，ソーシャルメディア上の情報伝播パターンを利用した手法などの 3 種類の代表的な手法が知られている．機械学習を用いたの手法は最も有効で広く利用されている．機械学習を用いたフェイクニュースの検出手法としては，シングルドメインフェイクニュース検出手法とマルチドメインフェイクニュース検出手法が知られている．前者は，検出対象となったニュースがフェイクニュースであるかどうかの判別の際に，そのニュースの属するカテゴリを考慮することなく，政治や経済，スポーツなど事前に定められたニュースドメインと呼ばれる一つのカテゴリだけに注目（または全カテゴリのニュースを１つのドメインに属すると仮定）して，フェイクニュースの検出を行うものである．後者は，YAHOO!ニュースなどのニュースサイトにおけるフェイクニュースの検出を想定し，ニュースが{政治，経済，スポーツ，…}といったドメイン集合のなかの一つのドメインに投稿されると仮定し，＜ニュース，ドメイン＞の形で与えられたニュースがフェイクニュースであるかどうかについての判別を行うものである．

本論文では，機械学習を用いたマルチドメインフェイクニュース検出を対称として，エキスパートのドメイン知識を強化した混合エキスパートモデル（Mixture of Experts，MoE）[2][3][4][5][6]によるフェイクニュースの検出手法の提案を行う．

†岩手県立大学大学院ソフトウェア情報学研究科
Graduate School of Software and Information Science, Iwate Prefectural University

2. 関連研究

マルチドメインフェイクニュース検出方法の一つとして MoE を利用した MDFEND （Multi-Domain Fake News Detection）法[7]が知られている．

MDFEND の基本的な考え方は，入力としてのニュース全体がドメインに割り振られることを考慮し，フェイクニュース検出の精度を高めるため，MoE を利用して，MoE 内のそれぞれのエキスパートが，それぞれ異なるドメイン知識からニュースを解析し，それぞれの専門知識を組み合わせて最終判断を行うといったものである．シングルドメインフェイクニュース検出手法と比べて，ドメイン知識の利用により大幅な性能改善が得られている．

MDFEND では，一件のニュースが，政治や経済，スポーツなどのドメインの内，ただ一つのドメインに属するといったことを仮定している．しかし実際には，一件のニュースが２つ以上のドメインの内容を含む場合が多い．例えば，「アメリカが今年，石油価格の下落を受けて、戦略的な石油備蓄を補充するために 1,200 万バレルの石油を購入する」[8]といったニュースは，政治と経済の二つのドメインに属している．この点を解消するための MDFEND の改善法として，FuDFEND[9]が提案されている．FuDFEND はファジィ推論機構を MDFEND に導入し，MDFEND の検出精度の向上を行った．

本論文も FuDFEND と同じく，上述した MDFEND の問題点の解消を目指す．本論文での提案手法の基本的な考え方は以下の通りである．学習モデルを作成する際に，ある学習データが政治と経済といった２つのドメイン内容を含んでおり，政治的内容が 95%，経済的内容が 5%であった場合には，政治的コンテンツだけを利用し学習モデルの訓練を行い，MoE の各々のエキスパートのドメイン知識を強化する．これにより，MoE によるフェイクニュースの効果的検出を行う．

3. エキスパートのドメイン知識を強化した MoE によるフェイクニュース検出手法

本提案手法のモデルを図 1 に示す．図 1 内の太線と点線で囲まれた部分が本提案独自の部分となる．

本モデルは MoE の利点を活用することにより，エキスパート$Expert_i$の集合を使用してニュースの様々な表現を抽出し，ニュースコンテンツの複雑な特徴を複数のエキスパートで把握する．つまり，複数の異なるドメインに特化したエキスパートが異なる観点から問題を解析し，それぞれのドメイン知識を組み合わせて最終的な判断を行うものである．

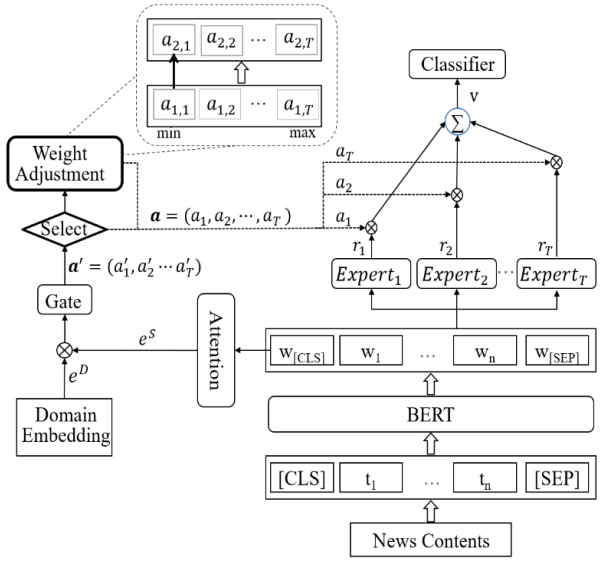

図 1 ドメイン知識強化型混合エキスパートモデルによるフェイクニュース検出手法

3.1 本提案手法の基本的な考え方（エントロピー視点）

エキスパートの総数をT，エキスパートの集合を$\mathrm{MoE} = \{ Expert_i \}_{i=1,2,\cdots T}$とする．MoE の各要素である$Expert_i$の分類器 Classifier への貢献度の差を調べるために，MDFEND のプロトタイプシステム[7]を用いて予備試験を行った．その結果，ほとんど全てのトレーニングサンプルに対して，大きな差を観測することができなかった．これは MoE の本来の効果が十分に発揮できていないことを意味していると考える．MoE に期待することは，異なるエキスパートが異なる領域（ドメイン）に特化し，能力を発揮することである．どのようなデータに対しても各エキスパートが同程度の貢献度となるようであれば，利点は感じられない．そこで著者らは，MoE の各エキスパートに付随するドメイン知識，つまり，得意なドメイン群の強化を行えば，検出精度を高めることができるのではないかと考えた．この考え方の実現方法として，分布のランダム性を測る指標のエントロピーの活用を試みる．即ち，MoE のゲーティングネットワーク（図 1：Gate を参照）の出力分布のエントロピーをより低い方向へ調整することにより，ゲーティングネットワークの出力により偏りを持たせ，学習モデルのトレーニングを行う．以下に具体的な手順を示す．

トレーニングサンプルの集合をX，第j番目のトレーニングサンプルを$X_j \in X$とする．X_jに対するゲーティングネットワーク(Gate)からの出力を$\boldsymbol{a^j}$，$\boldsymbol{a^j}$の第i成分（第i番目エキスパート$Expert_i$に対する第j番目のトレーニングサンプルの重要性を表す重み付け値）をa_i^jとする．従って，第j番目のトレーニングサンプルX_jに対する$\boldsymbol{a^j}$のエントロピーは式(1)で表すことができる．

$$H(\boldsymbol{a^j}) = \sum_{i=1}^{T} \left(-a_i^j \log\left(a_i^j\right) \right) \qquad (1)$$

式(1)において，エキスパートを適切に選び，対応したa_i^jを無効(0)に設定すれば，$H(\boldsymbol{a^j})$が減少する方向へ調整でき，MoE のエキスパートのドメイン知識の強化が行えると考える．

3.2 エキスパートの専門知識の強化手法

本節では，図 1 のモデルの説明（詳細は Nan Qiong らの論文[7]を参照）に加え，我々の提案手法の説明を行う．

学習モデルの作成において，先ず，ニュース文を BERT 専用トークナイザーでトークン化する．その結果（[CLS]，t_1，…，t_n，[SEP]）は BERT モデルに渡され，入力シーケンスの各トークンに対応する単語埋め込み表現ベクトル($w_{[\mathrm{CLS}]}$，w_1，…，w_n，

$w_{[SEP]}$）が生成される．このベクトルには，トークンの意味的な表現や文脈の情報が埋め込まれている．

続いて，単語埋め込み表現ベクトルは Mask-Attention ネットワークにより処理され，ニュース文の特徴を表す文埋め込みベクトル$\boldsymbol{e}^s$が得られる．この際，同様に，単語埋め込み表現ベクトルは Domain Embedding ネットワークにより処理され，ニュースの属するドメインを表現したドメイン埋め込みベクトル$\boldsymbol{e}^D$が生成される．

その後，エキスパートの選択を行うドメインゲーティングネットワーク（Gate）にドメイン埋め込みベクトル$\boldsymbol{e}^D$と文埋め込みベクトル$\boldsymbol{e}^s$が渡され，各エキスパートの重要性を示す重み付けベクトル$\boldsymbol{a}' = (a_1', a_2' \cdots a_T')$が生成される．この重み付けベクトルは，選択プロセスにおいて各エキスパートの貢献度や信頼度を反映したものとなっている．この重みに基づいて意思決定が行われる．つまり，ドメインゲーティングネットワークはドメインに応じて，エキスパートの重み付けを動的に調整する役割を担っている．

そして，特定ドメインへの集中度を高め，ニュース内容の集約性を高めるために$H(\boldsymbol{a}^j)$が減少するよう調整する．そのため，エキスパートを適当に選択し，対応したa_i^jを無効(0) に設定する．この際，どのトレーニングサンプルに対してこのような調整を行うべきかは，チューニングを行う必要がある．仮に重み付けベクトルの分散が大きくない場合には，検出精度に対するマイナスの影響が生じてしまう．

調整対象の重み付けベクトル$\boldsymbol{a}' = (a_1', a_2' \cdots a_T')$を選択するために，その標準偏差値を計算する．予め設定された閾値を超えた重み付けベクトルのみを調整対象とし，Weight Adjustment に入力する．続いて，調整対象となった重み付けベクトル$\boldsymbol{a}' = (a_1', a_2', \cdots, a_T')$の成分を昇順にソーティングし，$\boldsymbol{a_1} = (a_{1,1}, a_{1,2}, \cdots, a_{1,T})$を生成する．$\boldsymbol{a_1}$の第 1 成分から第 m 成分までを無効 (0) に設定し，$\boldsymbol{a_2} = (a_{2,1}, a_{2,2}, \cdots, a_{2,T}), a_{2,1} = a_{2,2} = \cdots = a_{2,m} = 0$を生成後，$\boldsymbol{a_2}$の成分をソーティング前の順($\boldsymbol{a}'$の成分順)に並び替え$\boldsymbol{a} = (a_1, a_2, \cdots, a_T)$として出力する．なお，$m$の値は予め実験にて決定し，パラメータ調整として行う．$\boldsymbol{a}$は最終的な異なるエキスパートの重要性を示す重みづけベクトルである．

以上により，エキスパートの専門知識の強化を行い，よりドメイン知識に特化したエキスパートネットワークを構成する．

次に，各エキスパートネットワークの出力$r_i (1 \leq i \leq T)$とマージして，ニュースの最終的なニュース特徴ベクトルを次式により算出する．

$$\mathrm{v} = \sum_{i=1}^{T} a_i \, r_i$$

最後に，このニュース特徴ベクトル v を分類器 Classifier にかけ，フェイクニュースかリアルニュースかの判別を行う．

3.3 調整時の閾値と成分数決定アルゴリズム

3.2 節で述べた検出手法では，標準偏差値の閾値を超えた場合，調整対象の重み付けベクトルとして選択し，調整を行っている．以下に，この調整で使用する閾値（Qと記す）と調整対象の重み付けベクトル内の調整する成分の数（mと記す）の決定アルゴリズムを示す．以下，評価指標のF値をFSと記す．

INPUT： データセット D，標準偏差値の範囲$[S_1, S_2]$，増分ΔK，エキスパートの総数T
OUTPUT： 閾値Q，成分数m
Step 1. 初期化： $FS \leftarrow 0$
Step 2. 成分数m'を順次に $1, 2, \ldots,$ Tに設定し，Step3 を行う．
Step 3. 標準偏差 S を順次に $S_1, S_1 + \Delta K, S_1 + 2\Delta K, \ldots, S_2$に設定し，Step4 を行う．
Step 4. m'とS を用いて実験を行い，F値を求める．もしF値 $> FS$， $FS \leftarrow$ F値， $Q \leftarrow S$， $m \leftarrow m'$.
Step 5. Qとmをそれぞれ標準偏差値の閾値と調整の対象となる成分の個数として出力する．

我々は，一般的に標準偏差値の閾値と調整対象となる成分の個数が，検出の対象となるニュースデータの構成に依存すると考えている．作成された判別モデルの実運用をする際，検出の対象となるニュースデータの構成は一定となるので，上記アルゴリズムにより，その最適化された値を予め決められると考える．

以下，本アルゴリズムの動作例を示す.

標準偏差値の範囲$[0.0,\ 0.3]$，増分 $0.025(\Delta K = 0.025)$，エキスパートの総数$5(T = 5)$とする.

Step 1. 初期化： $FS \leftarrow 0$

Step 2. 成分数m'を順次に $1, 2, \ldots, 5$に設定し，Step3 を行う.

Step 3. 標準偏差 S を順次に $0.0, 0.0 + 0.025, 0.0 + 0.05 \ldots, 0.3$に設定し，Step4 を行う.

Step 4. m'とS を用いて実験を行い，F値を求める．もし F 値 $>$ FS， $FS \leftarrow$ F 値， $Q \leftarrow S$，$m \leftarrow m'$.

Step 5. Qとmをそれぞれ標準偏差値の閾値と調整の対象となる成分の個数として出力する.

4. 性能評価と考察

本提案手法の有効性を確認するため，性能評価実験を行った．関連研究との性能差の比較を行うため，エキスパートの数Tを関連研究と同じく $5(T = 5)$に設定した．本節では，実験結果を示し，考察を行う.

4.1 利用したデータセットと性能評価指標

本性能評価では，Nan Qiong らの使用したデータセット[7]（データセット 1 と記す）とインターネットからダウンロードした英語データセット(データセット 2 と記す) を利用した.

データセット 1 は，Weibo[10]上の 2014 年 12 月から 2021 年 3 月までの科学・軍事・教育・災害・政治・健康・財経・娯楽・社会ドメインの 4,488 件のフェイクニュースと 4,640 件のリアルニュース，合計 9,128 件のニュースを格納している．train データ 5,751 件，validation データ 1,918 件を用い，学習，パラメータの調整後，test データ 1,923 件を使い本評価結果の算出を行った.

データセット 2 は，GitHub や IEEE DataPort からダウンロードした３つのフェイクニュースに関するデータセットをマージした英語のマルチドメインデータセット[11][12][13][14][15]である．このデータセットは，技術，セレブ，教育，政治，経済，スポーツ，娯楽，Covid19（新型コロナウイルス）などのニュースドメインを有しており，14,147 件のフェイクニュースと 14,857 件のリアルニュース，合計 29,004 件のニュースを格納している．train データ 17,447 件，validation データ 5,762 件を用い，学習，パラメータの調整後，test データ 5,801 件を使い，本評価結果の算出を行った.

性能評価の指標として，適合率（Precision），再現率（Recall）と F 値（F-score）を利用した．適合率は，正と判定した結果のうち、実際に真の値と一致しているかを表す指標である．次式で算出する.

$$\text{Precision} = \frac{TP}{TP + FP}$$

TP (True Positive，真陽性)は予測値を正として，その予測が正しい場合で，FP(False Positive，偽陽性)は予測値を正として、その予測が誤りの場合である.

F値とは，適合率と再現率（Recall）のトレードオフ関係に着目し，2 つの値を調和平均した値のことである．次式で算出する.

$$\text{F} - \text{Score} = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

再現率（Recall）は，実際に正であるものの中から、どれだけ正と予測できたかを表す指標で，網羅率を意味する指標である．次式で算出する.

$$\text{Recall} = \frac{TP}{TP + FN}$$

FN (False Negative，偽陰性)は予測値を負として，その予測が誤りの場合である.

4.2 調整対象の設定とmの値の決定について

利用環境（つまり、データセットの種類）により調整対象の重み付けベクトルの標準偏差値が変わる可能性があるため，データセット 1 とデータセット２のそれぞれに対し，3.3 節で述べた方法で，実験により最適化された閾値とmの値を求めた．この際，性能評価の指標としてF 値を使用した.

データセット 1(DS1)に対しては，標準偏差値の範囲を[0.10,0.35]，増分を 0.025 とし，3.3 節で述べた方法で，実験により閾値を決定する．図 2 に示すように，横軸は標準偏差値であり，縦軸は全体F値である．標準偏差値の最小値 0.10 から始め，0.025 の増分で標準偏差値の閾値を設定し実験を行った結果，閾値を 0.30 に設定したとき，全体の検出精度F値が一番高く，0.9137 となる．そのため，データセット１の閾値を 0.30 とした.

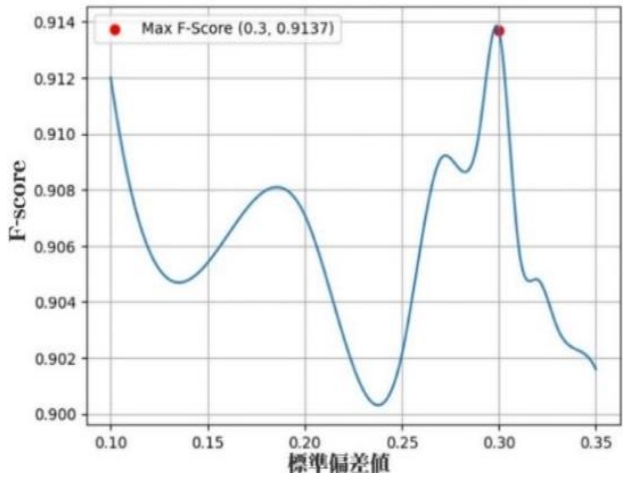

図 2　標準偏差値の閾値によるF値の変化曲線(DS1)

同様にデータセット 2(DS2)に対して実験を行い，図 3 に示す結果が得られた．閾値を 0.25 に設定したとき，全体の検出精度F値が一番高く，0.7751 となる．そのため，データセット 2 の閾値を 0.25 とした．

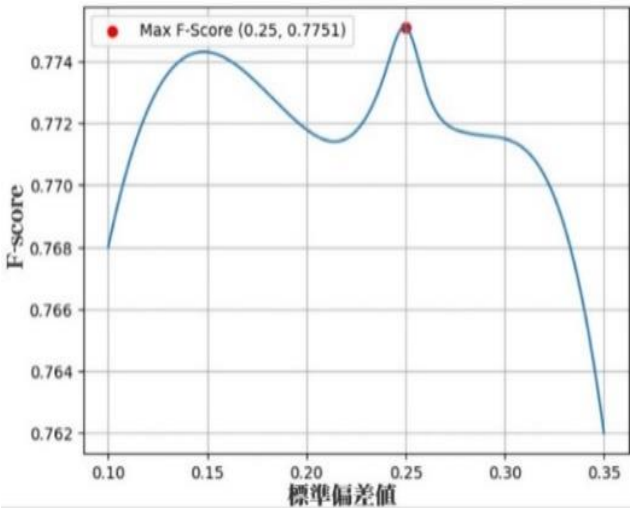

図 3　標準偏差値の閾値によるF値の変化曲線(DS2)

エキスパートの総数を 5（T＝5）に設定した実験環境では，データセット 1 に対して，閾値が 0.30 となるとき，表 1 に示されるようにmの値の変化はF値に影響を与える．$m=1$の場合は，F値が一番高く，0.9137 となる．そのため，mを 1 とした．

表 1　DS1 に対するmの値によるF値の変化

mの値	1	2	3	4	5
F値	0.9137	0.9016	0.9059	0.9130	0.8813

データセット 2 に対して，閾値が 0.25 となるとき，表 2 のようにmの値の変化がF値に影響する．$m=1$の場合は，F値が一番高く，0.7751 となる．そのため，mを 1 とした．

表 2　DS2 に対するmの値によるF値の変化

mの値	1	2	3	4	5
F値	0.7751	0.7748	0.7706	0.7684	0.5177

4.3 実験結果と考察

データセット 1 の場合，本提案手法の財経ドメイン部分はやや低い値となったが，他 8 ドメインでは改善がみられ，全体のF値としては 0.9137 となった．MDFEND のF値 0.9115 に比べ，僅かながら改善している．

データセット 2 の場合，本提案の教育ドメインだけはやや低い値となり，他 7 ドメインは改善がみられ，全体のF値は 0.7751 となった．MDFEND のF値 0.7489 に比べ，改善が見られた．

以下の実験結果により，エキスパートのドメイン知識を強化することで，効果的にフェイクニュースの検出ができることを確認した．

表 3　DS1 の MDFEND とのF値比較結果

	科学	軍事	教育	災害	政治	健康	財経	娯楽	社会	全体
MDFEND	0.81	0.91	0.88	0.90	0.89	0.96	0.90	0.90	0.90	0.9115
提案手法	0.86	0.94	0.90	0.91	0.90	0.96	0.87	0.91	0.91	0.9137

表 4　DS2 の MDFEND とのF値比較結果

	技術	セレブ	教育	政治	経済	娯楽	スポーツ	Covid19	全体
MDFEND	0.54	0.62	0.56	0.63	0.41	0.57	0.69	0.96	0.7489
提案手法	0.72	0.64	0.55	0.66	0.56	0.63	0.79	0.97	0.7751

5. おわりに

フェイクニュースの流布がますます容易になる現在，社会的な不安定化や影響の拡大を防ぐため，MDFEND や FuDFEND などの検出手法が提案され，高い検出精度の検出手法が求められている．

本研究では，混合エキスパートモデルを基にエキスパートのドメイン知識を強化する手法を提案し，Nan Qiong らの使用したデータセット(データセット１)とインターネットからダウンロードしてきた英語データセット(データセット２) を利用し評価実験を行った．評価結果により，検出結果の精度向上が認められ，効果的に複数ドメインの内容を含むフェイクニュースの検出ができることを確認した．これにより提案手法の有効性を示した．

但し，データセット２の実験結果では，ドメインによって性能が 0.55 から 0.97 までばらついており，平均で 0.7751 という結果となっている．この低い部分の分析など，検出精度をさらに向上させるために，今後も研究を続ける必要があると考える．

参考文献

[1] 総務省：令和元年版政策白書，フェイクニュースを巡る動向 (https://www.soumu.go.jp/johotsusintokei/whitepaper/ja/r01/html/nd114400.htm)（参照 2023-06-12）.

[2] Robert A. Jacobs, et al, "Adaptive Mixtures of Local Experts", Neural Computation, Vol.3, No.1(1991).

[3] Ma, Jiaqi, et al, "Modeling Task Relationships in Multi-task Learning with Multi-gate Mixture-of-Experts.", Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (2018).

[4] Lepikhin, Dmitry, et al,"Gshard: Scaling giant models with conditional computation and automatic sharding." arXiv preprint arXiv:2006.16668 (2020).

[5] Chattopadhyay, R., & Tham, C. K,"Mixture of Experts based Model Integration for Traffic State Prediction." 2022 IEEE 95th Vehicular Technology Conference (VTC2022-Spring) (2022).

[6] Fedus, William, Barret Zoph, and Noam Shazeer,"Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity." Journal of Machine Learning Research 23.120 (2022).

[7] Nan Qiong, et al,“MDFEND: Multi-Domain Fake News Detection.” CIKM ’21: Proceedings of the 30th ACM International Conference on Information & Knowledge Management(2021).

[8] Ruijianshiyao (http://www.newsverify.com) (参照 2023-06-14)

[9] Liang, Chaoqi, et al,“FuDFEND: Fuzzy-Domain for Multi-Domain Fake News Detection.” NLPCC 2022: Natural Language Processing and Chinese Computing(2022).

[10] Sina Weibo(SINA Corporation, 中国), (https://weibo.com) (参照 2023-06-01).

[11] CONSTRAINS(covid19),(https://github.com/diptamath/covid_fake_news?tab=readme-ov-file) (参照 2024-04-16).

[12] CONSTRAINS(covid19),(https://competitions.codalab.org/competitions/26655) (参照 2024-04-16).

[13] POLITICS(FNID: Fake News Inference Dataset | IEEE DataPort),(https://ieee-dataport.org/open-access/fnid-fake-news-inference-dataset) (参照 2024-04-16).

[14] TECHNOLOGY, EDUCATION, BUSINESS, SPORTS, POLITICS,ENTERTAINMENT, CELEBRITY,(https://aclanthology.org/C18-1287/) (参照 2024-04-16).

[15] TECHNOLOGY, EDUCATION, BUSINESS, SPORTS, POLITICS,ENTERTAINMENT,CELEBRITY,(https://web.eecs.umich.edu/~mihalcea/downloads.html#FakeNews) (参照 2024-04-16).