フェイクレビュー対策のための特徴量の抽出と評価

井手 伊織[1]　豊坂 祐樹[2]　成　凱[1,a)]

概要：EC サイトで商品を購入する際, 他の顧客が購入した後に書き込んだ感想や評価, いわゆる, 顧客レビューを参考にすることが重要である. しかし, レビューの中にはフェイクレビューと呼ばれる嘘の書き込みも存在し, 商品に対する正当な評価が損なわれ,EC サイトの信頼性が失われるなど問題となっている. フェイクレビュー対策の一環として, フェイクレビューを素早く検出する必要がある. レビューというものは個人が主観に基づいているものであり, レビューテキストからそのレビューが本物であるかどうかということを見分けるのが難しいことで知られている. 本研究では, フェイクレビューを検出するために有効な特徴量を特定し, 評価を行う. 具体的には, 投稿時間の集中度, 評価値の偏り, ユーザー毎の投稿件数, レビュー文の長さ等の抽出と評価を行った. 結果として, 評価値の偏り, 投稿件数が少ないユーザー毎の投稿件数, レビュー文の短さは通常のレビューより, フェイクレビューとして 10%程度傾向が高い数値が見られ, 特徴量として有効であることがわかった. また, 投稿日の集中を時系列グラフに可視化したが, 極端な特徴は見られなかった.

Feature Extraction and Evaluation for Combating Fake Reviews

1. はじめに

昨今の IT 技術向上により, 人々の生活には情報で溢れかえるようになった. 特に SNS の普及によって自分自身の意見を気軽に投稿できたり, 不特定多数の他人の意見を見ることができるようになった. そうした中, アマゾンや楽天といった EC サイトでは商品を購入した後に感想などを書くレビューを見かけるようになった. レビューは本来購入者が商品に対する感想を書き, それをサイトで閲覧したユーザーが参考にして購入を検討するものである. しかし, レビューの中にはフェイクレビューと呼ばれる嘘の書き込みも存在し, 商品に対する正当な評価が損なわれ,EC サイトの信頼性が失われるなど問題となっている.

フェイクレビュー対策の一環として, フェイクレビューを素早く検出する必要がある. しかし, あるレビューがフェイクレビュー (嘘のレビュー) であったとしても, 第三者から見てフェイクレビューであると立証することはほぼ不可能である. 理由としてレビューは個人の主観に基づいて書くものであることから, 嘘をついているかを判断するのは困難である [12][13][14].

本研究では, フェイクレビューを検出するために有効な特徴量を特定し, 評価を行う. 具体的には, 投稿時間の集中度, 評価値の偏り, ユーザー毎の投稿件数, レビュー文の長さ等の抽出と評価を行う.

[1] 九州産業大学
Kyushu Sangyo Univeristy,
2-3-1, Higashi-ku, Fukuoka, 813-8503, Japan
[2] 九州工業大学
Kyushu Institute Technology
[a)] chengk@is.kyusan-u.ac.jp

2. 偽情報としてのフェイクレビュー

インターネット社会が発展すると同時に近年の偽情報問題が深刻化している. 画像や映像による偽情報が増えている中,EC サイトなどでは購入者が感想を述べる商品レビューの文章などにも偽情報が存在する.

2.1 フェイクレビュー問題

フェイクレビューとは, 商品の評価を偽って良く (または悪く) 見せようとするレビューであり, 最近ではフェイクレビューを投稿する業者が金銭的な報酬を与える代わりとして多数のレビュアーを募って, フェイクレビューを投稿する行為が多発している [1][2][3]. これらの行為により,EC サイト上では本来の商品の評価では無く, 操作がおこなわれた評価を元に信頼しなければならない. 特に全体の評価された回数が大きく, かつ 5 段階評価中 4 以上の評価が多い

ほど信頼されやすい. その為 EC サイト上で販売する会社では評価が注視され, 上記のような業者に依頼し, 本来の商品の評価を偽って売れやすくしようとする企業も存在し, 昨今ではそういった企業がフェイクレビューを使ってマーケティングする傾向が問題となっている. フェイクレビューの定義として以下の点が挙げられる.

- 嘘のレビューが書かれていること
- 人の判断に影響を与える効果があること

好評か悪評のどちらかが書かれており, さらに複数件で意見が偏って載っていること.(好評の事例が多いため, 研究対象を好評のみに絞る) 嘘のレビューは人間の主観から見て嘘かどうかを見分けられない為, この研究では意図的に発生させたものと思わしきレビューかつ, レビュー効果の条件を満たしていることをフェイクレビューの可能性の高いものと定義とする.

2.2 フェイクレビュー対策

現状のフェイクレビューの対策として挙げられるのは, 企業側が行っている対策とユーザー側が行える対策がある. 企業側は例として Amazon が行っている「強力な機械学習のツールと熟達した調査担当者が, 毎週 1,000 万件を超す投稿を分析し, レビューの悪用を公開前に阻止する取り組み」が挙げられる [9][10]. ユーザー側が行える対策としてはインターネット上で言われているフェイクレビューの特徴の知識を身に付けている対策とサクラチェッカー[*1]などのウェブサイトに URL を貼るとフェイクレビューの確率が表示される.

フェイクレビューの特徴について様々な仮説がある. 最近のフェイクレビューの内容は上記のような単純明快な特徴はほとんど無く, 通常のレビューの内容と見分けがつかないものが増えつつあるが, よく知られている仮説をあげると以下になる.

(1) 一定期間内で募集をかけていることで, 短期間でのレビュー投稿が集中している.
(2) レビューを書くことが目的になっている為, 閲覧数, 購入数に対するレビューの比率が一般の商品よりも大幅に高くなり易い.
(3) レビューの効果を発揮させる為, 高評価率が高くなり易い.
(4) レビューの内容として商品の説明が多い. などが挙げられる.

3. レビューデータにおける特徴量抽出

フェイクレビューの検出を行う上で必要な項目として適切な特徴量であるかを分析して有効であるかを評価し, 精度の高いモデルを使用してフェイクレビューを検出しなければならない. その為, レビューとフェイクレビューの中にある違いを特徴量にし, 精度の高いフェイクレビューの検出を機械学習にさせていくことが最終的なフェイクレビュー対策の研究となるため, レビューとの違いがある特徴について検討していく.

3.1 データの前処理

データの前処理では何も手を加えていない加工前のデータを用いて機械学習が解釈できるように最終的にデータを正規化しなければならない. 本研究のデータの前処理では Python の Pandas というライブラリを使用し, 前処理を行っていく. 前処理の手順として (1)Pandas ライブラリの導入,(2) データ読み込み, (3) データ型の確認・変更,(4) データのクレンジング, (5) 特徴量の抽出, (6) スケーリング, 等が基本となる.

データの読み込みは Excel,CSV,DB などに格納されている表形式のデータ群を Python にデータフレームとして読み込むことでデータの操作を行うことができる. データ型の確認は主に日付型や文字型が数値型になっているかなどを確認する. データ型に不備があればプログラムで変更を行う. 次にデータのクレンジングとは欠損値や外れ値を整理することである. 欠損値はデータフレームの一部にデータが入っていない箇所があることがあり,Null や NaN(Not a Number) などが該当する.Null はデータが空の状態,NaN は浮動小数点における考え方で実数が異常な値であることを示している. 外れ値はデータにある異常な値である. これらの除去方法としてデータを除外するか前後のデータから平均を代入するなどして補間し, データを整理する.

3.2 特徴量の抽出と評価

特徴量の抽出は 2.2 で述べたデータの特性から有効な可能性のある特徴を見つけ出すことである. 有効でない特徴だった場合, この工程で選びなおす必要がある. 特徴量抽出までの工程が完了したらスケーリングを行う. スケーリングとは特徴量を正規化や標準化することで, 主に機械学習モデルに入力する前のデータに対して行われる. 例えばデータの列（特徴量）によっては, その範囲が 0〜1 の場合もあれば,−50〜+5000 の場合もあり得るが, このように列によって数値の範囲が違いすぎると機械学習がうまくできなかったり, 学習により多くの時間がかかったりする可能性がある. よって多くの場合では, データの前処理として正規化／標準化を行った方がよいとされる [11].

レビューデータから特徴量を抽出するにあたって, 始めに 2.2 で述べたフェイクレビューの特徴が有効であるかを実際に本研究で調査し, 有効であるかを検討していく. 使用するレビューデータには Spam という項目があり, スパムである内容なら 1, そうでなければ 0 が入っている. 本研究ではこの項目をフェイクレビューであるかを基準とする項

[*1] https://sakura-checker.jp/

表 1　Yelp レビューデータのデータ例

	User_id	*Product_id*	*Rating*	*Date*	*Review*	*Spam*	*Sentiment*	*Features*
0	923	0	3	2014/1/30	The food at snack is a selection of popular Gr...	1	Positive	['appetizer tray', 'greek salad', 'main courses']
1	924	0	3	2011/5/5	This little place in Soho is wonderful. I had ...	1	Positive	['little place', 'soho', 'lamb sandwich', 'soh...
2	925	0	4	2011/12/30	ordered lunch for 15 from Snack last Friday. A...	1	Positive	['snack', 'regular company lunch list']
3	926	0	4	2012/10/4	This is a beautiful quaint little restaurant o...	1	Positive	['beautiful quaint', 'pretty street', 'great p...
4	927	0	4	2014/2/6	Snack is great place for a A , A casual sit down...	1	Positive	['snack', 'great place', 'Ã ¢ casual', 'cold wi...

目に設定し, 研究を進めていく.

まず, 特徴量を検討していく上でデータの内容を理解し, それぞれのデータの特性を理解した上で特徴量として有効性のあるデータを分析していく.

3.3　データセット

今回，Yelp Labelled Review Dataset with Sentiments and Features というデータセットを対象に抽出実験を行った．口コミサイト Yelp が公開している 2011 年 3 月 11 日から 2014 年 8 月 31 日までのホテルとレストランのレビュー 355, 210 件をまとめたデータセットである[*2]. データ形式は Excel である. 表 1 は Yelp レビューデータの例である.Yelp レビューデータの項目について説明する.

- *User_id*: レビューしたユーザーのアカウント番号
- *Product_id*: 対象の商品番号
- *Rating*: 商品評価値.1 から 5 までの 5 段階評価
- *Date*: レビュー投稿日時
- *Review*: レビューテキスト (英語)
- *Spam*: Yelp が独自のアルゴリズムでフィルタリングした結果であり，フェイクレビューなら 1，そうでないなら 0
- *Sentiment*: レビューテキストについて Yelp の感情分析結果，肯定的なレビューなら Positive，否定的なレビューなら Negative, 無関心なら Neutral
- *Features*: レビューテキスト内の特徴的な単語

データの特性として User_id,Product_id,Spam,Sentiment は属性であり,Date は時系列データ,Rating は数値,Review,Features は文字列データである. 元のデータには他にも使用できそうな部分があり, たとえば Review の文章の長さは数値として扱える. また, 統計データを数値として扱うことでより多くの特徴量を検討することができる.

4.　実験結果

週間・月間投稿件数の推移, 評価毎の投稿件数, ユーザー毎の投稿件数,Spam の投稿件数が多い商品, レビュー文章の長さをまとめた. 特徴量としては評価の偏りや Spam が 1 である割合が高いかなどを基準に確認していく.

4.1　週間・月間投稿件数の推移

週間・月間投稿のうち,Spam が 1 である投稿と 0 である投稿の件数をそれぞれ時系列データグラフとして比較を行う. 図 1 と図 2 は週間・月間投稿件数の時系列図であり, 青が Spam が 1, 赤が Spam が 0 であり, 左右の軸の色に対応している.

比較して,Spam が 0 であるグラフと 1 であるグラフの傾向は全く違うものだとわかるが特徴的な違いがわかる部分があまりなく, 形状が似ている部分が多く, 時期的に Spam が 1 であった投稿が集中しているものでは無いと推測する.

4.2　評価点数毎の件数

評価点数（Rating）毎の件数を出力し, 比較を行う.value_counts(normalize=True) 関数を実行し, 出力された数値を表 2 に記す. なお,value_counts でデータフレーム内の Rating の項目別で件数を集計し,(normalize=True) で全体の割合として出力する. 出力された数値を表 2 に記す.

表 2　評価点数毎の Spam 率

評価点数	1	2	3	4	5
Spam=0	0.1118	0.1319	0.1138	0.3191	0.3235
Spam=1	0.1633	0.1376	0.0723	0.2404	0.3864

Spam が 0 の場合として評価が 5 と 4 が同じくらいであるのに対し,Spam が 1 の場合は 5 が大きく偏っていることがわかる. また,3,2,1 を見て 0 ものは平均的に分かれているが,Spam が 1 のものは 1 と 2 が多く, 逆に評価が 3 の場合が極端に減少していることがわかる. 全体的に Spam が 1 の場合の評価が大きく偏っている傾向が見られる.

4.3　ユーザー毎の投稿件数

ユーザー毎に投稿している件数を出力し, 比較を行う. ユーザーの投稿件数を出力するために, データフレーム内

[*2] https://www.kaggle.com/datasets/ilhamfp31/yelp-review-dataset

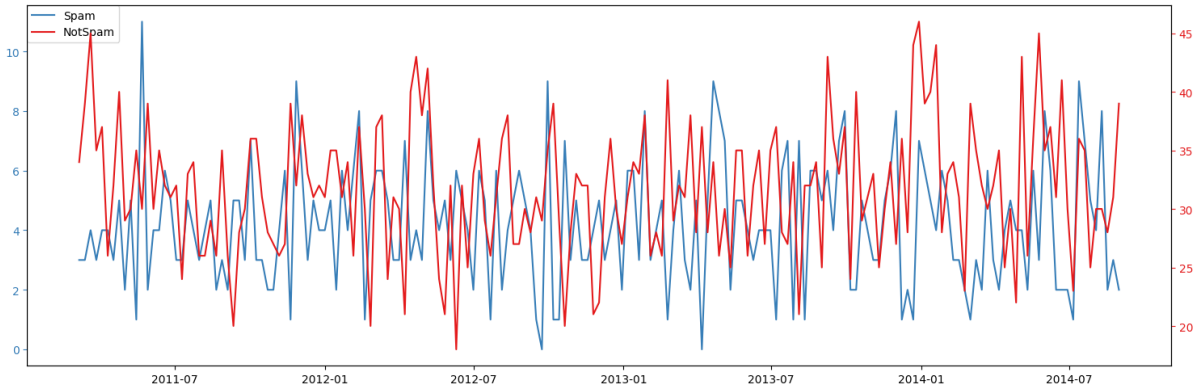

図 1 Spam と非 Spam 別のレビューの週間投稿数

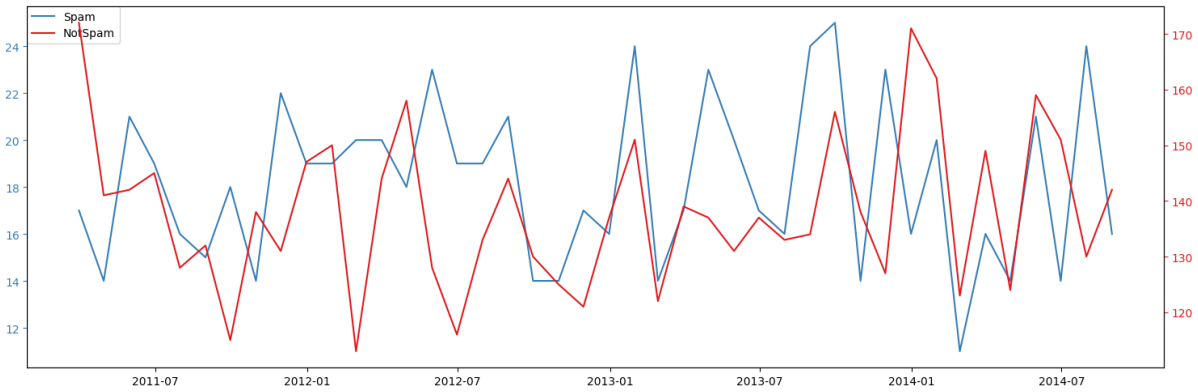

図 2 Spam と非 Spam 別のレビューの月間投稿数

にあるユーザー ID を集計し, 新たに counts の項目として出力する. 最も多く投稿しているユーザーは 181 件数であるが, 図 3 を見ると投稿件数が 1 件のものが約 66%,2 件が 15%と急激に減少した. そのため基準として (1 件,2 件,3〜10 件, それ以上) に分け,Spam の割合を表 3 に表示する.

表 3 ユーザーの投稿件数毎の Spam 率

投稿件数	1	2	3〜10	11〜
Spam=0	0.7784	0.8824	0.9509	0.9836
Spam=1	0.2216	0.1176	0.0491	0.0164

表 4 は投稿件数毎の評価の割合を示している. 高評価の傾向を見て, 投稿件数が 1 件の場合は 5 が 41%で最も高く, 投稿件数が 11 件以上の場合は 5 が 22%に減少し, 評価が 4 の場合が最も高くなった. 逆に低評価の傾向を見て, 全体的には近い数値で 10〜15%程度であるが, 投稿件数が 1 件で評価が 3 の値が 6%とかなり低い傾向が見られる.

表 4 投稿件数毎の評価の割合

評価点数	1	2	3	4	5
件数=1	0.1472	0.1363	0.0676	0.2371	0.4118
件数=2	0.1156	0.1326	0.0918	0.2904	0.3697
件数=3〜10	0.1063	0.1326	0.1169	0.3303	0.3139
件数=11〜	0.0952	0.1277	0.1608	0.3871	0.2292

Spam の投稿件数が多かった商品を 10〜49 件,50〜99 件,100〜249 件,250 件以上の 4 グループに分け,Rating の割合を調べ, 得た Spam の投稿件数が多い Product_id のみのデータフレームに戻すために 2 つのデータフレームから Product_id で結合し, 一つのデータフレームを作成する. 結合したデータフレームから以下のようなソースコードを実行し, 表 5 の数値を出力した.

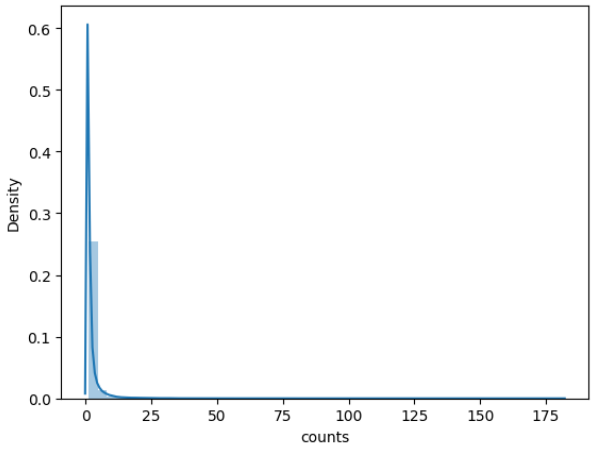

図 3 ユーザー毎の投稿件数の分布図

結果全体とし評価が 5 の割合が多かったものの, 中には 4 と同じくらいのものもあり, あまり特徴的なデータを抽出することはできなかった.

表 5　商品グループ別の評価の割合

評価点数	1	2	3	4	5
商品グループ 1	0.1107	0.1244	0.1073	0.3235	0.3341
商品グループ 2	0.1378	0.1515	0.1063	0.3092	0.2952
商品グループ 3	0.1033	0.1274	0.1197	0.3265	0.3328
商品グループ 4	0.1339	0.1536	0.1122	0.2508	0.3496

4.4　レビュー文章の長さ

レビュー文章の長さ毎に Spam の割合を算出する. 算出する際に統計データから算出した 25%(0〜169),50%(170〜385),75%(386〜746),max(747〜) を参考に４つに分類し Spam の算出を行う. レビューの長さが 0〜169 で全体の Rating の評価の割合数値と Spam の割合を出力する. 出力結果は表 6 と表 7 に示している.

表 6　レビュー文章長さ毎の評価別の割合

レビューの	1	2	3	4	5
長さ=0〜169	0.2634	0.2691	0.0495	0.1753	0.2427
長さ=170〜385	0.0932	0.1059	0.1001	0.3236	0.3772
長さ=386〜746	0.0546	0.0732	0.1320	0.3729	0.3672
長さ=746〜	0.0559	0.0809	0.1568	0.3733	0.3331

算出した結果としてレビューの長さが短いものは評価が低評価の割合が高くなり,Spam が１である割合も高くなった. 逆にレビューが長いものは評価が高くなり,Spam が１である割合が小さくなった.

表 7　レビュー文章長さ別の Spam 率

文章長さ	0〜169	170〜385	386〜746	747〜
Spam=0	0.8559	0.8804	0.9130	0.9440
Spam=1	0.1441	0.1196	0.0870	0.0560

本研究のレビューセットでは英語のレビューを使用したが, 言語によってはレビューの長さによって特徴が変化する可能性があることから英語のみ有効であると判断する.

4.5　フェイクレビュー判断に寄与しそうな特徴量

まず, 評価の偏り, ユーザー毎の投稿件数, レビューの長さはフェイクレビューとして傾向が高いと言える.

評価の偏り方としては 5 が多い傾向があり, 前述の「レビューの効果」として高い評価を付けてその商品に信頼性を持たせるためにこのような偏り方をしていると考える. 特に 1 と 5 の最大値での評価が最も閲覧者が注視する評価値であるため Spam が 1 である投稿の割合が多い傾向となった.

ユーザー毎の投稿件数に関しては投稿件数が 1 件のユーザーが傾向として多く, 考えられる理由として１人の人間が多数のアカウントを駆使して投稿していることと１つアカウントで似たようなレビューをしていることを悟られないために 1 件のみを投稿していることが考えられる.

また, 全体の約レビューの長さについては, 研究開始時点での見立てとして長いほうがフェイクレビューである可能性が高いと見立てていたが, 短いほうが Spam 率が高いという予測に反した結果となった. 理由としては今回のデータセットは 2011 年〜2014 年までの範囲であり, フェイクレビューという概念がまだ浸透していない状態の中でレビューに対するリテラシーが低かったため, 短いレビューでも問題視されていなかったと考えられる.

4.6　フェイクレビュー判断に寄与しなさそうな特徴量

逆にあまり特徴的な傾向が確認できなかったものは月毎の投稿件数と Spam の投稿が多い商品である.

月毎の投稿件数は Spam が 0 と 1 の投稿件数をグラフにし,2 軸での比較を行ったが, 傾向としては同じ形に近い部分があったが, 全く違う部分が混在していたため傾向としての評価が困難であると分析する.Spam の投稿が多い商品では全体的にグループの数値が近い値となり, 傾向があまり見られなかった. しかし, グループ 4(Spam の投稿が 250〜) の評価が 4 の数値が他の 3 つと比べて低くなったことを見て Spam が 1 である投稿件数が大きく影響していることが確認できた. このことから Spam が 1 の評価の集中した投稿には一定の影響力がある可能性が高いと言える.

4.7　考察

最終的に実際に, 仮説となる特徴と同じ内容と結果を得ることができたが, 機械学習モデルで使用していないため, 実際の評価が変わる可能性があり, 今後の課題として実際の特徴量の評価が求められる. また, レビューの長さで考察した利用者のリテラシーの向上により, 昔の傾向と現在のフェイクレビューに対して有効であるかは追加検証が必要であると考える.

今回得られた結果としてはレビューとフェイクレビューの違いをある程度分類できる特徴があり, 人間の主観以外からの判別ができたことが最終的な評価結果として言える. しかし, 確率としては 20〜25%程度での判別のみであるため, この特徴量から有効な機械学習モデルを作成し, 統計的に判別性能の高い機械学習を行うことで最終的なフェイクレビューの判別が行えるものであると考える. そのため本研究での評価結果としては 20%程度のフェイクレビューを判別することができる特徴量を抽出できたといえる.

5.　終わりに

本研究では EC サイト上で問題となっているフェイクレビュー判別を行うため, データセットから有効な特徴量抽出と評価を行った. 主に前処理の工程を行い, レビューとフェイクレビューとの違いが見られる特徴量抽出とデータの可視化を出力することができた. 特に, 今回のデータセットでは評価別の割合と Spam が 0 と 1 の割合を確認しながら比較的簡単にレビューとフェイクレビューの違いの基準

を設定することができた.

今後の予定として, 今回使用しなかったレビュー文章を使用した感情分析を行い, より複雑な特徴から有効な特徴量を明確にすることである. また, 閲覧数や購入数などの今回使用しなかった項目が入っているデータセットを入手し, 特徴量として使用して, 最終的により精度の高いフェイクレビューを判別できる機械学習モデルを作成し, フェイクレビュー対策に貢献をしていく.

参考文献

[1] Nitin Jindai,Bing Liu, 2008, *Opinion Spam and Anslysis*, WSDM'08: Proceedings of the 2008 International Conference on Web Search and Data Mining, 219-230

[2] Liu, Wenqian, et al. 2019, *A Method for the Detection of Fake Reviews Based on Temporal Features of Reviews and Comments*. IEEE Engineering Management Review 47.4 (2019): 67-79.

[3] Liu, Bing. 2020. *Sentiment analysis: Mining opinions, sentiments, and emotions*. Cambridge university press, 2020.

[4] R. Mohawesh et al, 2021, *Fake Reviews Detection*：*A Survey*. In IEEE Access.vol.9, pp.65771-65802, 2021. Doi:10.1109 /ACCESS.2021.3075573

[5] Mukherjee, A., Venkataraman, V., Liu, B., and Glance, N. S. 2013, *What yelp fake review filter might be doing?*. In Icwsm pp. 409-418

[6] Savage, David, et al. 2015, *Detection of opinion spam based on anomalous rating deviation*. Expert Systems with Applications 42.22 (2015): 8650-8657.

[7] Karishma Sharma, Feng Qian, He Jiang, Natali Ruchansky, Ming Zhang, and Yan Liu. 2019, *Combating Fake News: A Survey on Identification and Mitigation Techniques*. ACM Trans. Intell.Syst. Technol. 10, 3, Article 21 (May 2019), 42pages DOI:https://doi.org/10.1145/3305260

[8] Xinyi Zhou and Reza Zafarani. 2020, A Survey of Fake News: Fundamental Theories, Detection Methods, and Opportunities. ACM Comput. Surv. 53, 5, Article 109 (September 2021), 40 pages. DOI:https://doi.org/10.1145/3395046

[9] 米アマゾン,「フェイクレビュー」業者を提訴 「消費者欺く」, 産経ニュース (2022/2/23) https://www.sankei.com/article/20220223-4LX3CNKVBVM3TAPV3XRHRL37G4/

[10] 平和博, フェイクレビュー 2 億件,Amazon が SNS を批判するわけとは？ Yahoo！ニュース (2021/6/18) https://news.yahoo.co.jp/byline/kazuhirotaira/20210618-00243547

[11] 正規化（Normalization）／標準化（Standardization）とは？ (2021/10/7) 正規化（Normalization）／標準化（Standardization）とは？

[12] 黒木 亮人，QU LIJING，成 凱，統計的特徴量に基づくフェイクレビュー検出，第 14 回データ工学と情報マネジメントに関するフォーラム (DEIM2022) 論文集，2022 年 2 月 27 日〜3 月 2 日

[13] 黒木 亮人，成 凱，フェイクレビュー対策のためのメタ情報を利用した外れ値検出，第 20 回情報科学技術フォーラム (FIT2021) 論文集，2021 年 8 月 25 日〜27 日

[14] 黒木 亮人，成 凱，フェイクレビュー対策のためのレビュアー特定，情報処理学会第 83 回全国大会論文集，2021 年 3 月 18 日〜20 日