Fraud Review Detection: Methods, Challenges, and Analysis

Saeedreza Shehnepoor*, Roberto Togneri, Wei Liu, Mohammed Bennamoun,

***Abstract*—Social reviews have dominated the web and become a plausible source of product information. To gain more customers, businesses can hire a single user, groups of users, or a bot trained to generate fraudulent content. To provide an exhaustive survey on fraud review detection, we propose a framework that categorises the large volume of works in this area using three key components: the review itself, the user who carries out the review, and the item being reviewed. The literature is reviewed based on behavioral, text-based features and their combinations. With this framework, a comprehensive overview of approaches is presented including supervised, semi-supervised, and unsupervised learning. The supervised approaches for fraud detection are introduced and categorized into two sub-categories; classical, and deep learning. The lack of labeled datasets is identified with potential solutions suggested. To assist new researchers in understanding the field, a summary of future directions is included at each stage of the proposed systematic framework.**

***Index Terms*—raud Review, Components, Features, Classification, Deep Learningraud Review, Components, Features, Classification, Deep LearningF**

I. Introduction

User generated data in the form of social reviews is one of the defining features of Web 2.0. This has motivated businesses to not only understand, but also take advantage of the opportunity presented by social reviews to promote their products and services, resulting in their own benefit and potentially the loss of their competitors. This unfortunately also motivates some businesses to employ fraudsters to spread fraudulent information through social review platforms such as Amazon and Yelp.

A. Definitions

To facilitate understanding, we will first define the following key concepts:

- **Fraud Review: A review intentionally written by fraudsters to promote/demote services to boost gain/loss of the businesses providing the services.**
- **Fraudster: A user paid by a business to write fraud positive reviews on their own business to promote services and products or to write fraud negative reviews on their competitors' services and products.**
- **Fraudster Group: Fraudsters may form a group to collectively promote (demote) their own (competitors') services to dominate the sentiment of a service.**
- **Target Item: The service or product that a fraudulent review is written about.**

B. Fraud Review in Comparison with Other Types of Spam

Compared with other spam contents (e.g., email spams, insults, threats, malicious links, and fake news), fraud review detection is more challenging.

Insults and threat detection can rely on *sentiment analysis* to find abusive comments [1], [2], [3]. Malicious links are detectable through blacklists [4], [5], [6], [7]. Building blacklists are among the most popular approaches to block malicious links. The blacklists are provided through several websites such as jwSpamSpy[1], PhishTank[2], and DNS-BH[3]. Such blacklists are built using user feedbacks and mechanisms to detect malicious URLs.

Fake news refers to articles intentionally written to convey false information for financial and political purposes. Though knowledge of political science, journalism, and psychology is needed to detect fake news, such contents are still detectable through fact-checking techniques as factual truths are available [8], [7], [9].

Fraud reviews, on the other hand, are written or automatically generated in a similar way to genuine reviews. Hence, it is reported that even expert human judges find it hard to ascertain whether a review is fraud or not [10]. Therefore, characterizing and detecting fraud reviews is not as simple as other types of spams. Additionally, studies [11] show that fraud reviews increased in Yelp by 5% to 25% from 2005-2016, while a rating increase of 1-star in Yelp may lead to a 5-9% increase in revenue for a restaurant [12]. Hence, fraud detection is important to social review services such as Yelp and Amazon to provide spam-free platforms for users. As illustrated in Fig. 1, a social review detector can be viewed as the interaction of three types of constituents, namely, User, Review and Item. Computationally, a social review detector can be represented as a set of triplets ⟨User, Review, Item⟩, which denotes the fact that a *user* wrote a *review* about an *item*. To facilitate using a different annotation, their description is given in Table I.

TABLE I: List of symbols and their corresponding descriptions.

Symbol	Description
R_u	Review written by user u
X^i	i^{th} element of component X (e.g. R_u^i is i^{th} review of user u)
N_{X_u}	Number of X attribute on user u
NR_u	Negative review written by user u
PR_u	Positive review written by user u
$Rating_u$	Rating given by user u
$avg(X_u)$	Average of X attribute on user u
R_{u-p}	Review written from user u on item p

S. Shehnepoor (*corresponding author), R. Togneri, M. Bennamoun, and W. Liu is with the University of Western Australia, Perth, Australia. emails: {saeedreza.shehnepoor@research.uwa.edu.au, roberto.togneri@uwa.edu.au, wei.liu@uwa.edu.au, mohammed.bennamoun@uwa.edu.au.}

[1]http://www.jwspamspy.net
[2]http://www.phishtank.com/
[3]http://www.malwaredomains.com

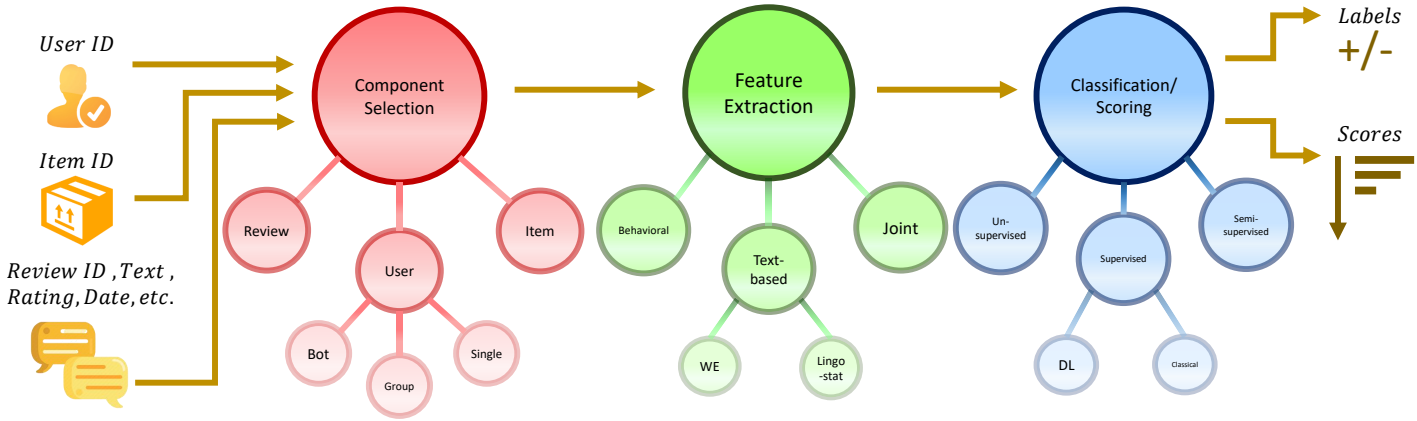

Fig. 1: A systematic framework of steps in a typical fraud detector.

C. Fraud Review Categorization and Identification - Past Literature Surveys

In the research community, recent studies focused on the identification of fraud reviews and their characteristics. Fig. 2 summarizes the main topics covered by different fraud detection review papers.

1) Fraud Definitions and Categorization: **Crawford *et al.* [13], and Dewang *et al.* [14] adopted an analytical approach to identify fraud reviews. Dewang *et al.* [14] considered fraud reviews as one of the many spam types including email spams, web spams (which refers to techniques used to manipulate the rank and the orientation of web pages), and blog spams (defined as messages posted randomly on different discussion boards), while Crawford *et al.* [13] survey different types of fraud reviews, using the definitions provided by Jindal *et al.* [15]:**

- **Untruthful Reviews: Such reviews are written to promote/defame products intentionally.**
- **Fan Reviews: The second type is reviews written by extreme/radical fans of specific brands or products.**
- **Junk Reviews: These reviews contain no relevant information regarding the targeted item; e.g., advertisements.**

Crawford *et al.* [13] then focus on the detection of untruthful reviews, similar to the current study. Mohawesh *et al.* [16] proposed a similar categorization to that of Jindal *et al.* [15] and provided two examples to explain the challenging task of labeling the reviews by human. To provide a more general category for fraud detection, Paul *et al.* [17] categorized fraud reviews as a type of opinion spam, alongside other categories such as advertisements, unrelated random texts, and product questions.

Given all definitions and different types of fraud reviews investigated by various review papers, some new types of fraud reviews are still missing. Bot generated reviews, are still missing as an important type of fraud review which are particularly difficult to spot [18]. In this paper we also consider bot generated reviews as another type of fraud and investigate relevant recent studies.

2) Review Data: **Heydari *et al.* [19] categorized review data into two major subcategories: 1) the content of reviews and 2) the metadata of reviews. The content of the reviews is used to extract text-based features, and the metadata is employed to extract behavioral features (to be discussed in Sec. I-C4). Most datasets provide both review text and metadata (e.g., rating, date of the review, user ID, and item ID) alongside the ground-truth. The statistics of major datasets is provided in Table II. Mohawesh *et al.* [16], however, proposed a different categorization where metadata is considered as a subcategory of the text-based features. The latest review by Paul *et al.* [17] surveyed models based on the features extracted by different approaches. Their features were extracted from both the review texts and metadata.**

Although previous survey papers provided a thorough review on the data that could be utilized for fraud review detection, they do not consider new modalities of data (IP, MAC address, etc.) that could be employed for fraud detection. This multi-modal data has been utilized before for untruthful reviews in Twitter [35]. We will discuss the potential of such data in providing a better overview of the user activity on a platform in Sec. V. The multimodal dataset helps the fraud detection algorithm to deal with more complicated tasks such as the cold-start problem [36]. The "cold-start problem" for fraud detection refers to the challenge of detecting fraud when there is little or no previous data available to use for training a model. This can make it difficult to accurately identify fraudulent activity.

Few recent studies [37], [38] also focused on components involved in fraud review detection, failing in delivering a systematic framework used in fraud review detection.

3) Components of Interest: **Heydari *et al.* [19], Dewang *et al.* [14], and Crawford *et al.* [13] also categorized the fraudster detection tasks into different categories. Heydari *et al.* [19] and Dewang *et al.* [14] considered two categories of single and group fraudster detection, while Crawford *et al.* [13] mostly focused on two different tasks of 1) review-based fraud detection and 2) user-based fraud detection.**

None of the current literature surveys examined the importance of targeted items (products or services) in identifying the fraud contents and preventing them to spread. In contrast to previous surveys, we consider targeted items as another component of interest and examine their potential to address the recent challenges on fraud detection, as utilized by Ji *et al.* [39] for the cold-start problem.

4) Features and Techniques: **Heydari *et al.* [19] and Crawford *et al.* [13] elaborated on different features for the detection task. Dewang *et al.* [14] specifically divided the features into linguistic and non-linguistic features. Heydari *et al.* [19] then divided fraud review detection techniques into two major groups: duplicating detection methods and content-based methods. Crawford *et al.* [13], Dewang *et al.* [14], and Vidanagama *et al.* [40] divided the classification approaches into supervised, unsupervised, and semi-supervised. In contrast to previous surveys, Vidanagama *et al.* [40] provided a critical review on features and techniques through comparison and also considered network-based approaches, overlooked by previous studies. Mohawesh *et al.* [16] surveyed the features with more focus on the modality of extracted features. Mohawesh *et al.* [16] were the first to survey word embedding techniques and explored different studies to cover the fraud detection approaches that employed such techniques. Paul *et al.* [17] specifically focused on the models based**

TABLE II: Details of datasets in previous studies.

Datasets	Reviews	Users	Items	Metadata	Ground-truth
Yelp	608,598	260,277	5,044	review text, rating (1-5), date, user ID, item ID	Yes (Yelp recommender system) [20]
Amazon	53,777	42,655	6,822	review text, rating (1-5), date, user ID, item ID, profile name, feedback (number of likes)	Yes (human judges) [15]
TripAdviser	1600	-	20	review text, rating (like, dislike), date, user ID, item ID	Yes (human judges) [10]
Tencent	302,097	82,542	7,584	review text, rating (1-5), date, user ID, item ID	Yes (near ground-truth) [21]
ResellerRating	408,470	343,063	14,561	rating (1-5), user ID, item ID	Yes (human judges) [22]
MT	2,836	-	-	review text, rating (1-5), user ID, item ID	Yes (Amazon Mechanical Turk) [23]
Dianping	21,255	504	14,187	review text, rating (1-5), date, user ID, item ID	Yes (human judges) [24]
Epinions	-	189,028	-	rating (1-5), trust, user ID	No [25]
WeiDai	36,851	-	-	review text, user ID	Yes (recommender) [26]
Lending	18,405	-	-	user ID	Yes (recommender) [27]
TREC	707,664	-	-	review text, user ID	Yes (TREC spam tracker) [28]
Expedia	1,800	-	-	review text, review rating, user ID, item ID	Yes (human judge) [29]
Google	341,993	265,724	718	date, rating (1-5), user ID, item ID	Yes (human fraud workers) [30]
Priceline	4,427	-	118	date, review text, user ID, item ID	Yes (Amazon MT) [31]
Weibo	9,726	-	-	date, review text, user ID, item ID	Yes (human judge) [32]
JD	48,562	-	100	date, rating (like, dislike), user ID, item ID, uselessVoteCount, isMobile	Yes (human judge) [33]
SWM	1,132,373	966,842	15,094	date, rating (1-5), user ID, item ID	No [34]

on the extracted features. Paul *et al.* [17], therefore, categorized Fake Review Detection (FRD) models into those which employed Content Similarity (FRD-CS), Writing Footprints (FRD-WF), Behavioral Footprints (FRD-BF), Network Footprints (FRD-NF), Rating Footprints (FRD-RF), and finally Collusion Footprints (FRD-CF).

Previous review papers considered different semi/unsupervised learning approaches to show their importance in dealing with fraud review detection when there is a lack of data. However, many recent studies utilized graph representation learning to show their potential in addressing such a challenge, such as Graph Convolutional Network [41], graph-based inductive learning [42], etc. We will discuss such techniques in Sec. IV-B and Sec. IV-C to provide a better overview of such techniques.

5) Summary: **Fraud detection is a relatively new and evolving area, some challenges are only partially addressed while new challenges have recently emerged. For example, there are also long-standing challenges that are still relevant (e.g. lack of data), and new challenges to tackle (bot-generated reviews, cold-start problem, etc.). Exploring all these topics under a unified framework helps researchers to acquire new insights of future directions in fraud review detection.**

D. Scope of This Review

Although previous studies have contributed to a better understanding of fraudulent reviews and fraudsters, there is currently no comprehensive framework that covers all aspects of fraud detection. The framework should be simple and easy to understand while helping define the scope of different studies and to clarify future challenges. Hereby, we devised a framework based on three essential steps in fraud detection tasks, as illustrated in Fig. 1.

1) Component Selection: **In the first step, a component should be selected for information collection. The framework includes three different components: review, user, and item. Review, single-user, and group users are mentioned in previous review papers. Previous review papers only considered the fraud and fraudster (single, group) detection task, but not targeted item identification. The targeted items recently attracted a lot of researchers' attention [43], [20], and were detected in combination with fraud review and fraudsters. Fig. 1 depicts the different components which can be selected by a fraud review detection algorithm. This review filled the gap by surveying works on targeted item detection. Another overlooked emerging area of study includes bot review generation. Such generators cause serious problems in social review platforms. Recent studies on fraudsters indicated that they are now using bot-generated reviews. Bots simply generate reviews similar to human written reviews, and simultaneously bots hide their footprints [18].**

2) Feature Extraction: **The second step extracts features from and for the selected component.**

Text-based features refer to features that are directly extracted from the text using language models [44] or text statistics [45], [46], [31]. To improve the review and user representations, review metadata is used to extract behavioral features [15], [47], [48]. Text-based and behavioral features are combined as "joint" features to achieve a better representation. Subsequent studies also used the same joint representation to achieve better components' representation [49], [20]. Given the fair performance of the lingo-statistic features and their combination with other types of features they can still be manipulated by fraudsters to escape detection. Recent advances in neural language models, such as Word Embeddings (WEs) can be used to produce a vector representation of review texts [50], [51], [52], [53] to overcome the limitations of text-based lingo-statistic features in achieving a global representation. In other words, lingo-statistic features suffer from a limitation to extract the sentiments from different aspects of a text, while WE is capable of extracting sentiment from various aspects of a given text. Fig. 1 provides an overview of some of the utilized features in fraud detection. We extend on the previous reviews by covering a wider range of review text representation learning.

3) Classification/Scoring: **In the third and final step, a classifier or scorer is applied to the extracted features for the final labeling (scoring). Such approaches are categorized as unsupervised, semi-supervised, or supervised. Due to the scarcity of trusted labeled data in fraud detection, recent studies employed unsupervised learning [42], [20] and semi-supervised [49] approaches. Supervised learning is mostly applied on the Yelp dataset [50], [54] for fraud detection. Fig. 1 displays an overview of fraud detection approaches. Deep learning is the**

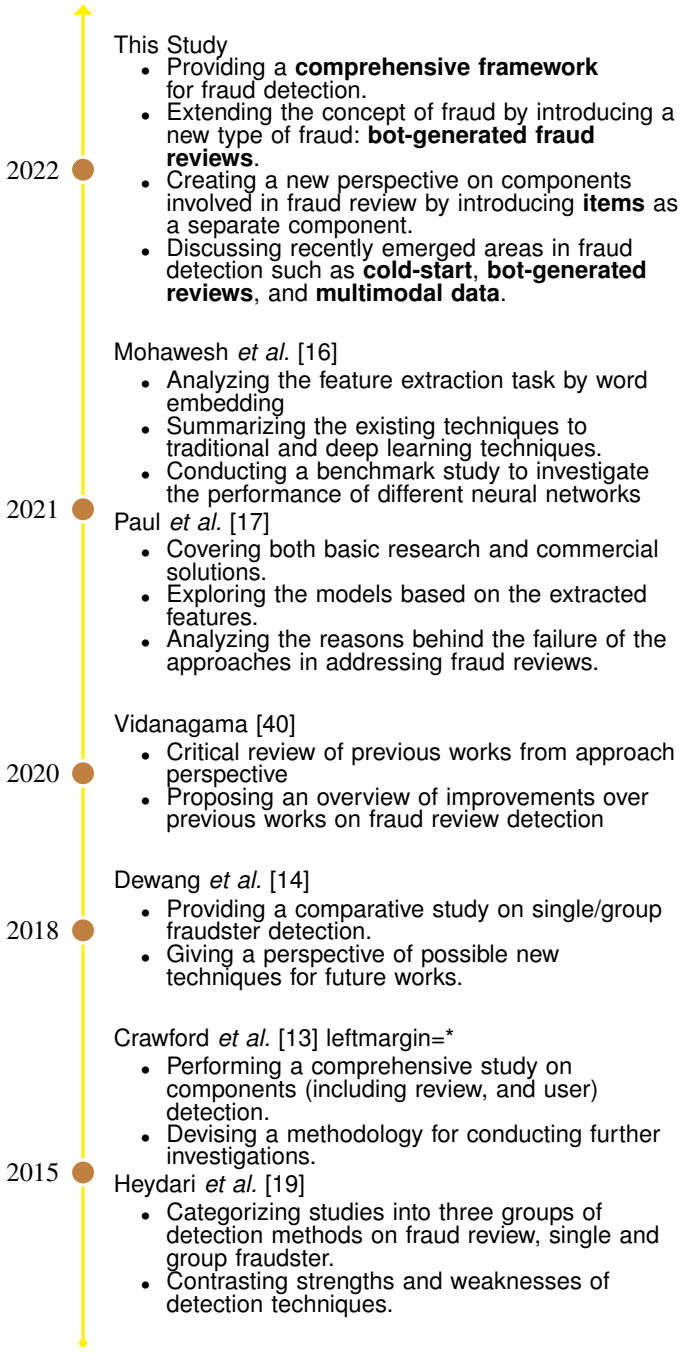

Fig. 2: The timeline of recent review papers.

most recent popular supervised learning approach for improving fraud detection accuracy [51], [55], [50].

To highlight the novelties of this survey, Fig. 1 presents each step of the framework with a focus on the overlooked aspects in previous review papers. In summary, this survey paper offers the following contributions:

- **For the first time, we introduce bot fraud generation as a newly emerging area in fraudulent reviews.**
- **We also extend the components to items targeted by fraudsters and explore the studies in the target item detection.**
- **We provide a more complete overview of open problems and identify three categories of classical, ongoing, and hot topics to facilitate better understanding of future directions of the fraud review detection research. In each category, we then explain the usefulness of such approaches to address future challenges, such as cold-start [36], [56], [57], bot-generated reviews [42], transfer learning, fraud detection through multimodal data [35], end etc.**
- **We provide the first comprehensive survey that covers various research by following a systematic framework for fraud detection in social review platforms.**

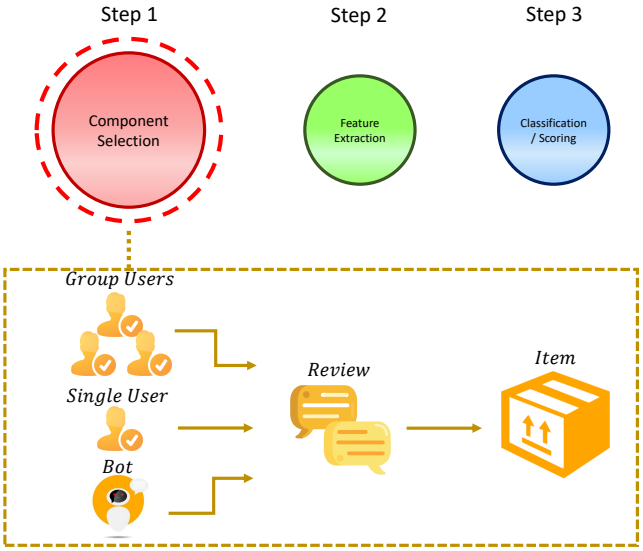

Fig. 3: An overview of component selection step in a fraud detection framework.

In the following sections, we first review the fraud detection studies focusing on each of the three steps of our proposed fraud detection framework, including components in Sec. II, features in Sec. III, and approaches in Sec. IV, respectively. Next we provide discussions on the datasets, topic analysis in fraud detection, and future directions based on the new identified challenges, in Sec. V. Finally we conclude our survey in Sec. VI.

II. First Step: Component Selection

Since the first review work by Jindal *et al.* [15] in 2008, the scope of most studies are limited to fraud review and fraudster detection. Here, we broaden the scope by introducing targeted item detection while expand the concept of fraudsters to include software bots. Bots are capable of generating contents, in a fashion that are highly similar to those written by human. Fig. 3. shows the three key components involved in fraud review detection. In the following, we discuss specific studies which considered a single component as their main focus.

A. Review

Early studies on fraud review detection mostly focused on fraud review detection as the main task. For the first time, Jindal *et al.* [15] proposed a categorization for different types of fraud reviews, namely, untruthful opinions, reviews on brands only, and non-reviews (as discussed in Sec. I-C1). Jindal *et al.* [15] employed different features such as review rating, review feedbacks, rating deviation, review ranking (behavioral), and review feedback (linguistics) to deal with fraud reviews. It is worth mentioning that Jindal *et al.* split up features into three sets of categories; reviews, reviewer, and product-centric features. The results on the Amazon dataset showed an AUC of 98.7% for fraud review detection.

TABLE III: An overview of some of the features employed by Barbado *et al.* [59]. H/L depicts if a High/Low value of the the feature is more likely to be associated with the fraud.

Features	Explanation	H/L	Formula
Review Count	Number of reviews by a specific user	H	N_{R_u}
Negative Ratio	Ratio of negative review vs. total number of reviews	H	$\frac{N_{Rating_u=1,2}}{N_{R_u}}$
Self Deviation	Deviation of ratings given by the fraudster from the rating average value	H	$\frac{1}{N_{R_u}}\sum_i(Rating_i - avg(Rating_u))$

Jindal *et al.* [15] were the first to provide a list of employed features for fraud review detection. Afterward, researchers proposed various methods to improve the performance of the fraud review detection.

Singleton fraud reviews are reviews written by the same person using different account names. Sandulescu *et al.* [58] was the first to propose a framework to spot singleton reviews. The proposed framework specifically employed two different methods: first, the review similarity; second, the topic similarity using bag-of-words and bag-of-opinion-phrases. For topic modelling, the proposed framework utilized Latent Dirichlet Allocation (LDA). The proposed approach showed a maximum precision of 80% on the Yelp dataset.

Barbado *et al.* [59] proposed a framework that extracts several types of features for fraud review detection. The proposed features include personal features (e.g., a profile description, bookmarks, and updates), social features (e.g., popularity, and compliments), review activity (e.g., review count, negative ratio, and self-deviation), and trust (e.g., relative deviation, and content similarity). An explanation of a selected few of features is provided in Table III.

Barbado *et al.* utilized an adaBoost classifier which yielded an F-measure of 82% on the Yelp dataset. With the advance in context-aware applications, Ruan *et al.* [60] proposed a Geolocation-based Account Detection Model (GADM) to improve the fraud review detection performance. GADM utilized geolocation information alongside account information (e.g., review rating) to capture location-based information in three phases. In the first phase, account features and geolocation features were extracted from the reviews and used as input. The geolocation information was obtained from visited restaurants and hotels by users. The geolocation information is then fed to a Long-Short Term Memory (LSTM) to model the user location series. A classification step is then carried out on account information and the outputs were then combined and fed to SVM for final classification. GADM showed an accuracy of 85.8% on the Yelp dataset.

He *et al.* [26] proposed a scheme considering the temporal relationship between reviews. He *et al.* [26] proposed a PU learning based on positive (P) samples (fraud reviews) and unlabelled (U) ones. The PU learning refers to an approach, where unlabeled samples are scored based on the samples with positive labels, in a binary classification task. The proposed PU scheme includes three consecutive steps designed to score reviews being fraud or genuine.

In the first step, the time interval between every two reviews was used as a feature to label unlabelled reviews as Real Negative (RN) reviews. Next, an SVM classifier is trained using the Positive (P) and RN set. Finally, text-based features were combined with outputs of previous outcomes into an Expectation-Maximization (EM) algorithm for the final clustering. Data collected from the WeiDai-Financial[4] website were used as the main dataset. The proposed approach showed an F1-measure of 75.3%.

The analysis of studies on fraud review detection shows the importance of spatial and temporal modeling of reviews in a social review platform to detect fraud reviews.

B. User

Initially, fraudsters worked in isolation when using a social review platform. As businesses recognised the significant influences of social reviews on sales and profits, fraudsters started working in groups to gain more impact. With the advances in Deep Learning techniques in recent years, bot fraudsters started to play a non-negligible role in generating and propagating fraud reviews. In this section, we take into account these new trends and focus on fraudster group and bot fraudster detection after a review of the traditional single fraudster detection.

1) Single User: **Single user refers to a fraudster who actively is involved in fraudulent activities as an individual. *Single fraudster detection* techniques mostly rely on a users' behavior pattern and their actions. Jiang *et al.* [61], [62] proposed a suspicious behavior formulation through a combination of matrix decomposition (for finding unexpected values) and KL Divergence for K-mode speciousness. A matrix is initialized first with each user as the row and features as the columns. For each user, the ID, IP, time of the written reviews and retweet comment were used to find fraud reviews on Twitter.**

Chiu *et al.* [63] also utilized user behavior analysis to find suspicious activity, using *multimodal data*. *Multimodal data* refers to data gathered from different sources such as security devices, networks, servers, and applications. Chiu *et al.* [63] employed three modalities [64] from data which refer to data derived from different layers of the review writing process (from IP address at the lowest layer to review text at the highest layer). Twitter messages obtained from streaming API, certain user data (e.g., User ID, Tweet ID, and Time), Network Information (e.g., IP Address), and Device Information (e.g., GPS position). Chiu *et al.* [63] characterized fraudsters as users with the same IP tweets who retweet with a different username for different purposes (political or commercial manipulation). The multimodal data was then mapped to a 3D tensor, representing data evolution through time. The matrix in each time step represents the user ID, review ID, and the date of the written review as data modalities. Chiu *et al.* [63] provided an intuition of how modeling the behaviors of the user through time helps to improve fraudster detection.

To detect single fraudsters, Kumar *et al.* [65] proposed to first identify possible fraudster groups to find single fraudsters. This is based on the observation that behavioral clues on singleton reviews are scarce. To obtain information on a group of reviewers' collective behavior, a review-item matrix is constructed. The output of the proposed approach is a score given to each fraudster group. They achieved a recall of 88.79% on the YelpZip dataset.

The graph Convolutional Network (GCN) was adopted by Wang *et al.* [21] to detect fraudulent users in review platforms in a framework called FdGars. Wang *et al.* [21] categorized fraudsters based on their motivations into three types: camouflaged users, crowdsourcing (group fraudsters), and fraudsters. FdGars performs both tasks of fraud and fraudster detection using review text statistics (e.g., review length and review symbol number) and user-based features (e.g., Review Quantity (RQ), Time-based Quantity Distribution (TQD), and Score-based Quantity Distribution (SQD)).

Table IV lists these features. The behavioral features connected users as nodes in a graph based on the users' proximity. The users' proximity was calculated based on the users' similarity in terms of extracted features. The GCN then labeled the reviews and users based on the similarity between them through a semi-supervised approach. FdGars achieved a performance of 93.8% for F1-measure on a dataset collected from Tencent Inc.[5]

[4]http://weidai.investorroom.com/

[5]https://www.tencent.com/en-us

TABLE IV: An overview of some of the features employed by Wang *et al.* [21]. H/L depicts if a High/Low value of the feature is more likely to be associated with fraud.

Features	Explanation	H/L	Formula
Review Length	Length (L) of review as the number of words	L	$L(R_u^i)$
Review Symbol Number	Number of non-alphabetic characters (NAC) in a review	H	N_{NAC_u}
Review Quantity	Number of reviews written by a fraudster	H	N_{R_u}
Time-based Quantity Distribution	Distribution of the number of reviews over time	-	N/A
Score-based Quantity Distribution	Distribution of the number of reviews over different ratings	-	N/A

TABLE V: An overview of some of the features employed by [69]. H/L depicts if a High/Low value of the feature is more likely to be associated with fraud.

Features	Explanation	H/L	Formula
Group Time Window	The time (T) window in which the members of a group write the reviews	L	$Max([Max(T_{R_u}) - Min(T_{R_u})]\forall u \in group)$
Group Product Tightness	Number of common products with reviewers in a group	H	$N_{R_{u-p} \bigcap R_{u'-p}}$

with 302,097 reviews written by 82,542 users on 7,548 different applications. Recently, Danilchenko *et al.* [66] employed few shot learning in combination with a classical and a graph-based approach (specifically, belief propagation) and demonstrated that the combination of the two outperformed the individual approaches. Danilchenko *et al.* [66] first created a network of users connected to each other based on a set of products they have co-reviewed. To obtain the ground truth labels for reviewers, the proposed approach employed active learning via few shot learning. The potential of each user to be benign or fraudsters was set to [0.5,0.5]. Each user was represented with a feature vector based on features such as Positive Ratio (PR), Rate Deviation (RD), etc. The experiments showed that the proposed approach outperformed the previous study with an AUC of 72% on the Yelp dataset.

In summary, the GCN provided a great opportunity for researchers to characterize the behavior of users in social review platform. Recently, multimodal data was also used to aggregate different aspects of user activity and hence improved the performance of fraudster detection.

2) User Groups: **A fraudster group refers to a group of users coordinating to attack the same item. Most studies focused on employing graphs to model a group of users. Candidates of a potential fraudster group are first generated followed by a refinement step to improve the group's constitution.**
Graphs in the fraud detection task are typically categorized into three sub-categories based on the node type:

- **Monopartite: The nodes are from the same type (users, or reviews) [67], [49], [68]**
- **Bipartite: The nodes are from two different types (both users and reviews, or users and items) [34]**
- **Multipartite: The nodes are from more than two different types (users, reviews, and items) [20].**

The connections between the nodes are defined based on the downstream task.

***NEST* and *BEST* were proposed by Zhu *et al.* [68] to address the overlooked small dense groups (modeled as a complete graph in the fraudster group detection task). A bipartite graph was utilized to model the users as a node type and users' activities as another node type. Users' activities are represented based on their social profile. Users with similar activities are more probable to be in a small group of fraudsters, representing explicit relations. Implicit relation was then defined as a relation between two users obtained through other users with similar activities. NEST and BEST both employ an iterative scorer to score the groups. BEST achieved an F-measure of 71% on the Yelp dataset.**

Bitarafan *et al.* [69] employed a Heterogeneous Information Network (HIN) to utilize the merits of a graph in both the feature and classification step. The proposed approach; $SPGD\_HIN$; employed a bi-connected candidate group detection to detect the users with at least one co-reviewed item.

The weight between two users was therefore computed based on the time-interval between the reviews written by the users. The users with strong ties (similar rating, and reviews in a time burst) were defined as a bi-connected fraudster groups. Several features were extracted such as Group Time Window (GTW) and Group Product Tightness (GPT). Table V describes these features. Next, the metapath concept was employed to obtain the final weight between each of the two groups. Groups were labeled using a semi-supervised approach based on the calculated weights. The results yield a precision of 70% on the Yelp dataset. Similar to single fraudster detection multimodal data were incorporated alongside graph-based methods to improve the detection task.
***GraphRfi*, proposed by Zhang *et al.* [41], utilized different features such as the number of rated products and the length of a username. Extracted features were combined with the rating of each review into a single vector and then fed to a Multi-Layer Perceptron (MLP) for the final prediction. *GraphRfi* showed an F-measure of 99% on the Yelp dataset.**

In summary, the earlier studies on fraudster group detection mostly relied on graph-based techniques to model the relations between members in a group, while later researches employed graphs to learn the vector representation of users. Most recent studies focused on multimodal data to better represent a group and for an improved performance.

3) Bot Fraudsters: **As deep learning has advanced, fraudsters have recently started using bots to generate fraudulent reviews. In contrast to human-written fraud reviews, bot-generated reviews can be produced in a high volume, cutting the cost of human fraudster employment. Most importantly, bot-generated reviews leave no behavioral trace, as such bots can manipulate the behavioral clues (e.g., number of written reviews in a time burst, date of written reviews, the volume of reviews, review rating, etc.), which are often used in behavioral feature extraction. In this section, we explore the studies on the newly introduced topic of bot-generated review detection. Bot-generated attacks were first investigated by Yuanshun *et al.* [18] using a pattern analysis method on the generated texts. Yuanshun *et al.* [18] proposed an approach to generate reviews based on the character distribution modeling with a simple Recurrent Neural Network (RNN). The generation process followed three consecutive steps. In the first step, the RNN was trained using characters instead of words, due to the low computational cost and the memory required for the training. Next, reviews were generated by the model. The limited size of the model led to information loss during back propagation. Hence, in the third step, the generated reviews were customized by replacing certain words with domain terminology.**

To detect generated reviews, an RNN classifier was trained based on the character distribution of the review test. To this end, two separate RNNs were trained, one with the generated reviews, and the another with the real reviews. The reviews in the test set were then fed to both models and the probability of each review being a fraud or real was calculated through a log-likelihood measurement. The likelihood was using a similar

approach to that of the Siamese network where the outputs of two different trained neural networks are given to a scorer to calculate the final score. Fig. 4 shows the framework proposed by [18]. The results of the proposed approach showed an F-measure of 86% on the Yelp dataset.

Afterwards, several studies utilized Generative Adversarial Networks (GAN) to simulate the process of the review generation through bots. GAN was proposed by Goodfellow *et al.* [70] to originally generate captions for images. The GAN consists of two main blocks: a Generator and a Discriminator. In bot review generation/detection context, before training the blocks, the generator and the discriminator are pre-trained using the samples from the training set derived from the TripAdvisor dataset. In the adversarial training phase, the pre-trained generator produces negative samples for the discriminator. In a zero-sum game between the two blocks, the generator generates fake samples and the discriminator learns to discriminate between real samples and fake ones. Both models are updated based on the backpropagation loss from the discriminator. Such a structure makes the GAN a promising choice for both bot generation and fraud detection tasks.

Aghakhani *et al.* [50] proposed FakeGAN which employed a generator and two Convolutional Neural Networks (CNNs) as discriminators, one to discriminate between the generated reviews and human written fraud reviews, and the other one to discriminate between the set of fraud reviews (the combination of generated reviews and human written fraud reviews from the dataset) and real reviews. Two discriminators were proposed to address the limitation of the GANs in handling the "mode collapse" problem. The mode collapse problem refers to a situation where the generator generates the reviews only from a single mode of distribution due to the multi-dimensionality of the data.

Before the adversarial training, the generator was pre-trained using Maximum Likelihood Estimation (MLE) and the discriminators were pre-trained using a cross-entropy loss function. FakeGAN was evaluated on TripAdvisor with 1600 reviews, containing 800 fraud reviews and 800 real reviews, and showed an accuracy of 78% (See Sec. V-B).

To better simulate bot reviews, Shehnepoor *et al.* [67] proposed ScoreGAN, a GAN-based approach where the generator utilized review ratings to improve the quality of the generated reviews. Their objective function introduced an additional term to maximize the correlation between a generated review for a given rating. Such a generator improves the quality of the generated reviews, which ultimately enhances the detection accuracy. ScoreGAN showed an accuracy of 84% and 77% on the Yelp and TripAdvisor datasets, respectively.

In summary, although only a few studies investigated bot review identification as fraud detection, recent approaches have demonstrated promising results. The use of GAN significantly improved the detection performance through bot review generation and the simultaneous training of one or two discriminator(s) to detect bot-generated reviews.

C. Item

Some services or products may be more prone to fraud review attacks than others, referred as targeted items. As soon as a new product is introduced into markets, fraudsters may start to propagate fraud reviews. Detecting targeted items, therefore, is as important as fraudster and fraud review detection. Target item detection can indirectly help improve fraudster and fraud review detection. However, few studies investigated target item detection.

Given the significance of item classification in fraud detection, FraudEagle, proposed by Akoglu *et al.* [34], uses Loopy Belief Propagation (LBP) to propagate the fraudulent score through users and items. LBP is an iterative message passing, shown to be effective for various real-world applications [71]. As existing datasets contained no groundtruth, Akoglu *et al.* [34] removed the users with a high probability of being fraudsters and then measured the items' rating difference before and after removing the suspected groups. Akoglu *et al.* [34] concluded that fraudsters significantly affected the average rating given to target items. Thus, the target items were identified by removing fraudsters while observing the impact on the SoftWare Marketplace dataset (SWM)[6].

Another work that uses item features is SPeagle [34]. To improve the fraudster detection performance, SPeagle [20] used both metadata and text to extract features from reviews, users, and items. The extracted vector (the combination of behavioral features such as negative ratio, positive ratio, and linguistic features such as content similarity) from review, user, and item was used to calculate prior knowledge for each component. The prior knowledge was then utilized to initialize a representation of each component in a multi-partite graph. Nodes in the graph represent user and items. Each edge represents a review written by a user on an item. LBP was employed to predict fraud reviews, fraudsters, and target items. SPeagle showed an AUC of 69.07% and an AP of 42.45%, on the Yelp dataset.

Ji *et al.* [39] utilized target item detection to improve fraudster group detection performance. Ji *et al.* [39] employed the item-related features (product rating distribution, product average rating distribution, and suspicious score) to detect candidate fraudster groups in the Amazon dataset[7]. To this end, the fraudster group features (group rating deviation, group size, group review tightness, group one-day reviews, group extreme rating ratio, group co-Activeness, and group co-active review ratio), individual fraudster features (ratio of extreme rating, rating deviation, the most reviews one-day, review time interval, account duration, and active time interval reviews), and item related features were first extracted. Then, a threshold was applied to the overall score, resulting in a representation to initially determine the targeted items. A Kernel Density Estimation (KDE) method was then applied to calculate the burstiness of items. KDE is a technique that could asymptotically converge to any density function with sufficient samples. As such, KDE was employed to model the review sequence of a target item to find groups of reviewers in a review burst. A review burst refers to a sudden increase in the popularity of a product due to a potential fraud attack. Next, candidate groups were determined from the KDE output. The items with burstiness value over a threshold were considered target items. The approach demonstrated an F1-measure of approximately 75% on the Amazon dataset.

D. Summary

To better illustrate the evolution of research on component detection, we have provided a timeline of previous studies on various components in Fig. 5. As new techniques have advanced, more researchers have focused on components as the primary subject of study. Hence, Fig. 5 shows a boost toward studies focusing on each of the three components (especially between 2019-2021).

III. Second Step: Feature Extraction

After choosing a component of study, features of the selected component need to be extracted, either from the review text or metadata. Features extracted from the review text are referred to as text-based features, while features making use of metadata are referred to as behavioral features. To achieve a better representation, text-based features and behavioral features can be concatenated or jointly learned through deep learning to obtain *joint* features. Fig. 6 displays the detailed steps involved in the feature extraction stage.

[6]http://odds.cs.stonybrook.edu/swmreview-dataset/

[7]https://jmcauley.ucsd.edu/data/amazon/

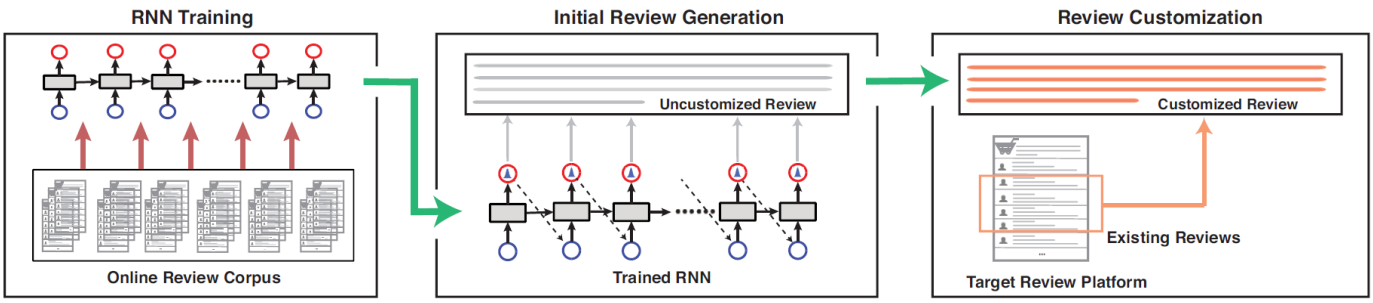

Fig. 4: The overall framework proposed by [18].

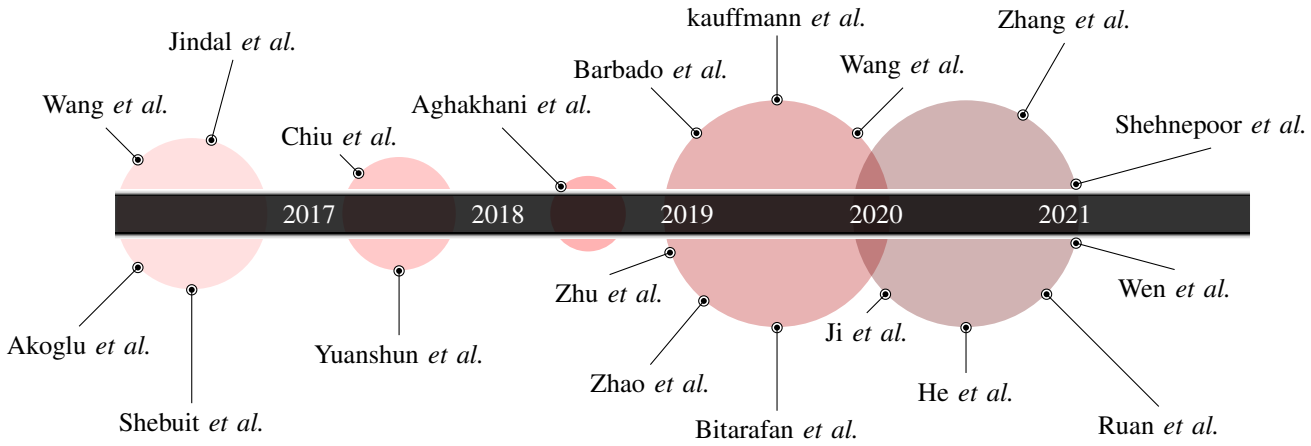

Fig. 5: The volume timeline of the studies on different components in fraud review. Diameter of each circle is proportional to the number of studies for one year period [2017-2018, 2018-2019, etc.]

A. *Text-based*

Features extracted from the text for fraud detection can be grouped into two categories, namely lingo-statistics features and vector space features (i.e., word embeddings).

1) Lingo-Statistics: **Lingo-statistics text-based features were proposed by Lai *et al.* [28] to detect different types of *Spam* contents, including fraud reviews. As one of the first text-based features, a unigram language model was developed based on Part of Speech (POS) tagging. To calculate the similarity between two different reviews, the Kullback-Leibler (KL) divergence measure was used, where the likelihood of a review generating the contents of the other review was used. The performance of the proposed approach was evaluated on the TREC dataset and showed a precision of 94%.**

Mukherjee *et al.* [72] also employed Part Of Speech (POS) tagging to extract the bigram and unigram features from review text based on a specified frequency. An SVM model was trained on the extracted POS features and showed an accuracy of 85.1% on the Yelp dataset.

Banerjee *et al.* [29] extended the concept of the lingo-statistic text-based features using *understandability*, *level of details*, *writing style*, and *cognition indicators*. For each, a set of features was proposed to describe the review. Several classifiers were employed such as Logistic Regression (LR), Decision Tree (DT), Neural Network (NN), Naive Bayes (NB), Random Forest (RF), Support Vector Machine (SVM), and voting. A dataset of 900 genuine reviews was collected from different hotel websites including *Hotels.com*, *Expedia.com*, and *Agoda.com*. To provide positive samples, 60 fake reviews were also deliberately written by experienced reviewers. LR yielded the best performance with an AUC of 81.5% on the Expedia dataset. Though singletone text-based features (i.e., features from a single review) provide a good level of characterization, the similarity between reviews (known as pair-wise features) provided a more powerful means for fraud detection and is considered as one of the most important lingo-statistic text-based features.

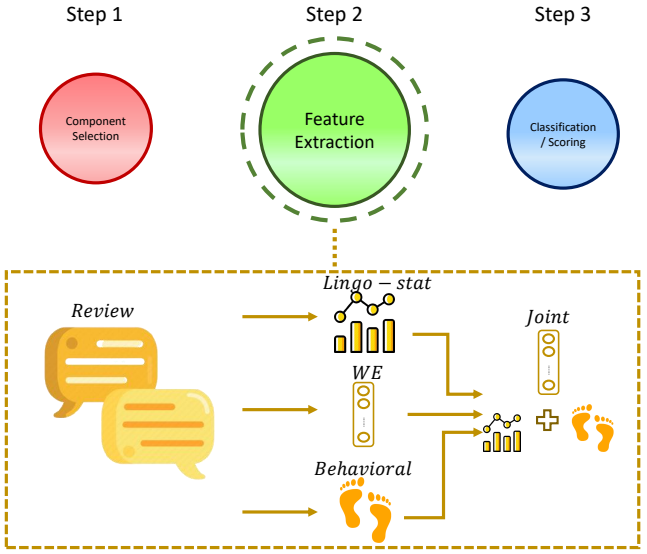

Fig. 6: An overview of the feature extraction stage in a fraud detection framework.

With many lingo-statistic text-based features proposed in the early years, Crawford *et al.* [23] decided to evaluate the importance of each word in a review for the fraud detection task. To this end, Crawford *et al.* [23] designed an approach to determine the keywords for the fraud detection task. Several feature selection techniques were utilized to determine the importance of each word in a review text for the fraud detection task. The techniques included word frequency, Chi-Squared (CS), and Mutual Information (MI). Then the presence of the 100 most frequent words in a review was given as a feature vector. The word vector was then fed to different classifiers such as LR, NB, SVM, etc. The best result was achieved for SVM on the Mechanical Turk (MT) dataset with an AUC of 87%.

As the most recent study employing lingo-statistic features, Abri *et al.* [73] extracted new linguistics features such as lexical diversity, emotiveness, etc., to detect fraud reviews. To evaluate the performance of different classifiers, the extracted features were fed to an SVM, Naive Bayesian, Random Forrest, etc. The best performance was demonstrated by the Multi-Layer Perceptron (MLP) with an accuracy of 79.09% on reviews crawled from online restaurants.

The obvious shortcoming of lingo-statistic text-based features is that indicators researchers adopted can easily be manipulated by fraudsters. However, some lingo-statistic text-based features such as language models, are employed to reduce the need for explicit indicators to deal with fraud detection. A language model can be employed in different ways as lingo-statistic features. Lai *et al.* [28] employed the unigram model to calculate the likelihood of each review is a fraud, while Shebiut *et al.* [20] employed the language model to represent each review as a bag-of-bigrams and then calculate the similarity between two reviews. With the advent of Deep Learning, different neural networks such as RNN were used to consider the language model in a review and improve the detection task.

2) Word Embeddings: Lingo-statistic text-based features suffer not only from sparsity and subjectivity but also from easy manipulation by fraudsters. On the other hand, the newly introduced Word Embedding (WE) techniques utilized vector representation of words to overcome the such limitation. Yafeng *et al.* [51] claimed that a genuine review contained more contextual embedding, i.e., it is more informative than fraud reviews, as demonstrated in previous studies [74].

Accordingly, Li *et al.* [24] extracted different text-based features such as review content similarity (pair-wise features) between two different reviews and then extracted a WE for each one. Emotion modeling was proposed based on the sentiment of the review text. The sentiment diversity, i.e., means square deviation of the negative, positive, and neutral word ratios is obtained based on their squared difference. The combination of all text-based features was then fed to different classical classifiers for labeling. The performance showed an F1-score of 93% on the Yelp dataset with a Decision Tree (DT) as the classifier, shown to be more effective than lingo-statistic features. Given the success of WE in extracting sentiment, researchers began to employ WE in the fraud detection task. Zhang *et al.* [75] claimed that the reviews written by fraudsters were different from real ones since fraudsters typically could not reflect their experience of using the product in their reviews, potentially due to the lack of experience with the products.

Ren *et al.* [51] demonstrated that the Recurrent Convolutional Neural Network (RCNN) outperformed traditional classifiers. The RCNN takes distributed word embeddings as an input to learn a continuous document representation for each review. The Continuous Bag of Words (CBOW) word embeddings were obtained using Word2Vec pre-trained on the dataset and then fed to an RCNN for extracting a document representation. Finally, the representation is fed to a softmax layer for classification. The results showed an accuracy of 83.6% on the Mechanical Turk dataset. Jia *et al.* [76] also employed term frequency and LDA (Latent Dirichlet Allocation) alongside WE. LDA is an approach that extracts review topics by highlighting important words (the most frequent ones). To calculate the term frequency, 5000 of the most frequent words (using unigram) were extracted from reviews and each word's frequency was calculated. The Word Embeddings (WE) were pre-trained for each word using the skip-gram model [77] and then an average of the words' embeddings was used as a vector representation of a review document. Jia *et al.* exploited 5 topics for fraud and genuine reviews and each topic was described in 8 words. Such a topic modelling helped to find the most popular trends among fraudsters and genuine users. A Multi-Layer Perceptron (MLP) was applied to the final representation to achieve an accuracy of 81.3% on the Yelp dataset.

In summary, the most recent studies employed word2vec techniques for fraud detection to extract WE for reviews. The proposed approaches extracted WE either independently from the classification phase in a pre-training phase (e.g., skip-gram, Glove, etc.) to obtain the representation, or employed an end-to-end deep learning technique (e.g., RCNN). To improve the performance of WE in improving fraud detection accuracy, the representations were also combined with behavioral features. As the most recent approach, new pre-trained architectures are used to provide finetuned representations, which can be easily transformed to a different scope. Bidirectional Encoder Representations from Transformers (BERT) [78] has been used to bi-directionally learn the representations using transformers. The key difference is that BERT considers both directions for learning the embeddings instead of either left or right. Such techniques can be helpful for obtaining a better representation in the fraud detection task.

B. Behavioral

Behavioral footprints play an undeniable role in detecting fraud review. Such features can be used to represent a user, a review, or an item.

With the demonstrated importance of rating as a behavioral feature in early studies, Luca *et al.* [79] employed the rating from reviews and applied a regression model to determine the fraudulent score of the reviews. As the review rating was the only feature employed by Luca *et al.* [79], the regression (R^2) value between review ratings and predictions was calculated to evaluate the performance of the proposed approach. The results demonstrated a regression of 0.68 between the review rating and the probability of the review is a fraud on the Yelp dataset.

Goswami *et al.* [80] proposed an approach with user-based features as the input to an artificial neural network (ANN) to spot fraud reviews on a dataset crawled from the Yelp website. Goswami *et al.* employed different behavioral features such as photo count, compliment, friend count, tips, followers, and funny and useful votes as extracted features. To determine the importance of each new feature a decision tree was used. The framework showed a precision of 91.53% and a recall of 99.98% on a manually collected dataset.

To incorporate effective behavioral features, an Unauthorized Optimization based De-Anonymization (UODA) approach was proposed by Hernandez *et al.* [30] to deal with fraudsters with multiple accounts on social review platforms. *Inter-review time* was also employed as a key feature in determining fraudsters (also referred to as *burstiness* in different studies). *Rating difference* between two reviews was used as another behavioral feature to reflect the correlation between two different reviews. An iterative classifier was then applied to the features extracted for each user to calculate the similarity between different users and then calculated the fraudulent score for each user. The UODA showed an F1-score of 93.84% on the Google play dataset.

Shaalan *et al.* [81] proposed a two-step approach concentrating on the key features of fraudsters such as adding redundant

TABLE VI: An overview of some of the features employed by Kumar *et al.* [82]. H/L depicts if a High/Low value of the feature is more likely to be associated with the fraud.

Features	Explanation	H/L	Formula
Review Count	Number of reviews written by a user	H	N_{R_u}
Review Gap	The time gap between written reviews of a user	L	$T_{R_u^i} - T_{R_u^{i-1}}$
Entropy	The difference between users' reviews	L	N/A

information in written review text or writing reviews in bursts. In the first step, a Deep Boltzmann Machine (DBM) was used as the aspect-level sentiment model. In the second step, a Long Short Term Memory (LSTM) was applied to the extracted sentiment aspect-level representation. The output of the LSTM was the label of a review to be fraudulent/genuine. The approach shows an accuracy of 80.85% on the Yelp dataset.

In summary, despite being utilized in fraud detection since the early studies, behavioral features are still considered as one of the most effective indicators for the fraud detection task. However, behavioral features may have limitations in handling bot-generated reviews. Different modalities of data, such as IP, MAC, etc., may provide a better overview of users' activity in social review platforms.

C. Joint

Joint features refer to the combination of text-based and behavioral features; through either a simple concatenation or a joint learning process. Joint features were proposed to improve the representation of the components in social review platforms.

Given the promising performance of the joint representation, researchers recently employed the joint learning of the text-based and the behavioral representations to address challenging problems such as the cold start problem.

The cold-start problem was first investigated by Wang *et al.* [36] and refers to the limitation of the filtering algorithm to detect fraud reviews written by a new user. Wang *et al.* [36] was presumably the first study to address the cold-start problem. This is a challenging task since there is no information history of a new user, resulting in the failure of the detection algorithm to detect new fraudsters.

Wang *et al.* [36] claimed that features proposed by early studies fail to describe user behaviors. Therefore, a *TransE* model was employed to encode a graph structure between an item, a user, and a review (as a head/translation/tail relation). Lingo-statistic text-based features (e.g., review length and maximum cosine similarity) and behavioral features (e.g., rating deviation) were extracted for components in the Yelp and Amazon platform. Continuous Bag of Words (CBoW) was employed as a Word Embedding technique to obtain the word embedding from the review text. The extracted features were then jointly learned through an objective function on both the text-based and behavioral features. Finally, a CNN was used to classify the users as genuine/fraudster on the Yelp dataset. The results demonstrated an accuracy of 58.3% on the Yelp dataset.

To address the limitations of the framework proposed by Wang *et al.* [36], an attribute-based framework, namely AEDA (Attribute Enhanced Domain Adaptive), was proposed. Three types of relations were defined: attribute-attribute, entity-attribute, and entity-entity. Pairwise features were extracted to better handle the cold-start problem and the objective function was defined based on the relations in the TransE model. AEDA demonstrated an accuracy of 80.0% on the Yelp dataset.

The joint representation learning of the text-based and behavioral features has also been broadly employed to address the data challenges in recent years. Kumar *et al.* [82] proposed an unsupervised approach to overcome the limitations of supervised learning. The proposed approach first fits the best distribution to the different behavioral features such as *Review Count (RC)*, *Review GAP (RG)*, and *rating entropy* as shown in Table VI. The length of the review (i.e., the number of words) was also extracted as a lingo-statistic text-based feature. To incorporate the inter-dependencies between reviews, extracted features were modeled as different Dirichlet distributions. The distributions were modeled using a Gaussian Mixture Model. According to Kumar *et al.*, a fraudster was more likely to be deviant in a distribution compared to a genuine user. The proposed approach demonstrated an AUC of 70% on the Yelp dataset.

As one of the most recent studies, Xiang *et al.* [83] proposed a framework to represent linguistic information using BERT. Behavioral features such as the Maximum Number of Reviews (MNR), Positive Ratio (PR), Negative Ratio (NR), etc. were extracted from the reviews and categorized into two categories of product-based and user-based features. Such features were then fed to the Graph Convolutional Network (GCN). Both BERT and behavioral features were then fused into one global feature and fed to a Softmax layer for final classification. The proposed approach demonstrated an accuracy of 69.9% on the Yelp dataset.

The joint representation extraction provided an opportunity to deal with more challenging problems recently introduced in social review platforms. However, the performance of a framework with a joint representation of components as the input highly depends on how the features are combined.

D. Summary

To provide a general overview of the evolution of features, a volume timeline is depicted in Fig. 7 showing the most recent studies on different types of features for fraud detection. Earlier studies (before 2017) focused on proposing new features, while deep learning provided the opportunity to use neural models to represent features in vector spaces. Fig. 7 shows a boost in the number of studies that focused on features in 2018, with more studies concentrating on features in the subsequent years. A summary of employed features is shown in Table VII

IV. Third Step: Classification/Scoring

In the final step of the proposed fraud detection framework, the extracted features or feature representations are used to classify each component. The approaches are categorized into three main sub-categories; supervised, semi-supervised and unsupervised. These categories can output the binary labels indicating the class of the component, or ranked scores showing the probability of the component belonging to a specific class. An overview of approaches is given in Fig. 8.

A. Supervised

The supervised approaches use labeled samples to train the model. Early studies mostly employed supervised learning to classify the components. Nonetheless, early studies [15], [87] used datasets labeled as near-ground-truth or preferred to use a small set of manually labeled samples. More recent approaches focused on unsupervised and semi-supervised approaches to overcome several limitations such as the lack of labeled data, uncertainty of near-ground-truth datasets, etc. Supervised approaches employed either classical techniques or deep learning to perform the classification.

TABLE VII: Studies based on feature categorization.

Feature Categories		Features	Paper	Component
Text-based	Lingo-stats	Unigram language model	Lai *et al.* [28]	Users
		POS-based features	Mukherjee *et al.* [72]	Reviews
		Pairwise features	Xu *et al.* [84]	Users
		puasality, lexical diversity, emotiveness	Abri *et al.* [73]	Reviews
	WE	WE through skip-gram, positive and negative ratio	Li *et al.* [24]	Reviews
		CBOW	Ren *et al.* [51]	Reviews
		Skipgram	Zhang *et al.* [75]	Users
		Skipgram + word frequency	Jia *et al.* [76]	Reviews
Behavioral		The Ratio of First Reviews, Rating Extremity, etc.	Mukherjee *et al.* [85]	Reviews
		review rating	Luca *et al.* [79]	Reviews
		photo count, compliment, friend count, tips, followers, etc.	Goswami *et al.* [80]	Reviews
		inter-review time, rating difference, etc.	Hernandez *et al.* [30]	Users
		review rating	Li *et al.* [86]	Users
		Burstiness	Shaalan *et al.* [81]	Reviews
Joint		unigrams, bigrams, second-person pronouns ratio, number of objective and subjective words in a text, and the cosine similarity between two review texts, rating deviation	Li *et al.* [46]	Reviews
		inter-review time, rating difference, etc.	Hernandez *et al.* [30]	Users
		Review Length (RL), Maximum Cosine Similarity (MCS)	Wang *et al.* [36]	Reviews
		rating difference, date difference, CBOW	You *et al.* [56]	Users
		Review Count (RC), Review GAP (RG), and Rating entropy, Review Length (RL)	Kumar *et al.* [82]	Users
		Maximum Number of Reviews, Positive Ratio, Negative Ratio, BERT	Xiang *et al.* [83]	Reviews

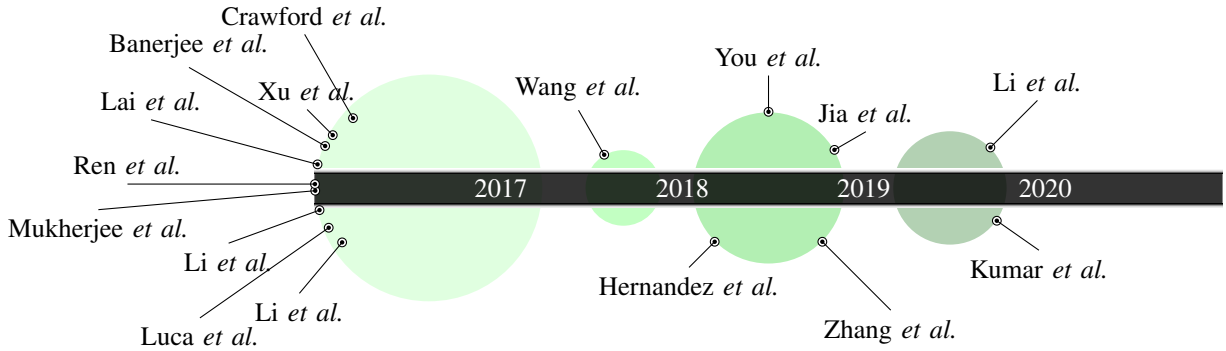

Fig. 7: The volume timeline of the studies on features employed in fraud review detection. Diameter of each circle is proportional to the number of studies for one year period [2017-2018, 2018-2019, etc.]

1) Classical: **Early studies adopted classical classifiers such as the Naive Bayesian Model, SVM (Support Vector Machine), Multi-Layer Perceptron (MLP), and Logistic Regression (LR). Such classifiers utilized samples with ground-truth labels from human experts to train the detection models.**

Peng *et al.* [88] proposed a relationship-based method to detect fraudsters. Hence, Peng *et al.* [88] considered the reviews as a relationship between the item and the user in a network. The sentiment of each review was obtained as the first step in the feature extraction step. The sentiment score of the review was then compared to the rating of the review as the first indicator, called Interior Difference (ID).

The second and third indicators were Rating Deviation (RD) and Sentiment Deviation (SD), respectively. The final fraudulent score of a user was calculated by a linear regression classifier. The experiments on the Reseller_rating dataset showed an NDCG of 90%.

Previous studies [31], [89], [88] revealed the potential of classical approaches for *multi-component* classification. Therefore, Yoo *et al.* [25] adopted the Loopy Belief Propagation (LBP) on a *multi-component graph* consisting of users, items, and reviews. The algorithm stated different *edge potential* assumptions on links between the components in the graph. Yoo *et al.* [25] then applied the assumptions in an iterative algorithm called "Supervised Belief Propagation" (SBP). SBP calculated the final probability of each review being a fraud.

The performance of the SBP showed an AUC of 93% on the Epinions dataset with 131,828 components (users, items, and

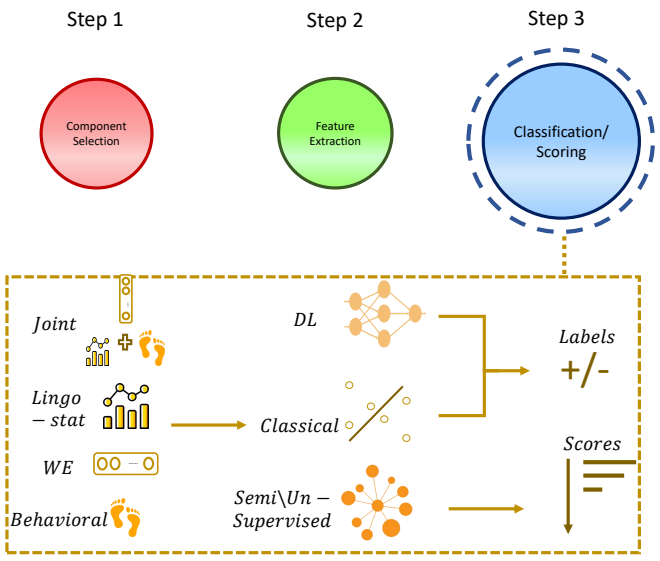

Fig. 8: An overview of different modules, feature types, and outputs used in a fraud detection framework.

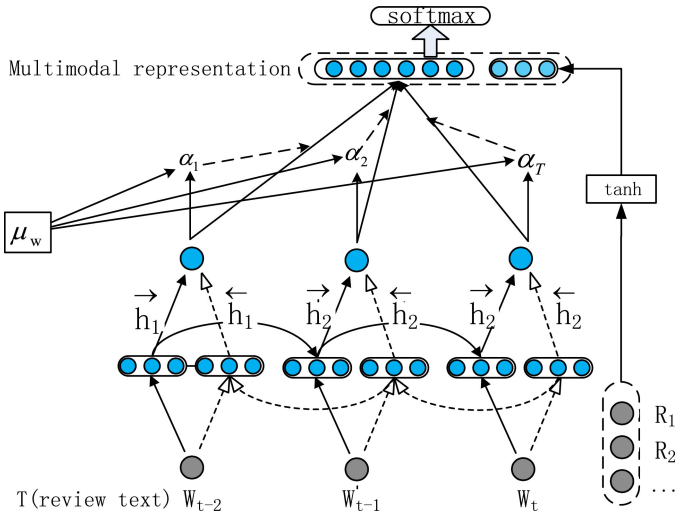

Fig. 9: Learning multi-modal embedding representation of the n-gram textual and rich behavior features [35].

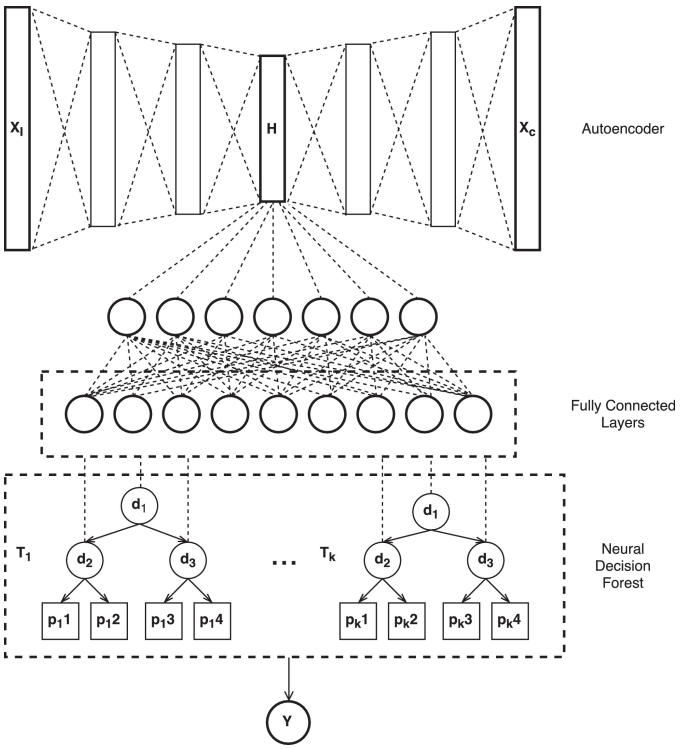

Fig. 10: Proposed neural autoencoder decision forest in [91].

reviews) and 841,372 edges.

To better utilize the potentials of the different classical supervised approaches, Kumar *et al.* [90] proposed *hierarchical supervised learning* to detect fraudsters on the Yelp dataset. User-based features such as Review Gap (RG) and Review Count (RC) were extracted. The extracted features were then used as the input to different classical classifiers such as LR, K-Nearest Neighbor (KNN), NB, AdaBoost, RF, and SVM. LR showed an AUC of 72.3% as the best classifier.

Classical supervised approaches showed a significant potential to handle fraud detection in different studies. However, such approaches suffer from a limitation in handling small datasets and require human experts to label the training dataset. Nevertheless, supervised learning has shown promising performance in fraud detection.

2) Deep Learning: **Since 2016 [51], the fraud detection research community has witnessed the success of deep learning approaches. Mostly utilized as an end-to-end architecture, deep learning approaches take texts as the input for a prediction task of either classification or regression. In other words, deep learning mostly does not have explicit feature extraction, and instead combines feature representation learning and the classification step into one end-to-end network.**

To address the domain dependent limitation of classical approaches, Liu *et al.* [35] investigated traditional features for fraud review detection on a Chinese website. Liu *et al.* [35] employed texts and metadata as multi-modal data to jointly learn a representation of each review. The architecture is given in Fig. 9. The joint representation was then fed to a bi-directional Gated Recurrent Unit (GRU) to learn the review text representation. Behavioral features such as the Ratio of Positive rating (PR), Burstiness (BST), and Rating Deviation (RD) were combined with the joint features to form a unique representation for each review. A CNN was then trained on these features to predict the class for each review. The dataset was collected from two large Chinese websites, namely Dianping hotels and restaurants. The results showed an F1-score of 68% on the collected dataset.

To utilize the potential of both classic and deep learning approaches, Dong *et al.* [91] proposed a technique to combine deep learning and traditional classification. Both user's behavioral data and text semantic features were extracted in the feature extraction stage. For the former, *the entropy of scores*, *review time entropy*, *the entropy of ratings*, and *the entropy of the product's comment time* were extracted and combined with an indicator called the *same date* indicator. For the latter, the review length and the review text semantics were extracted. An *AutoEncoder* (AE) was applied to reduce the noisy components from the initial representation. An overview of AE is given in Fig. 10. The bottleneck features were then used as the input to a *Neural Decision Tree* (DT); i.e., a decision tree where each node represents a probabilistic distribution. The approach showed an accuracy of 95.85% on the Amazon dataset.

To compare the effectiveness of different deep learning approaches, Hajek *et al.*[92] proposed an approach to compare

a Deep Feed Forward Neural Network (DFFNN) with a CNN. Review ratings and word embedding (learnt using skip-gram) were used as the features. Such features were then fed to a CNN and a DFFNN for the final classification. The CNN model achieved an accuracy of 81.30% on the Amazon dataset.

Early studies [51], [35] on deep learning mainly focused on performance improvement, recent studies [91], [92] focused on different deep learning techniques to compare and analyze the potential of such approaches in the case of challenging fraud detection tasks. The state-of-the-art study [35] demonstrates the combination of deep learning with other aspects (e.g., multimodal data) resulted in a significant improvement.

B. Semi-supervised

Supervised approaches have the limitation of relying upon a quality labelled dataset by human experts which is expensive to obtain. Semi-supervised learning provides an opportunity to utilize the existing small set of labels to propagate the available knowledge to unlabelled samples. Different from supervised learning, there is no training set to train a model, but the model progressively learns the trend in data and tunes the hyper-parameters. Given the effectiveness of deep learning in handling multimodal data, Deng *et al.* [33] also proposed a framework with features extracted from both metadata and the content of the review text. Several features were extracted from the metadata as behavioral features: *User-Level (UL)* and *User Mobility* from the IPs which were concatenated with the Bag of Word (BOW) as a text-based feature. The joint representation was then fed to an autoencoder for dimensionality reduction. K-Nearest-Neighbor (KNN) was applied to the features to achieve different feature clusters. A new dataset was collected from JD.com to evaluate the effectiveness of the proposed approach on different domains. The proposed approach showed an accuracy of 89.3% on the JD dataset. The best performance was obtained with an accuracy of 83.75% on the TripAdvisor dataset.

Recently introduced semi-supervised approaches mostly utilized graph-based techniques to deal with fraud review. Shehnepoor *et al.* [49] proposed NetSpam, combining text-based and behavioral features and obtaining the importance of each feature using a weighting method based on the *Metapath* concept. The weights were used in a Heterogeneous Information Network (HIN) to link the reviews with similar weights. A novel scoring function was then proposed to calculate the probability of a review being a fraud.

$$Pr_{u,v} = 1 - \prod_{i=1}^{L} 1 - mp_{u,v}^{p_i} \times W_{p_i} \quad (1)$$

where $Pr_{u,v}$ is the probability of user u, v being fraudster based on the metapath value between two users on feature p_i ($mp_{u,v}^{p_i}$) and W_{p_i} is the weight of the feature. The results on the Yelp dataset demonstrated an AUC of 77%.

Given the effectiveness of the graph-based techniques, Yilmaz *et al.* [93] proposed a semi-supervised approach called SPR2EP. SPR2EP incorporated two types of features; *review-based* features, and *user-item* features. The former was extracted through an algorithm called Doc2Vec, and the latter was obtained through an algorithm called Node2Vec. Both approaches relied on Word2Vec to generate the embeddings. The proposed Doc2vec converted each document to a vector embedding with a size of 384. Node2Vec mapped the review text to the same vector embedding space of size 384. The concatenated features were then fed to a Linear Regression model for the final scoring. The best performance was obtained with an AUC of 83.18% on the Yelp dataset.

As one of the latest studies, Wang *et al.* [94] proposed a vote-based integration model to harness the power of different heterogeneous information rank-based algorithms. The proposed framework, *SpamVote*, took ranked lists generated by different unsupervised detection models as inputs, computed the weighting of ranked reviews based on the similarity between two ranked lists, and output the score of a review is a fraud review. *SpamVote* showed NDCG@100 of 93.68% on the Yelp dataset.

C. Unsupervised

In addition to their promising performance in semi-supervised models, graph-based approaches were also employed for unsupervised learning.

Liu *et al.* proposed HoloScope [95] that models data as a bipartite graph with users as the source nodes and items as the sink nodes.

Nodes were linked through a directed edge signed by either rating or the timestamp of the review. Holoscope then applied a mutually exclusive iterative algorithm inspired by [100], [42] to score the users as fraudsters or genuine. The results on the Yelp dataset demonstrated an AUC of 99.5%.

Hooi *et al.* [42] also proposed a novel unsupervised approach to deal with a recent hot-topic challenge in fraud detection: the *camouflaged fraudsters*. *Camouflage* refers to an act of writing genuine reviews by fraudsters to escape detection. Hooi *et al.* mapped the components to a bipartite network with items and reviews as two different node types in the graph. An iterative algorithm, inspired by [43], was applied to a bipartite network to find the score of a new user potentially being a fraudster. Results on Amazon showed an accuracy of 89%. Recently, unsupervised approaches are also employed to address new challenging topics such as the cold-start problem. Kumar *et al.* [43] proposed REV2 to incorporate a Bayesian Belief Network to perform inference on three components (fairness of the user, the goodness of the product, and the reliability of a reviewer) from five different datasets including Flipkart, Bitcoin OTC, Bitcoin Alpha, Epinions, and Amazon. To handle the cold-start problem, Kumar *et al.* utilized a Laplacian smoothing. To score a user as a fraudster or honest, a mutually recursive algorithm was applied to the components. Results demonstrated an accuracy of 64.89% for the multi-component classification on the Amazon dataset.

As unsupervised approaches are not limited by the labeling quality of the reviews, recent research mostly relied on unsupervised methods to address new challenging topics (such as cold start and camouflage). The recent approaches mainly employed graph-based techniques to model the components. The potential of unsupervised approaches is not only to improve the detection accuracy in fraud detection but also to address the current and future challenges.

D. Summary

The volume timeline of previous studies on different approaches is provided in Fig. 11. Similar to Fig. 7, there is a boost in the number of studies focusing on models as the main contribution in 2018, coupled with the introduction of deep learning for fraud detection. Thereafter, most studies concentrated on deep learning for the feature representation learning step. A summary of approaches is given in Fig. VIII.

V. Analysis

This section includes a detailed analysis of the findings in various subsections. We provide an analysis of three types of topics: *classical*, *ongoing*, and *future* challenge. In the first section, we provide an analysis of classical topics already investigated in previous studies. After that, we provide an overview of possible future directions. In the subsequent section, we examine data as the primary challenge, in previous and future studies. We also suggest future steps to address the shortage of data. Finally, we discuss different future topics and possible solutions.

TABLE VIII: Studies based on approaches.

Approach	Type	Model	Output	Paper
Supervised	Classical	Support Vector Machine	Binary classes	Myle *et al.* [31]
		Bayesian Network	Ranked scores	Li *et al.* [89]
		Linear Regression	Binary classes	Peng *et al.* [88]
		Linear Regression	Binary classes	Kumar *et al.* [90]
	Deep Learning	Convolutional Neural Network	Binary classes	Wang *et al.*[96]
		Convolutional Neural Network	Binary classes	Liu *et al.*[35]
		Neural Decision Tree	Binary classes	Dong *et al.* [91]
		Convolutional Neural Network	Binary classes	Hajek *et al.* [92]
Semi supervised	Tabular	Linear Regression	Binary classes	Wang *et al.* [32]
		K-nearest Neighbor	Clusters	Deng *et al.* [33]
		Linear Regression	Binary classes	Rout *et al.* [97]
		Linear Regression	Binary classes	Yilmaz *et al.* [93]
		Voting	Binary classes	Wang *et al.* [94]
	Graph	Heterogeneous Information Network	Ranked scores	Shehnepoor *et al.* [49]
Unsupervised	Tabular	Expectation Maximization	Ranked probabilities	Xu *et al.* [74]
		Maximum Likelihood Estimation	Binary classes	Dong *et al.* [98]
		Ranking	Ranked scores	Xu *et al.* [99]
	Graph	Iterative Mutually Exclusive	Ranked scores	Hooi *et al.* [42]
		Bayesian network	Ranked scores	Kumar *et al.* [43]

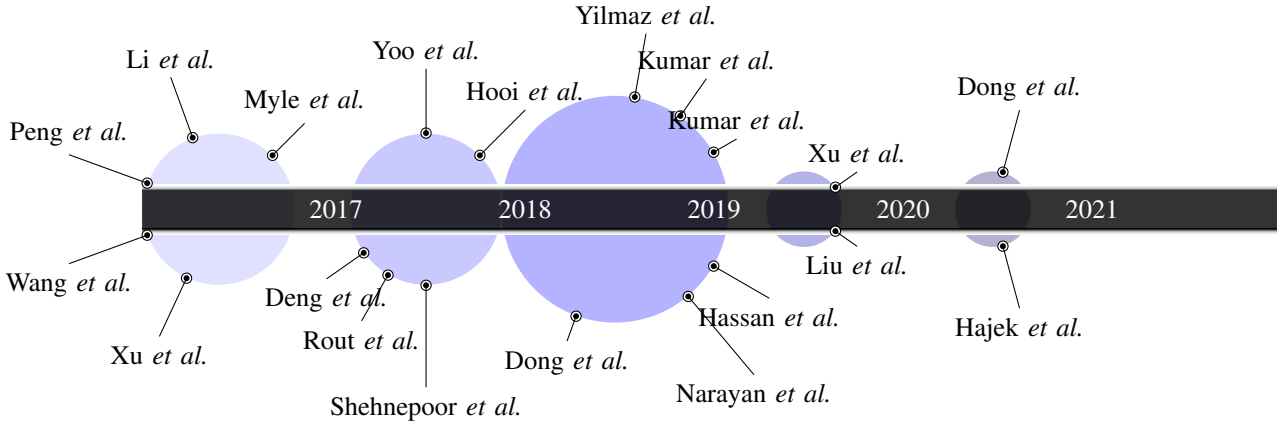

Fig. 11: The volume timeline of the most recent studies using different approaches for fraud detection. Diameter of each circle is proportional to the number of studies for one year period [2017-2018, 2018-2019, etc.]

A. Discussion on Topics: Classical Challenging Topics

This survey studied 58 recent works with 80% after 2016 (46 out of 58) to properly present the most recent techniques and challenges in fraud detection. Fig. 12 shows the research study statistics for different topics: components, features, and approaches. As Fig. 12 suggests, user and review detection were the main focus of researchers in recent years (94% combined), while item (as a separate component) takes up about 6% of the research efforts.

1) Targeted Item Detection: **Target item detection benefits fraud detection in different aspects. For example, it facilitates fraudster group detection, as suggested by Ji *et al.* [39]. It has also shown to be effective in dealing with the cold-start problem [36], [56] by finding items with a burst of reviews written in a short period of time. This encourages researchers to prioritize target item detection in order to address potential future challenges.**

2) Feature Extraction Importance: **The uniform distribution over the studies on fraud features (Fig. 12) suggests an equal importance of each category for researchers. It is worth mentioning that early studies employed text-based features to address fraud detection. In recent years, and with the significant advances in deep neural networks, deep learning techniques were utilized to represent text-based features implicitly using word**

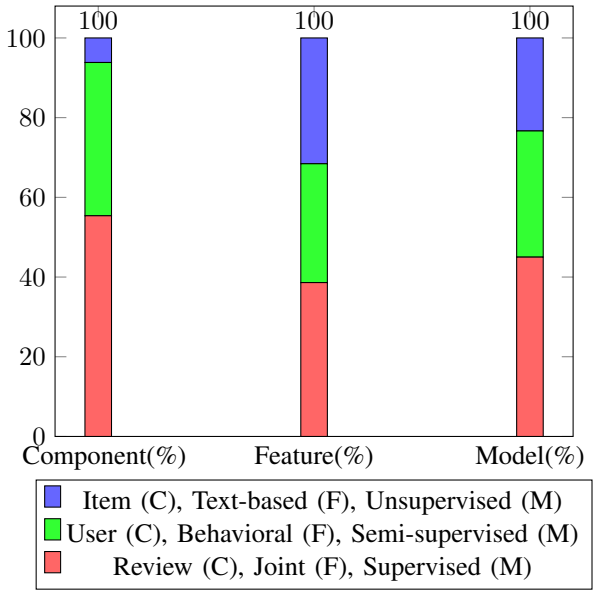

Fig. 12: Percentage breakdown of studies on components, features and classification approaches.

embeddings. Deep learning was also utilized as a classifier in the classification step or in the joint learning task in end-to-end neural networks. End-to-end approaches provide an opportunity to address fraud detection, by combining the feature representation and classification step. Despite behavioral features' effectiveness in handling fraud review, in recent years, and with the growth of bot-generated reviews, the behavioral features' representation may result in an unreliable representation of each component. However, given that new datasets contain information of the lower layers in the OSI model (such as IP, MAC address, and other basic information of users), these will likely compensate for the shortcomings in the application layer metadata and review text. Joint feature learning is considered to be an active area of research, since advances in both text-based and behavioral features result in improved joint representations.

3) Graph-based Approach Importance: According to Fig. 12, almost half of the approaches adopted supervised learning, one-third of the approaches utilized semi-supervised learning and the rest used unsupervised learning.

Fig. 12 suggests that there is a tendency towards using un/semi-supervised approaches. More than half of the studies relied on unlabelled data (unsupervised and semi-supervised approaches combined), while the majority of the remaining approaches, with supervised learning as their main focus, relied on the Yelp dataset.

With the lack of data and explanations provided in previous sections, a set of suitable approaches with a capability to correctly populate the labels for the unlabelled samples is required. Graph-based approaches, however, empower the models to utilize the merits of the connection between components to better generalize the classifications. Conclusively, the findings for the topic analysis demonstrate that graph-based approaches, as the latest techniques to address the lack of labels, can be utilized to employ a message forwarding [34] or inductive learning [42] to populate the labels based on different criteria. However, such propagation approaches (message forwarding, etc.) are well explored (Fig. 12) and researchers are using different refinement methods to deal with the representation tuning.

B. Lack of Data Problem: Ongoing Challenging Topics

The lack of data is the most challenging topic for researchers in fraud detection. The challenges are across a range of different tasks across all stages in our systematic fraud detection framework, such as the data collection for a component, data preparation for feature extraction, and most importantly the labels used to train or evaluate the framework performance for the classification/scoring step.

1) Labels: The most pressing challenge is the lack of trustworthy ground truth for fraud reviews, due to the challenging nature of fraud labeling. Due to the complexity of the task, even human labeling accuracy is no better than a random classifier [20], [18]. Some platforms (e.g., TripAdvisor) employed review experts as human judges to label the reviews [50]. Furthermore, the difficulty of labeling depends on the task. For example, a group labeling task is less challenging than labeling individual users, since the collaboration among members of a group provides a reasonable context for comparison and judgement [39]. Similarly, some platforms use detection algorithms to label the datasets, e.g., Yelp [69], [41], [93], [96], [36], [56], [86], [79]. Datasets labeled by the detectors, are called "near ground-truth" datasets, and the labels are mostly used in semi-supervised or unsupervised approaches [20], [49]. Apart from review labeling challenges, the other challenge is the lack of ground truth for the other components, such as users (fraudster or genuine user) and items (targeted by fraudsters or non-targeted). Some studies label the users with at least one fraud review as fraudsters [20]. Such labeling would result in misclassification, given that the labels are provided as near-ground truth.

There is a growing trend towards using datasets populated by web crawling applications, to deal with the lack of data problems. The Tencent dataset has been extensively used in recent studies [21], [101], and provided labels as near-ground truth. Fig. 13 shows that a considerable number of studies use the Yelp dataset. The first possible explanation is that the Yelp platform is a public dataset and is accessible for different purposes. The Yelp dataset also provides the ground truth for reviews based on the platform recommender. Similarly, the Amazon dataset provides a variety of datasets for different businesses with near-ground truth labels. TripAdvisor relies on a balanced distribution of fraud and genuine reviews for hotels in Chicago, providing 800 reviews for each class of fraud and genuine reviews. Such a dataset is either used in studies with classic classifiers [97], [102], [103] or for data augmentation to overcome data paucity [50], [67]. The remaining datasets are crawled using web crawlers, mainly to provide an insight for different fraud detection topics.

Labeling the review texts requires different levels of knowledge. Similar studies in related areas might be interesting and inspiring in fraud detection. Studies on detecting fake news, as another type of spam, employed fact-checking to improve the accuracy of the labeling [104], [105], [106]. Hierarchical fact checking is employed to find the clues about the origin of the news. However, accessing that level of information requires specific permission to the data. Nonetheless, fake news detection accuracy somehow depends on different datasets presented to the algorithm. In fraud review detection, on the other hand, acquiring such a relationship between different reviews in multiple platforms is challenging. Fake news detection, however, can provide some useful insights into how hierarchical fact checking [107] can be helpful to provide a degree of confidence on the labeling. Such an exploration requires considerable efforts on the labeller end.

As another direction, incorporating expertise from other areas such as psychology could be useful in identifying key implications of written reviews and the intention of the user through the use of psychological indicators. Such indicators can be driven from the sentiment analysis of the review text. The indicators are then correlated with the analysis from a psychological point of view [108], [109]. Porshnev *et al.* [108] studied the emotional state of different users in Twitter using sentiment analysis. Word frequency is also used as a feature to analyze the psychological state of the people who experienced a sudden change in the isolation with the COVID-19 outbreak. The outcome of such psychological indications can be used to help the labeller.

Given all the explanations and directions, achieving a large amount of data requires considerable investment from companies. Marciano *et al.* [110] reported that fraud reviews in online shops cost $152 billion a year. Facebook introduced new machine learning techniques to address scam on Facebook in 2018 [111]. Although such contents are considered as opinion spam, finding the accurate label for the task requires significant investment to provide both budget and data to empower the labeling tools for finding the correct labels.

2) Multi-modal Data: Fortunately, several platforms (e.g., Yelp, Amazon, and TripAdvisor) provide datasets including the review text and the metadata (e.g., user ID, item ID, review ID, date of a written review, rating given by the review, and the label) for each review. However, the platforms also need to withhold important information to secure the users' privacy.

3) More specifically, Yelp, which has been the most studied dataset (as shown in Fig. 13) includes the necessary metadata such as user ID, item ID, date of the written review, and the rating given by the review that can be used to extract behavioral features for user activity at the application layer of the OSI model [112]. However, experienced fraudsters employ a camouflage strategy to simply cover up their behavioral footprints and manage their traceable metadata. Although different approaches were proposed to address the camouflage problem, the performance of the state-of-the-art approaches is limited due to the lack of information of the user behavior at the lower layer of the OSI

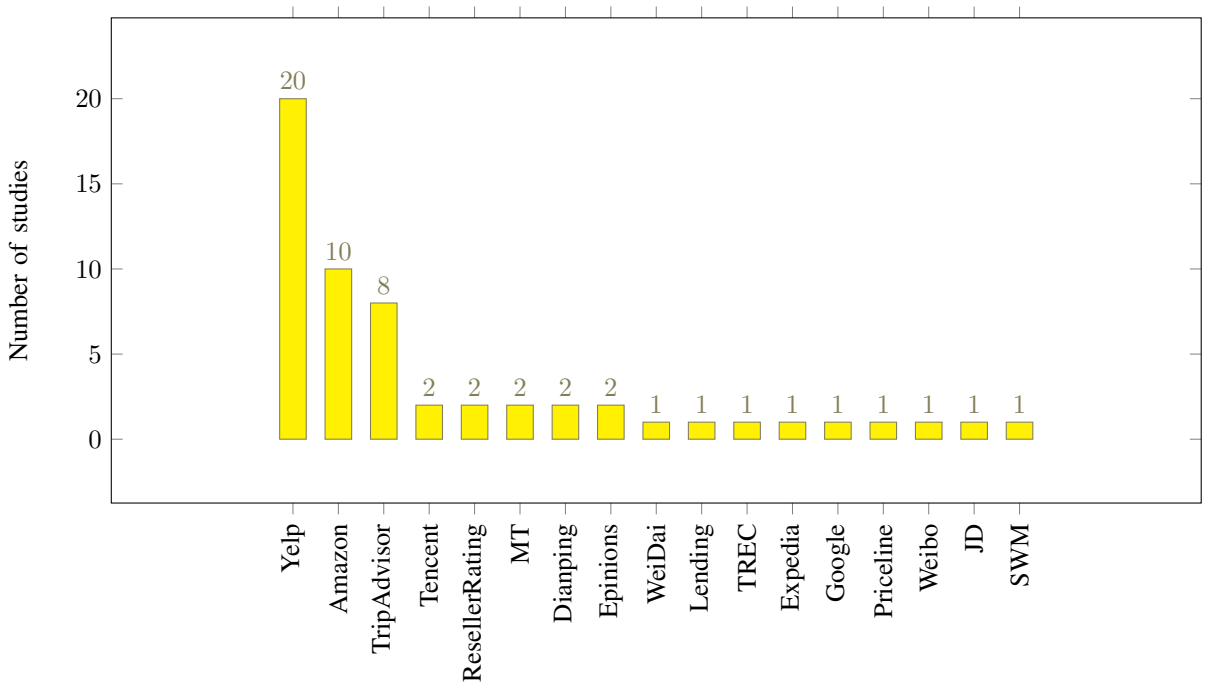

Fig. 13: Dataset distribution over different studies. Note that the most recent study on each dataset is provided alongside the dataset.

model (e.g., network layer). To deal with the camouflage problem more information is required to help the detection algorithm to fully trace the user's behavior. Unfortunately, the well-known platforms are only providing particular confidential information such as the user IPs in each session, location coordinates, etc. It is worth mentioning that some platforms provide privileges (e.g. APIs) for the crawlers to collect data from different data modalities, also referred to as *Multi-modal* data [63]. Gathering information from different modalities for feature extraction is an unexplored and a potential topic to consider for future studies [63]. Similarly, various studies employed multi-modal data to address spam tweets [61], [62] on different domains of the social review platforms. Using different data models, e.g., the Denial of Service (DoS) containing different attributes, is an effective approach in detecting fraud reviews. Liu *et al.* [35] suggest that feature types can be regarded as different data modalities. The features are likely extracted from one dataset with the metadata as the attribute to then extract text-based and behavioral features as different modalities. Attributes from the network layer (e.g., IP of the user) were utilized in a few studies [21], [101] to initialize a new area of research in fraud detection. Hence, future studies can extract multi-modal data and combine such features to achieve a better behavioral representation of each user.

For example, Wen *et al.* [101] and *Wang et al.* [21] collected data from the Tencent platform which includes IP and different physical layers alongside information from the application layer. Technically speaking, such multimodal data enables the detection model to learn a joint representation and then extract a unique vector for each user. The vector not only represents the traceable user's behavior at the application layer but also the background activities of the users. Such additional information help the detector with new fraudster detection in a shorter time compared to a situation in where the detector is provided with the information from only one data modality, and this can also address the cold-start problem [36].

However, the inconsistency between the different modalities of data could also be an interesting topic to explore. For example, different studies considered rating as a sentiment indicator, representing the tendency of the user towards the item. Surprisingly, with the sentiment analysis of a text, such assumptions are more likely to be inaccurate [36], [34], due to the inconsistency between the rating and the real sentiment of the review. Hence, to obtain accurate semantics, studies should solely rely on the sentiment analysis of the review text (and not rating). To train a joint representation of the text and metadata, several objective functions are proposed in different studies [36], [67].

3) Transfer Learning: **One of the most recent topics is to use transfer learning to improve fraud detection accuracy. Gupta *et al.* [113] employed BERT and transferred an NLP task to fraud review detection with different configurations. However, the study employed transfer learning as part of a representation extraction step from the text, though transfer learning can be applied to different steps of a machine learning task. To be specific, in a relatively similar area, such as speech recognition, studies applied transfer learning to improve the accuracy of the children's speech recognition using a model already trained with adult data [114], [115]. The transfer learning approach has shown to be useful in the case of data scarcity, as with a model trained with plentiful data on the other task, replacing the classification layer and initial training of the layer which requires a relatively smaller set of data [116].**

C. Hot Topics: Future Challenging Topics

There are some recent challenging topics addressed only by a few studies. We discuss such topics in the following and provide insights on possible solutions.

1) Bot Generated Review Detection: **Bot review generation was introduced by Yuanshun *et al.* [18] and then explored by different studies [50], [67]. Yuanshun *et al.* [18] claimed that bot-generated reviews are indistinguishable from human-written reviews from different aspects. Such reviews are manipulated by fraudsters, and generated in large volume, with minimum cost. A generator can be trained to generate authentic reviews, leaving neither human traces nor footprints. Consequently, behavioral**

features become less useful for bot-generated review detection. Yao *et al.* [117] reported an important observation on a bot trained by a character-based language model over the whole corpus. The study discussed one way to detect such fraud reviews is to parameterize the character distribution, as the distribution is one key factor in determining a bot review. Relying on a specific distribution, the study detected the bot-generated review using a likelihood measurement. This can be applied to different platforms (Yelp, TripAdvisor, etc.) using different criteria, as each domain has its own specific terminology.

Shehnepoor *et al.* [67] proposed ScoreGAN with synthetic reviews generated by a Generative Adversarial Network, and the performance of a model trained on the Yelp dataset is improved. Shehnepoor *et al.* thus suggested that there is a high probability of bot review presence in the Yelp dataset. A closer look at the generated review set reveals that generated reviews are often short. This is because generating long semantically coherent reviews requires a large training dataset. The syntax is not an issue, as there are only a relatively small set of syntax rules. A possible future direction is to distinguish between human-written fraud reviews and bot-generated fraud reviews. This will provide useful hints to characterize both human and bot-generated reviews. Few studies investigated such a problem and the studies mostly suffered from uncertainties in determining whether the final results can be validated unless the content is synthetically generated. Shehnepoor *et al.* also discussed the importance of using the rating in combination with the review text, as the only behavioral indicator of bots.

In summary, these findings suggest that bots can be identified with three key factors: first, certain distribution over characters, or words; whether synthetically correct, but semantically not; and finally, coherence between text and rating.

2) Cold-start Problem: The cold start problem can be expanded to a multi-domain (hotels, restaurants, online shops, etc.) problem, where different domains are involved in tracking the behavior of a user. A fraudster typically does not only write reviews on one platform, but multiple ones to maximize the impression. Similar studies are conducted in recent years to find the path of influential users in different social networks, namely multi-layer networks [118], [119]. As previously mentioned, fake news detection studies often employ a similar approach to verify the accuracy of a news article. In a similar study, Sivasankari *et al.* [120] used a multi-domain series of data to trace back the truth about suspicious news. Such cross-domain behavior analysis with a focus on similar patterns can be useful [56]. Although You *et al.* [56] employed a similar approach towards the cold-start problem, the study lacks the behavioral pattern exploration and analysis and solely relies on extracting attributes from multiple domains.

Graph-based approaches [22] are potentially suitable for the classification/scoring step to train the joint representation and address different challenges (e.g., cold-start problem). Graph-based learning considers the relations between the extracted features to refine the prior knowledge using a network-based inference algorithm. The inference algorithms mainly embed the network-based representations of the components into a single representation. Some previous graph-based approaches (e.g., Loopy Belief Propagation [34], [20]) suffered from limitations in normalizing the nodes' representation. The limitation leads to incorrect component predictions, since the feature aggregation is simply altered once the node's degree is changed. This allows fraudsters to manipulate their behaviors through camouflage. One possible solution is to employ graph-based inductive learning [121]. Inductive learning facilitates the representation learning for single or multiple components in a graph. The representations can then be used to find the primary representations of the newly introduced nodes, such as users or items in a social platform to handle the cold-start problem.

In summary, a potentially more optimized solution is to combine 1) user path prediction over a multi-layer network and 2) using the cross-domain features (as suggested by [56]) to find similar co-occurring behavioral patterns, or 3) graph-based approach to propagate the representation to unseen data.

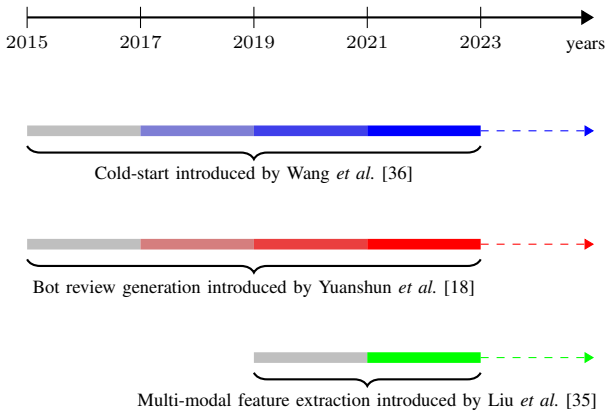

Fig. 14: Emerging areas timeline and potential future work.

We summarized the future works in a timeline depicted in Fig. 14, highlighting the three current hot topics for fraud detection. We picked an open problem in each of the three stages in our proposed systematic framework (component selection, feature extraction, classification/scoring) and elaborated on the current study. Bot-generated review detection was introduced as an open problem in the component selection step by Yuanshun *et al.* [18] and was primarily investigated by two studies by Aghakhani *et al.* [50] and Shehnepoor *et al.* [67]. As explained in the feature extraction step, the main challenge in the feature extraction step is to find a suitable representation for the selected component. Multi-modal data can effectively address such a challenge, as introduced by Liu *et al.* [35]. Extracting data from different network layers of the OSI model [21], [101] results in better activity modeling of reviewers in a social review platform, and thus provides a more informative feature representation. Graph-based approaches can be treated as a scoring method to address different challenges such as the cold-start problem (introduced by Wang *et al.* [36]) through its capability in modeling complex relationships between the components.

VI. Conclusion

With an increase in the popularity of social review platforms for businesses and their customers, the traditional marketing media (e.g., broadcast and print) are losing ground to such web and mobile-based social review platforms. Fraudsters seize such opportunities for their benefits and take advantage of the media to write fraud reviews on targeted businesses. Characterizing fraud reviews provides an opportunity for new researchers to better understand the nature of such reviews, and hence improve the detection algorithms' performance. Several studies provided an overview to define fraud and examined the detection algorithms from different perspectives [19], [13], [40]. However, researchers may still be confused about the overall fraud detection framework. In this study, we proposed a systematic framework for fraud review detection with three stages: component selection, feature extraction and representation, and classification/scoring.

- In the component selection step, all studies considered either review, user, or item as their selected component to explore. While primary studies focused on fraud review detection, user group detection has recently attracted the attention of researchers, but it is still far from being fully resolved. Bot generation is another challenging topic due to the capability of deep learning-based review generation. Targeted item detection is also a potential topic to consider, as it is

not investigated as a separate component and has always been studied alongside user and review for multi-component classification.

- We analyzed feature extraction across three different categories; text-based, behavioral, and joint features. With recent progress in deep learning, new word embedding techniques have been employed to extract features and address the limitations of lingo-statistic text-based features. Behavioral features were also investigated individually or in a combination with text-based features as joint features.
- Classification/Scoring techniques were categorized based on the approaches used to deal with the lack of reliable data. Semi-supervised and unsupervised learning deals with partial labeling data for detection, while deep learning can suggest a considerable classification improvement over classic approaches. Data augmentation approaches such as GANs are a possible solution to address data paucity.

In addition to the technical challenges, new problems such as cold-start, bot fraud generation, and user group detection are becoming increasingly important. One possible direction for user group detection is to use graph-based solutions (e.g., Graph Convolutional Network) due to their effectiveness in realizing the pair-wise tightness of members of a group. Multi-modal data can also be considered as an important future work since it provides a comprehensive representation of all the different OSI model layers associated with an online social review platform to describe a component. To conclude, in this study, we introduced a systematic framework for fraud detection tasks. The techniques outlined for each step showed improvements in fraud detection. It is hoped that this review will encourage researchers to deepen their understanding of fraud detection and provide solutions to emerging challenges.

REFERENCES

[1] S. Chiramel, D. Logofatu, and G. Goldenthal, "Detection of social media platform insults using natural language processing and comparative study of machine learning algorithms," in *24th International Conference on System Theory, Control and Computing, ICSTCC 2020, Sinaia, Romania, October 8-10, 2020*, L. Barbulescu, Ed. IEEE, 2020, pp. 98–101. [Online]. Available: https://doi.org/10.1109/ICSTCC50638.2020.9259730

[2] C. Van Hee, G. Jacobs, C. Emmery, B. Desmet, E. Lefever, B. Verhoeven, G. De Pauw, W. Daelemans, and V. Hoste, "Automatic detection of cyberbullying in social media text," *PloS one*, vol. 13, no. 10, p. e0203794, 2018.

[3] J. Salminen, M. Hopf, S. A. Chowdhury, S.-g. Jung, H. Almerekhi, and B. J. Jansen, "Developing an online hate classifier for multiple social media platforms," *Human-centric Computing and Information Sciences*, vol. 10, no. 1, pp. 1–34, 2020.

[4] H. Choi, B. B. Zhu, and H. Lee, "Detecting malicious web links and identifying their attack types." *WebApps*, vol. 11, no. 11, p. 218, 2011.

[5] M. S. I. Mamun, M. A. Rathore, A. H. Lashkari, N. Stakhanova, and A. A. Ghorbani, "Detecting malicious urls using lexical analysis," in *International Conference on Network and System Security*. Springer, 2016, pp. 467–482.

[6] R. E. Agbefu, Y. Hori, and K. Sakurai, "Domain information based blacklisting method for the detection of malicious webpages," *International Journal of Cyber-Security and Digital Forensics*, vol. 2, no. 2, pp. 36–48, 2013.

[7] H. Rashkin, E. Choi, J. Y. Jang, S. Volkova, and Y. Choi, "Truth of varying shades: Analyzing language in fake news and political fact-checking," in *Proceedings of the 2017 conference on empirical methods in natural language processing*, 2017, pp. 2931–2937.

[8] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, "Fake news detection on social media: A data mining perspective," *ACM SIGKDD explorations newsletter*, vol. 19, no. 1, pp. 22–36, 2017.

[9] A. Pathak and R. K. Srihari, "Breaking! presenting fake news corpus for automated fact checking," in *Proceedings of the 57th annual meeting of the association for computational linguistics: student research workshop*, 2019, pp. 357–362.

[10] M. Ott, Y. Choi, C. Cardie, and J. T. Hancock, "Finding deceptive opinion spam by any stretch of the imagination," in *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies-volume 1*, Association for Computational Linguistics. Stroudsburg, PA, USA: Association for Computational Linguistics, 2011, pp. 309–319.

[11] M. Luca and G. Zervas, "Fake it till you make it: Reputation, competition, and yelp review fraud," *Management Science*, vol. 62, no. 12, pp. 3412–3427, 2016.

[12] C. Team, "How to fight fake reviews: Saving your business from the epidemic of fake reviews online," Jul. 2018. [Online]. Available: https://medium.com/chlunetwork/how-to-fight-fake-reviews-2ff82e643fa3

[13] M. Crawford, T. M. Khoshgoftaar, J. D. Prusa, A. N. Richter, and H. Al Najada, "Survey of review spam detection using machine learning techniques," *Journal of Big Data*, vol. 2, no. 1, p. 23, 2015.

[14] R. K. Dewang and A. K. Singh, "State-of-art approaches for review spammer detection: a survey," *Journal of Intelligent Information Systems*, vol. 50, no. 2, pp. 231–264, 2018.

[15] N. Jindal and B. Liu, "Opinion spam and analysis," in *In Proceedings of the 2008 international conference on web search and data mining*. Palo Alto, California, USA: ACM, 2008, pp. 219–230.

[16] R. Mohawesh, S. Xu, S. N. Tran, R. Ollington, M. Springer, Y. Jararweh, and S. Maqsood, "Fake reviews detection: A survey," *IEEE Access*, vol. 9, pp. 65 771–65 802, 2021.

[17] H. Paul and A. Nikolaev, "Fake review detection on online e-commerce platforms: a systematic literature review," *Data Mining and Knowledge Discovery*, vol. 35, no. 5, pp. 1830–1881, 2021.

[18] Y. Yuanshun, B. Viswanath, J. Cryan, H. Zheng, and B. Y. Zhao, "Automated crowdturfing attacks and defenses in online review systems," in *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*. Dallas, Texas, USA: ACM, 2017, pp. 1143–1158.

[19] A. Heydari, M. ali Tavakoli, N. Salim, and Z. Heydari, "Detection of review spam: A survey," *Expert Systems with Applications*, vol. 42, no. 7, pp. 3634–3642, 2015.

[20] R. Shebuit and L. Akoglu, "Collective opinion spam detection: Bridging review networks and metadata," in *Proceedings of the 21th acm sigkdd international conference on knowledge discovery and data mining. ACM*. New York, NY, USA: ACM, 2015, pp. 985–994.

[21] J. Wang, R. Wen, C. Wu, Y. Huang, and J. Xion, "Fdgars: Fraudster detection via graph convolutional networks in online app review system," in *Companion Proceedings of The 2019 World Wide Web Conference*, ser. WWW '19. New York, NY, USA: Association for Computing Machinery, 2019, p. 310–316. [Online]. Available: https://doi.org/10.1145/3308560.3316586

[22] G. Wang, S. Xie, B. Liu, and S. Y. Philip, "Review graph based online store review spammer detection," in *2011 IEEE 11th International Conference on Data Mining*. IEEE, 2011, pp. 1242–1247.

[23] M. Crawford, T. M. Khoshgoftaar, and J. D. Prusa, "Reducing feature set explosion to facilitate real-world review spam detection," in *Proceedings of the Twenty-Ninth International Florida Artificial Intelligence Research Society Conference, FLAIRS 2016, Key Largo, Florida, USA, May 16-18, 2016*, Z. Markov and I. Russell, Eds. AAAI Press, 2016, pp. 304–309. [Online]. Available: http://www.aaai.org/ocs/index.php/FLAIRS/FLAIRS16/paper/view/12844

[24] Y. Li, X. Feng, and S. Zhang, "Detecting fake reviews utilizing semantic and emotion model," in *2016 3rd International Conference on Information Science and Control Engineering (ICISCE)*, 2016, pp. 317–320.

[25] J. Yoo, S. Jo, and U. Kang, "Supervised belief propagation: Scalable supervised inference on attributed networks," in *2017 IEEE International Conference on Data Mining (ICDM)*, 2017, pp. 595–604.

[26] D. He, M. Pan, K. Hong, Y. Cheng, S. Chan, X. Liu, and N. Guizani, "Fake review detection based on pu learning and behavior density," *IEEE Network*, pp. 1–6, 2020.

[27] P. Zhao, X. Fu, W. Wu, D. Li, and J. Li, "Network-based feature extraction method for fraud detection via label propagation," in *2019 IEEE International Conference on Big Data and Smart Computing (BigComp)*, 2019, pp. 1–6.

[28] C. Lai, K. Xu, R. Y. Lau, Y. Li, and L. Jing, "Toward a language modeling approach for consumer review spam detection," in *2010*

IEEE 7th International Conference on E-Business Engineering. IEEE, 2010, pp. 1–8.

[29] S. Banerjee, A. Y. K. Chua, and J.-J. Kim, "Using supervised learning to classify authentic and fake online reviews," in *Proceedings of the 9th International Conference on Ubiquitous Information Management and Communication*, ser. IMCOM '15. New York, NY, USA: Association for Computing Machinery, 2015. [Online]. Available: https://doi.org/10.1145/2701126.2701130

[30] N. Hernandez, M. Rahman, R. Recabarren, and B. Carbunar, "Fraud de-anonymization for fun and profit," in *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security*, ser. CCS '18. New York, NY, USA: Association for Computing Machinery, 2018, p. 115–130. [Online]. Available: https://doi.org/10.1145/3243734.3243770

[31] O. Myle, C. Cardie, and J. Hancock, "Estimating the prevalence of deception in online review communities," in *In Proceedings of the 21st international conference on World Wide Web*. New York, NY, USA: ACM, 2012, pp. 201–210.

[32] B. Wang, J. Huang, H. Zheng, and H. Wu, "Semi-supervised recursive autoencoders for social review spam detection," in *2016 12th International Conference on Computational Intelligence and Security (CIS)*, 2016, pp. 116–119.

[33] H. Deng, L. Zhao, N. Luo, Y. Liu, G. Guo, X. Wang, Z. Tan, S. Wang, and F. Zhou, "Semi-supervised learning based fake review detection," in *2017 IEEE International Symposium on Parallel and Distributed Processing with Applications and 2017 IEEE International Conference on Ubiquitous Computing and Communications (ISPA/IUCC)*, 2017, pp. 1278–1280.

[34] L. Akoglu, R. Chandy, and C. Faloutsos, "Opinion fraud detection in online reviews by network effects," in *Seventh international AAAI conference on weblogs and social media*. Massachusetts, USA: AAAI, 2013, pp. 100–109.

[35] Y. Liu, B. Pang, and X. Wang, "Opinion spam detection by incorporating multimodal embedded representation into a probabilistic review graph," *Neurocomputing*, vol. 366, pp. 276–283, 2019.

[36] X. Wang, K. Liu, and J. Zhao, "Handling cold-start problem in review spam detection by jointly embedding texts and behaviors," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2017, pp. 366–376.

[37] S. K. Maurya, D. Singh, and A. K. Maurya, "Deceptive opinion spam detection approaches: a literature survey," *Applied Intelligence*, pp. 1–46, 2022.

[38] A. Mewada and R. K. Dewang, "A comprehensive survey of various methods in opinion spam detection," *Multimedia Tools and Applications*, pp. 1–41, 2022.

[39] S.-j. Ji, Q. Zhang, J. Li, D. K. Chiu, S. Xu, L. Yi, and M. Gong, "A burst-based unsupervised method for detecting review spammer groups," *Information Sciences*, 2020.

[40] D. U. Vidanagama, T. P. Silva, and A. S. Karunananda, "Deceptive consumer review detection: a survey," *Artificial Intelligence Review*, vol. 53, no. 2, pp. 1323–1352, 2020.

[41] S. Zhang, H. Yin, T. Chen, Q. V. N. Hung, Z. Huang, and L. Cui, "Gcn-based user representation learning for unifying robust recommendation and fraudster detection," *arXiv preprint arXiv:2005.10150*, 2020.

[42] B. Hooi, K. Shin, H. A. Song, A. Beutel, N. Shah, and C. Faloutsos, "Graph-based fraud detection in the face of camouflage," *ACM Trans. Knowl. Discov. Data*, vol. 11, no. 4, pp. 44:1–44:26, Jun. 2017. [Online]. Available: http://doi.acm.org/10.1145/3056563

[43] S. Kumar, B. Hooi, D. Makhija, M. Kumar, C. Faloutsos, and V. Subrahmanian, "Rev2: Fraudulent user prediction in rating platforms," in *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*. ACM, 2018, pp. 333–341.

[44] F. Song, R. Banerjee, and Y. Choi, "Syntactic stylometry for deception detection," in *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2012, pp. 171–175.

[45] X. Chang and J. Zhang, "Combating product review spam campaigns via multiple heterogeneous pairwise features," in *In Proceedings of the 2015 SIAM International Conference on Data Mining*. Vancouver, British Columbia, Canada: ACM, 2015, pp. 172–180.

[46] F. Li, M. Huang, Y. Yang, and X. Zhu, "Learning to identify review spam," in *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence - Volume Volume Three*, ser. IJCAI'11, vol. 3. Barcelona, Catalonia, Spain: AAAI Press, 2011, pp. 2488–2493. [Online]. Available: http://dx.doi.org/10.5591/978-1-57735-516-8/IJCAI11-414

[47] A. J., Minnich, N. Chavoshi, A. Mueen, S. Luan, and M. Faloutsos, "Trueview: Harnessing the power of multiple review sites," in *In Proceedings of the 24th International Conference on World Wide Web*, 2015, pp. 787–797.

[48] M. Arjun, B. Liu, and N. Glance, "Spotting fake reviewer groups in consumer reviews," in *In Proceedings of the 21st international conference on World Wide Web*. Lyon, France: ACM, 2012, pp. 191–200.

[49] S. Shehnepoor, M. Salehi, R. Farahbakhsh, and N. Crespi, "NetSpam: A networkbased spam detection framework for reviews in online social media," *IEEE Transactions on Information Forensics and Security*, vol. 12, no. 7, pp. 1585–1595, 2017.

[50] H. Aghakhani, A. Machiry, S. Nilizadeh, C. Kruegel, and G. Vigna, "Detecting deceptive reviews using generative adversarial networks," in *2018 IEEE Security and Privacy Workshops (SPW)*. San Francisco, USA: IEEE, 2018, pp. 89–95.

[51] R. Yafeng and Y. Zhang, "Deceptive opinion spam detection using neural network," in *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*. New York, NY, USA: Elsevier Science Inc., 2016, pp. 140–150.

[52] T. Duyu, B. Qin, and T. Liu, "Document modeling with gated recurrent neural network for sentiment classification," in *Proceedings of the 2015 conference on empirical methods in natural language processing*. Lisbon, Portugal: Association for Computational Linguistics, 2015, pp. 1422–1432.

[53] L. Quoc and T. Mikolov, "Distributed representations of sentences and documents," in *International Conference on Machine Learning*. Beijing, China: JMLR.org, 2014, pp. 1188–1196.

[54] M. Ott, Y. Choi, C. Cardie, and J. Hancock, "Finding deceptive opinion spam by any stretch of the imagination," in *In Proceedings of Annual Meeting of the Association for Computational Linguistics*, 2011.

[55] J. Shaohua, X. Zhang, X. Wang, and Y. Liu, "Fake reviews detection based on LDA," in *4th International Conference on Information Management (ICIM)*. Oxford, UK: IEEE, 2018, pp. 280–283.

[56] Z. You, T. Qian, and B. Liu, "An attribute enhanced domain adaptive model for cold-start spam review detection," in *Proceedings of the 27th International Conference on Computational Linguistics*, 2018, pp. 1884–1895.

[57] S. Shehnepoor, R. Togneri, W. Liu, and M. Bennamoun, "Dfraud^3: Multi-component fraud detection free of cold-start," *IEEE Transactions on Information Forensics and Security*, vol. 16, pp. 3456–3468, 2021.

[58] V. Sandulescu and M. Ester, "Detecting singleton review spammers using semantic similarity," in *Proceedings of the 24th international conference on World Wide Web*, 2015, pp. 971–976.

[59] R. Barbado, O. Araque, and C. A. Iglesias, "A framework for fake review detection in online consumer electronics retailers," *Information Processing & Management*, vol. 56, no. 4, pp. 1234–1244, 2019.

[60] N. Ruan, R. Deng, and C. Su, "Gadm: Manual fake review detection for o2o commercial platforms," *Computers & Security*, vol. 88, p. 101657, 2020.

[61] M. Jiang, A. Beutel, P. Cui, B. Hooi, S. Yang, and C. Faloutsos, "A general suspiciousness metric for dense blocks in multimodal data," in *2015 IEEE International Conference on Data Mining*. IEEE, 2015, pp. 781–786.

[62] ——, "Spotting suspicious behaviors in multimodal data: A general metric and algorithms," *IEEE Transactions on Knowledge and Data Engineering*, vol. 28, no. 8, pp. 2187–2200, 2016.

[63] C. Chiu, J. Zhan, and F. Zhan, "Uncovering suspicious activity from partially paired and incomplete multimodal data," *IEEE Access*, vol. 5, pp. 13 689–13 698, 2017.

[64] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," 2017.

[65] D. Kumar, Y. Shaalan, X. Zhang, and J. Chan, "Identifying singleton spammers via spammer group detection," in *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, 2018, pp. 656–667.

[66] K. Danilchenko, M. Segal, and D. Vilenchik, "Opinion spam detection: A new approach using machine learning and network-based

algorithms," in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 16, 2022, pp. 125–134.

[67] S. Shehnepoor, R. Togneri, W. Liu, and M. Bennamoun, "Scoregan: A fraud review detector based on multi task learning of regulated gan with data augmentation," 2021.

[68] C. Zhu, W. Zhao, Q. Li, P. Li, and Q. Da, "Network embedding-based anomalous density searching for multi-group collaborative fraudsters detection in social media," *Computers, Materials & Continua*, vol. 60, no. 1, pp. 317–333, 2019. [Online]. Available: http://www.techscience.com/cmc/v60n1/28363

[69] A. Bitarafan and C. Dadkhah, "$spgd\_hin$: Spammer group detection based on heterogeneous information network," in *2019 5th International Conference on Web Research (ICWR)*, 2019, pp. 228–233.

[70] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, ser. NIPS'14. Cambridge, MA, USA: MIT Press, 2014, pp. 2672–2680. [Online]. Available: http://dl.acm.org/citation.cfm?id=2969033.2969125

[71] J. S. Yedidia, W. T. Freeman, and Y. Weiss, *Understanding Belief Propagation and Its Generalizations*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2003, p. 239–269.

[72] A. Mukherjee, V. Venkataraman, B. Liu, N. Glance *et al.*, "Fake review detection: Classification and analysis of real and pseudo reviews," *UIC-CS-03-2013. Technical Report*, 2013.

[73] F. Abri, L. F. Gutiérrez, A. S. Namin, K. S. Jones, and D. R. Sears, "Linguistic features for detecting fake reviews," in *2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA)*. IEEE, 2020, pp. 352–359.

[74] Y. Xu, B. Shi, W. Tian, and W. Lam, "A unified model for unsupervised opinion spamming detection incorporating text generality," in *Proceedings of the 24th International Conference on Artificial Intelligence*, ser. IJCAI'15. AAAI Press, 2015, p. 725–731.

[75] W. Zhang, Y. Du, T. Yoshida, and Q. Wang, "Dri-rcnn," *Inf. Process. Manage.*, vol. 54, no. 4, p. 576–592, Jul. 2018. [Online]. Available: https://doi.org/10.1016/j.ipm.2018.03.007

[76] S. Jia, X. Zhang, X. Wang, and Y. Liu, "Fake reviews detection based on lda," in *2018 4th International Conference on Information Management (ICIM)*, 2018, pp. 280–283.

[77] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, ser. NIPS'13. Red Hook, NY, USA: Curran Associates Inc., 2013, p. 3111–3119.

[78] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186. [Online]. Available: https://www.aclweb.org/anthology/N19-1423

[79] M. Luca and G. Zervas, "Fake it till you make it: Reputation, competition, and yelp review fraud," *Manage. Sci.*, vol. 62, no. 12, p. 3412–3427, Dec. 2016. [Online]. Available: https://doi.org/10.1287/mnsc.2015.2304

[80] K. Goswami, Y. Park, and C. Song, "Impact of reviewer social interaction on online consumer review fraud detection," *Journal of Big Data*, vol. 4, no. 1, pp. 1–19, 2017.

[81] Y. Shaalan, X. Zhang, J. Chan, and M. Salehi, "Detecting singleton spams in reviews via learning deep anomalous temporal aspect-sentiment patterns," *Data Mining and Knowledge Discovery*, pp. 1–55, 2021.

[82] N. Kumar, D. Venugopal, L. Qiu, and S. Kumar, "Detecting anomalous online reviewers: An unsupervised approach using mixture models," *Journal of Management Information Systems*, vol. 36, no. 4, pp. 1313–1346, 2019. [Online]. Available: https://doi.org/10.1080/07421222.2019.1661089

[83] L. Xiang, H. You, G. Guo, and Q. Li, "Deep feature fusion for cold-start spam review detection," *The Journal of Supercomputing*, pp. 1–16, 2022.

[84] C. Xu and J. Zhang, "Combating product review spam campaigns via multiple heterogeneous pairwise features," in *Proceedings of the 2015 SIAM International Conference on Data Mining*. SIAM, 2015, pp. 172–180.

[85] A. Mukherjee, V. Venkataraman, B. Liu, and N. S. Glance, "What yelp fake review filter might be doing?" in *ICWSM*. Ann Arbor, MI, USA: The AAAI Press, 2013, pp. 134–144.

[86] Q. Li, Q. Wu, C. Zhu, J. Zhang, and W. Zhao, "An inferable representation learning for fraud review detection with cold-start problem," in *2019 International Joint Conference on Neural Networks (IJCNN)*, 2019, pp. 1–8.

[87] A. Mukherjee, A. Kumar, B. Liu, J. Wang, M. Hsu, M. Castellanos, and R. Ghosh, "Spotting opinion spammers using behavioral footprints," in *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '13. New York, NY, USA: Association for Computing Machinery, 2013, p. 632–640. [Online]. Available: https://doi.org/10.1145/2487575.2487580

[88] Q. Peng, "Store review spammer detection based on review relationship," in *Advances in Conceptual Modeling*, J. Parsons and D. Chiu, Eds. Cham: Springer International Publishing, 2014, pp. 287–298.

[89] J. Li, C. Cardie, and S. Li, "Topicspam: a topic-model based approach for spam detection." in *ACL (2)*. The Association for Computer Linguistics, 2013, pp. 217–221. [Online]. Available: http://dblp.uni-trier.de/db/conf/acl/acl2013-2.html#LiCL13

[90] N. Kumar, D. Venugopal, L. Qiu, and S. Kumar, "Detecting review manipulation on online platforms with hierarchical supervised learning," *Journal of Management Information Systems*, vol. 35, no. 1, pp. 350–380, 2018. [Online]. Available: https://doi.org/10.1080/07421222.2018.1440758

[91] M. Dong, L. Yao, X. Wang, B. Benatallah, C. Huang, and X. Ning, "Opinion fraud detection via neural autoencoder decision forest," *Pattern Recognition Letters*, vol. 132, pp. 21–29, 2020.

[92] P. Hajek, A. Barushka, and M. Munk, "Fake consumer review detection using deep neural networks integrating word embeddings and emotion mining," *Neural Computing and Applications*, pp. 1–16, 2020.

[93] C. M. Yilmaz and A. O. Durahim, "Spr2ep: A semi-supervised spam review detection framework," in *Proceedings of the 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, ser. ASONAM '18. IEEE Press, 2018, p. 306–313.

[94] Z. Wang, H. Li, and H. Wang, "Vote-based integration of review spam detection algorithms," *Applied Intelligence*, pp. 1–12, 2022.

[95] S. Liu, B. Hooi, and C. Faloutsos, "Holoscope: Topology-and-spike aware fraud detection," in *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, ser. CIKM '17. New York, NY, USA: ACM, 2017, pp. 1539–1548. [Online]. Available: http://doi.acm.org/10.1145/3132847.3133018

[96] X. Wang, K. Liu, and J. Zhao, "Detecting deceptive review spam via attention-based neural networks," in *Natural Language Processing and Chinese Computing*, X. Huang, J. Jiang, D. Zhao, Y. Feng, and Y. Hong, Eds. Cham: Springer International Publishing, 2018, pp. 866–876.

[97] J. K. Rout, A. Dalmia, K.-K. R. Choo, S. Bakshi, and S. K. Jena, "Revisiting semi-supervised learning for online deceptive review detection," *IEEE Access*, vol. 5, pp. 1319–1327, 2017.

[98] L.-y. Dong, S.-j. Ji, C.-j. Zhang, Q. Zhang, D. W. Chiu, L.-q. Qiu, and D. Li, "An unsupervised topic-sentiment joint probabilistic model for detecting deceptive reviews," *Expert Systems with Applications*, vol. 114, pp. 210–223, 2018.

[99] G. Xu, M. Hu, C. Ma, and M. Daneshmand, "Gscpm: Cpm-based group spamming detection in online product reviews," in *ICC 2019 - 2019 IEEE International Conference on Communications (ICC)*, 2019, pp. 1–6.

[100] C. Zhu, W. Zhao, Q. Li, P. Li, and Q. Da, "Network embedding-based anomalous density searching for multi-group collaborative fraudsters detection in social media," *Computers, Materials and Continua*, vol. 60, no. 1, pp. 317–333, 2019. [Online]. Available: http://dx.doi.org/10.32604/cmc.2019.05677

[101] R. Wen, J. Wang, C. Wu, and J. Xiong, "Asa: Adversary situation awareness via heterogeneous graph convolutional networks," in *Companion Proceedings of the Web Conference 2020*, ser. WWW '20. New York, NY, USA: Association for Computing Machinery, 2020, p. 674–678. [Online]. Available: https://doi.org/10.1145/3366424.3391266

[102] R. Narayan, J. K. Rout, and S. K. Jena, "Review spam detection using semi-supervised technique," in *Progress in Intelligent Computing Techniques: Theory, Practice, and Applications*. Springer, 2018, pp. 281–286.

[103] R. Hassan and M. R. Islam, "Detection of fake online reviews using semi-supervised and supervised learning," in *2019 International Conference on Electrical, Computer and Communication Engineering (ECCE)*, 2019, pp. 1–5.

[104] A. Pathak and R. K. Srihari, "Breaking! presenting fake news corpus for automated fact checking," in *Proceedings of the 57th annual meeting of the association for computational linguistics: student research workshop*, 2019, pp. 357–362.

[105] H. Rashkin, E. Choi, J. Y. Jang, S. Volkova, and Y. Choi, "Truth of varying shades: Analyzing language in fake news and political fact-checking," in *Proceedings of the 2017 conference on empirical methods in natural language processing*, 2017, pp. 2931–2937.

[106] R. Vijjali, P. Potluri, S. Kumar, and S. Teki, "Two stage transformer model for covid-19 fake news detection and fact checking," *arXiv preprint arXiv:2011.13253*, 2020.

[107] S. Wang, W. Mao, P. Wei, and D. D. Zeng, "Knowledge structure driven prototype learning and verification for fact checking," *Knowledge-Based Systems*, vol. 238, p. 107910, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0950705121010650

[108] A. Porshnev, I. Redkin, and A. Shevchenko, "Machine learning in prediction of stock market indicators based on historical data and data from twitter sentiment analysis," in *2013 IEEE 13th International Conference on Data Mining Workshops*. IEEE, 2013, pp. 440–444.

[109] S. Li, Y. Wang, J. Xue, N. Zhao, and T. Zhu, "The impact of covid-19 epidemic declaration on psychological consequences: a study on active weibo users," *International journal of environmental research and public health*, vol. 17, no. 6, p. 2032, 2020.

[110] J. Marciano, "Fake online reviews cost $152 billion a year. Here's how e-commerce sites can stop them," https://www.weforum.org/agenda/2021/08/fake-online-reviews-are-a-152-billion-problem-heres-how-to-silence-them/, Aug 10, 2021, [Online; accessed 04-Aug-2022].

[111] S. Dickens, "Introducing New Machine Learning Techniques to Help Stop Scams," https://www.facebook.com/notes/10157814548136886/, March 26, 2018, [Online; accessed 04-Aug-2022].

[112] J. D. Day and H. Zimmermann, "The osi reference model," *Proceedings of the IEEE*, vol. 71, no. 12, pp. 1334–1340, 1983.

[113] P. Gupta, S. Gandhi, and B. R. Chakravarthi, "Leveraging transfer learning techniques-bert, roberta, albert and distilbert for fake review detection," in *Forum for Information Retrieval Evaluation*, 2021, pp. 75–82.

[114] M. Matassoni, R. Gretter, D. Falavigna, and D. Giuliani, "Non-native children speech recognition through transfer learning," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 6229–6233.

[115] P. G. Shivakumar and P. Georgiou, "Transfer learning from adult to children for speech recognition: Evaluation, analysis and recommendations," *Computer speech & language*, vol. 63, p. 101077, 2020.

[116] S. Shen, M. Sadoughi, M. Li, Z. Wang, and C. Hu, "Deep convolutional neural networks with ensemble learning and transfer learning for capacity estimation of lithium-ion batteries," *Applied Energy*, vol. 260, p. 114296, 2020.

[117] Y. Yao, B. Viswanath, J. Cryan, H. Zheng, and B. Y. Zhao, "Automated crowdturfing attacks and defenses in online review systems," in *Proceedings of the 2017 ACM SIGSAC conference on computer and communications security*, 2017, pp. 1143–1158.

[118] M. A. Al-Garadi, K. D. Varathan, S. D. Ravana, E. Ahmed, and V. Chang, "Identifying the influential spreaders in multilayer interactions of online social networks," *Journal of Intelligent & Fuzzy Systems*, vol. 31, no. 5, pp. 2721–2735, 2016.

[119] B. Oselio, A. Kulesza, and A. O. Hero, "Multi-layer graph analysis for dynamic social networks," *IEEE Journal of Selected Topics in Signal Processing*, vol. 8, no. 4, pp. 514–523, 2014.

[120] S. Sivasankari and G. Vadivu, "Tracing the fake news propagation path using social network analysis," *Soft Computing*, pp. 1–9, 2021.

[121] W. L. Hamilton, R. Ying, and J. Leskovec, "Inductive representation learning on large graphs," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, ser. NIPS'17. Red Hook, NY, USA: Curran Associates Inc., 2017, p. 1025–1035.