2019 IEEE 35th International Conference on Data Engineering (ICDE)

CATS: Cross-Platform E-commerce Fraud Detection

Haiqin Weng [1], Shouling Ji [1], Fuzheng Duan [1], Zhao Li [2], Jianhai Chen [1], Qinming He [1], Ting Wang [3]

[1] *Institute of Cyberspace Research and College of Computer Science and Technology, Zhejiang University*
{hq_weng, sji, 3150105576, chenjh919, hqm}@zju.edu.cn

[2] *Alibaba Group, Hangzhou, China*
lizhao.lz@alibaba-inc.com

[3] *Lehigh University, USA*
ting@cse.lehigh.edu

***Abstract*—Nowadays, the popularity of e-commerce has brought huge economic benefits to factories, third-party merchants, and e-commerce service providers. Driven by such huge economic benefits, malicious merchants attempt to promote items through inserting fraudulent purchases, fake review scores, and/or feedback, into them. Mitigating this threat is challenging due to the difficulty of obtaining internal e-commerce data, the variance of e-commerce services used by malicious merchants, and the reluctance of service providers in cooperation. In this paper, we present an efficient, platform-independent, and robust e-commerce fraud detection system, CATS, to detect frauds for different large-scale e-commerce platforms. We implement the design of CATS into a prototype system and evaluate this prototype on the world's popular e-commerce platform `Taobao`[1]. The evaluation result on `Taobao` shows that CATS can achieve a high accuracy of 91% in detecting frauds. Based on this success, we then apply CATS on another large-scale e-commerce platforms, and again CATS achieves an accuracy of 96%, suggesting that CATS is very effective on real e-commerce platforms. Based on the cross-platform evaluation results, we conduct a comprehensive analysis on the reported frauds and reveal several abnormal yet interesting behaviors of those reported frauds. Our study in this paper is expected to shed light on defending against frauds for various e-commerce platforms.**

I. Introduction

Nowadays, *e-commerce* efficiently connects customers (denoted as users in this paper) with factories, stores, and third-party merchants, providing them with a convenient, fast and reliable manner of shopping. Due to the numerous advantages of e-commerce, more and more people prefer online shopping over conventional shopping. The e-commerce retail sales are quickly expanding, and have brought huge economic benefits to factories, third-party merchants, and e-commerce service providers. For example, `Taobao`'s GMV (gross merchandise volume) is reported to reach US $320 billion (RMB $2,202$ billion) in the fiscal year of 2017[2], `Amazon`'s GMV is reported to reach US $149 billion (RMB 970 billion) in 2016[3], and `Jingdong`'s GMV is reported to reach US $101 billion (RMB 658 billion) in 2016[4].

Fraud Items on Various E-commerce Platforms. Naturally, the incredibly substantial economic benefits also attract malicious merchants to promote their items illegally. In reality, people are inclined to buy such items that are frequently purchased, have high review scores, and/or have positive feedback [1]. Hence, aiming to gain higher economic benefits, those malicious merchants always attempt to promote their targeted items through illegally inserting fraudulent purchases, fake review scores, and/or feedback, into them. For convenience, we name such illegally promoted items as *fraud items*. As reported in [2], [3], [4], fraud items commonly exist across many large-scale e-commerce platforms, including `Amazon`, `Taobao` and `Jingdong`.

E-commerce Fraud Detection. Fraud items are very harmful to the e-commerce ecosystem and cause unfair competitions, e.g., they provide fake information to users and can further mislead them to make improper decisions. However, understanding and detecting fraud items remains a challenge. For the e-commerce platform, who provide e-commerce service for users, stores and merchants, determining whether an item is fraududent or normal is to some extent bounded because of privacy commitments and ethical concerns. Even when the platform is willing to do so, it is difficult and improper for it to inspect other items outside of its platform. Effectively collaboration with other e-commerce platforms is by no means possible in practice due to business competitions. Exploring this issue becomes even more difficult if the internal e-commerce data (e.g., click records and user-item bipartite graph) is not available.

As a result, existing techniques cannot be directly applied to detect e-commerce frauds, simply because they are either dedicated to one specific platform [4], which cannot be easily extended to other platforms, or designed for other application domains, such as click fraud [5], content fraud [6] and account fraud [7], [8]. So far, little has been done to understand the characteristics of e-commerce frauds that are existing across many large-scale e-commerce platforms from the perspectives of cross-platform, compatibility, and third-party. A third-party, cross-platform, and compatible system tends to be a more feasible, justified and effective solution to e-commerce fraud detection since it does not show partiality for any e-commerce platforms, it can be extended to various e-commerce platforms, including `Taobao`, `Jingdong`, and `Amazon`, and it is a more efficient way to detect frauds as it is based on public e-commerce data, which can be directly affected by malicious promotions.

[1] https://en.wikipedia.org/wiki/Taobao
[2] http://www.alibabagroup.com/en/news/press_pdf/p170518.pdf
[3] http://phx.corporate-ir.net/phoenix.zhtml?c=97664&p=irol-reportsannual
[4] http://ir.jd.com/phoenix.zhtml?c=253315&p=irol-IRHome

2375-026X/19/$31.00 ©2019 IEEE
DOI 10.1109/ICDE.2019.00203

Our Methodology. Aiming at developing an efficient, platform-independent and robust third-party e-commerce fraud detection system, we present a **C**ross-platform **A**n**T**i-fraud **S**ystem (CATS) in this paper. CATS detects fraud items mainly based on a group of features identified through analyzing labeled fraud items. It mainly consists of four components: *a data collector*, *a semantic analyzer*, *a feature extractor*, and *a detector*. The data collector is used for collecting data from the public domain of e-commerce platforms. The semantic analyzer helps perform in-depth semantic analysis on e-commerce data, which will be used for feature preparation. Specially, it trains a *word2vec* model and provides a sentiment analysis model. The feature extractor prepares features for items using the open domain e-commerce data collected by the data collector. Based on the prepared features, the detector detects e-commerce frauds by using a binary classifier to determine whether an item is fraudulent or normal. In our design, CATS detects frauds through analyzing the public domain data of an e-commerce platform. It is a cross-platform e-commerce fraud detection system and can directly work on various large-scale platforms.

Applications. To evaluate the performance of CATS, we first use CATS to detect fraud items on the world's popular e-commerce platform, `Taobao`. After running CATS, we report the detection results of CATS on `Taobao` to `Alibaba`. Through the analysis of domain experts, `Alibaba` confirms that 91% of the results are truly frauds, which indicates that CATS is very effective on real e-commerce platforms.

Then, we apply CATS to another large-scale e-commerce platform to detect fraud items from millions of e-commerce items. To validate the detection results on this platform, we employ a methodology that combines manual analysis and statistical analysis. Through manual analysis, we confirm 96% of the reported fraud items. A further comprehensive measurement study on the reported fraud items demonstrates that these fraud items behave fraudulently in various aspects, including the reliability of the users who purchased these items, the characteristics of these items' feedback and the source of these items' orders.

Contributions. We summarize our contributions in this paper as follows.

- *Features.* We have identified a group of platform-independent features from the word level, the semantic level and the structural level to discriminate fraud and normal items on different e-commerce platforms.
- *Cross-platform Anti-Fraud System, CATS.* We develop a platform-independent and efficient third-party e-commerce fraud detection system, named CATS. CATS can effectively detect fraud items for different e-commerce platforms through inspecting their public domain data.
- *Implementation and Applications of CATS.* We implement the design of CATS into a prototype system and validate its performance on two different large-scale e-commerce platforms, on which CATS achieves high precision. We also recommend CATS to `Alibaba`, who confirms our findings and has partially incorporated CATS into its e-commerce platform `Taobao`.
- *Open Source.* Aiming at facilitating the e-commerce fraud detection research, we will gradually make the prototype of CATS be open source at [9]. It is also expected that our system can be helpful for more e-commerce platforms on defending against various online frauds.

Roadmap. The rest of the paper is organized as follows. Section II describes our analysis and findings on e-commerce frauds, and the design and implementation of CATS. Section III provides the experimental evaluation of CATS on `Taobao`. Section IV shows the experimental evaluation on the other tested platform and a comprehensive measurement study. Section VI discusses the deployment of CATS. We make further discussion in Section VII. Section VIII compares our work with related prior research, and Section IX concludes this paper.

II. Cross-Platform Anti-Fraud System

In this section, we first elaborate our analysis on fraud items and the features identified for discriminating fraud items and normal items. These features are used in our research to build our anti-fraud system, CATS. Then, we give an overview of CATS, followed by its detailed design and implementation.

A. Features of E-commerce Frauds

Based on the comments, we construct the features of e-commerce items. We use comments in our research simply because they are available to anyone from the public domain and they also become the primary means for malicious merchants to promote the targeted items. For example, in the e-commerce platform, an item will make a popular impression among users if it has a large number of positive comments, revealing that this item has a good quality and is deserved to buy. Thus, it is feasible and easy for malicious merchants to promote the targeted items through polluting comments.

Specially, we show the features identified for discriminating fraud items and normal items from three categories: the *word level feature*s, the *semantic feature*s, and the *structural feature*s. In the following, we first give some mathematical notations for facilitating the discussion, followed by introducing the word level, the semantic, and the structural features.

1) Notations: Let $I = \{I_i | i = 1, 2, \dots\}$ be the set of items. Given an item $I_i \in I$, let $C_i = \left\{C_i^j | j = 1, 2, \dots\right\}$ be the set of I_i's comments. Since each comment C_i^j is presented in the form of a short paragraph, let $C_i^j = \left\{C_i^j(t) | t = 1, 2, \dots\right\}$ denote the set of C_i^j's word segmentation result. For example, given a comment "我很喜欢这件商品 (I like this item so much)", the word set, {我, 很, 喜欢, 这件, 商品}({I, like, this, item, so much}), is the word segmentation result of this comment. Now, let $W = \{W_i | i = 1, 2, \dots\}$ be the set of e-commerce words. Let $P = \{P_i | P_i \in W$ and P_i is a positive word$\}$ and $N = \{N_i | N_i \in W$ and N_i is a negative word$\}$ be the sets of positive and negative words, respectively. For instance, 很好

TABLE I
THE POSITIVE SET AND THE NEGATIVE SET.

Type	Keywords
Positive Set	好评(good reputation), 好坪(good reputation) , 划算(cost-effective), 值得(deserve), 赞(like), 漂亮(beautiful), 好平(good reputation) , 很好(very good), 合适(suitable), 精致(delicacy), etc.
Negative Set	差评(negative reputation), 恶意(malevolence), 最烂(the worst), 不讲理(unreasonable), 太过分(a bit thick), 抵赖(deny), 可恨(hateful), 退货(sales return), 一星(one star), 威胁(threat), etc.

(very good) and 舒适 (comfort) are two examples of positive words, and 差评 (bad reputation) and 糟糕 (terrible) are two examples of negative words.

2) Word Level Features: Given an item, its comments are direct feedback from the users who have purchased it. It is intuitive that: positive words (e.g., good quality) contained in the comments will draw more attention from users while negative words (e.g., bad quality) may probably drive potential buyers away. Hence, we attempt to inspect the word difference between fraud and normal items' comments, through which we construct several word level features.

A key observation from our study is that *the comments of a fraud item tend to be filled with positive words and tend to have no negative words.* This observation is consistent with our intuition: fraud items need positive words in their comments, e.g., 很好 (very good) and 舒适 (comfort), to make a false impression that they are popular among users; whereas, normal items' comments are true feedback from e-commerce users and are expected to contain positive words, negative words and many neutral words. In our research, we call this the *deceptive* characteristics of a fraud item. According to this observation, we construct two word level features: *averagePositiveNumber* and *averagePositive/NegativeNumber*, to capture the deceptive characteristics of fraud items. The *averagePositiveNumber* describes the average number of positive words in an item's comments. The *averagePositive/NegativeNumber* measures the difference between the average numbers of positive and negative words in an item's comments. To this end, we first need to identify the set of positive words, P, and the set of negative words, N. Equipped with P and N, we can then calculate the *averagePositiveNumber* and the *averagePositive/NegativeNumber* of an item.

Positive and Negative Sets. We employ the *word2vec* model [10] to construct the positive set P and the negative set N based on a few seed words. Our construction steps of P and N are shown as follows. First, we train a word2vec model on a corpus of over 70 million records of comments, collected from the `Taobao` platform during August 2017. Second, we utilize this trained word2vec model to search for words similar to the seeds iteratively. For building P as an example, we initialize the seeds as a few positive words, e.g., 好评 (good reputation). Then, we search the k-nearest neighbors of the seeds, followed by iteratively search the k-nearest neighbors of these neighbors. Finally, with the help of the trained word2vec model, we obtain P based on several positive words, e.g., 好评 (good reputation), and obtain N based on several negative words, e.g., 差评 (bad reputation). In total, P contains ~ 200 positive words, and N contains ~ 200 negative words, as shown in Table I. From Table I, we can see that the word2vec model can even find homograph words, e.g., 好评 (good reputation), 好坪 (good reputation), and 好平 (good reputation), which may even be difficult for human experts to figure out. Note that, for computation efficiency, we limit the sizes of both the positive and the negative sets in our research.

After building P and N, we construct the two word level features: *averagePositiveNumber* and *averagePositive/NegativeNumber*. Refer that *averagePositiveNumber* describes the average number of positive words in an item's comments. Given an item $I_i \in I$, its *averagePositiveNumber* is measured by $\frac{\sum_j |C_i^j \cap P|}{|C_i|}$, where $|\cdot|$ denotes the size of the set. Refer that *averagePositive/NegativeNumber* measures the difference between the number of positive words and the number of negative words within an item's comments. Given an item $I_i \in I$, its *averagePositive/NegativeNumber* is measured by $\frac{\sum_j \big||C_i^j \cap P| - |C_i^j \cap N|\big|}{|C_i|}$, where $\|\cdot\|$ denotes the absolute value.

The word level features based on a single positive word may not carry the entire *deceptive* characteristics of fraud items. Therefore, we also consider using the *n*-gram of words for better characterizing a fraud item. In the field of computational linguistics, the *n*-gram is defined as *a contiguous sequence of n words from a given sample of text.* Specially, for our purpose, we introduce two other word level features: *averageNgramNumber* and *averageNgramRatio*. The *averageNgramNumber* describes the average number of positive *n*-grams in an item's comments. In our research, we focus on the 2-gram of words. Let $G = \{(W_i, W_j) | \exists W_i, W_j \in W\}$ denote the set of positive 2-grams, where (W_i, W_j) is a 2-gram and at least one word of W_i and W_j is from the positive set P. Given an item $I_i \in I$, its *averageNgramNumber* is measured by $\frac{\sum_j \sum_t \delta\big((C_i^j(t), C_i^j(t+1)) \in G\big)}{|C_i|}$, where $\delta(\cdot)$ is the indicator that gives 1 when the condition is true, 0 otherwise. The *averageNgramRatio* describes the average ratio of positive *n*-grams in an item's comments, which is measured by $\sum_j \sum_t \frac{\delta\big((C_i^j(t), C_i^j(t+1)) \in G\big)}{|C_i| \times (|C_i^j| - 1)}$.

3) Semantic Features: In addition to the word level features, we find that most of the fraud items' comments convey the intense emotion that these items are truly deserved to buy. For example, we randomly pick $5,000$ fraud items with $\sim 70,000$ comments, and $5,000$ normal items with $\sim 70,000$ comments from `Taobao` (we will describe the used dataset in details later). The sentiment distributions of the comments for fraud and normal items are shown in Fig. 1, where a large value represents a positive sentiment while a small

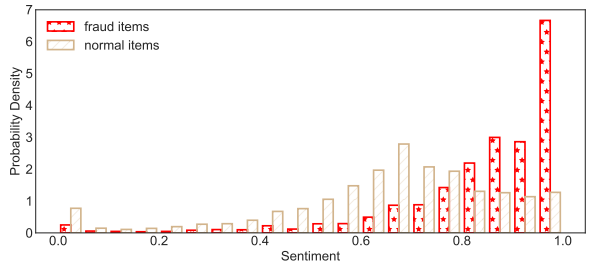


Fig. 1. Distribution of comments' sentiments.

value represents a negative sentiment. From Fig. 1, we can see that the sentiment distribution of fraud items' comments concentrates on large values near 1, while the sentiment distribution of normal items' comments concentrates on small values near 0.7. The reason is evident: for fraud items, a large portion of their comments are generated from the promotion activities of malicious users, and thus the sentiments of their comments are more positive and concentrate on large values; whereas, for normal items, nearly all of their comments are generated from the activities of benign users, and thus the sentiments of their comments are comparatively neutral and concentrate on relatively small values.

Based on the above observation, it is possible for us to employ the average sentiment of items' comments, denoted as *averageSentiment*, to differentiate fraud items from normal items. Specifically, in our research, we use the pre-trained sentiment model to calculate the sentiment for e-commerce comments. The sentiment model, provided by SnowNLP [11], is trained on a large-scale e-commerce data collected from real platforms, e.g., `Taobao` and `Amazon`, etc.

4) Structure Features: Finally, we explore the internal structure of comments. Listing 1 shows two representative comments: one is for a fraud item, and the other is for a normal item. From Listing 1, we can see that (1) the fraud item's comment is longer than that of the normal item; (2) the fraud item's comment is organized in a more chaotic way than that of the normal item; (3) the fraud item's comment has more punctuations than that of the normal item; and (4) the fraud item's comment has more duplicate words than that of the normal item. We conjecture the reasons as follows. Most of the fraud items' comments are from malicious promotions. These comments are composed long for carrying more promotive information and are prepared complicatedly to be more vivid and attractive. Such long and complicate organization of comments leads to more punctuations and repeated words. Based on this analysis, we attempt to extract structure the features from the writing style of comments.

Given a comment C_i^j, how chaotic C_i^j is organized can be measured by C_i^j's entropy, which is defined as $-\sum_t p(C_i^j(t)) \log p(C_i^j(t))$, where $p(C_i^j(t))$ denotes the frequency of $C_i^j(t)$ in this comment. Furthermore, to dive deeper into the internal structure of comments, we again randomly pick $5,000$ fraud items with $\sim 70,000$ comments, and $5,000$ normal items with $\sim 70,000$ comments from `Taobao`. Fig. 2 shows the distributions of the number of punctuations in the fraud and normal items' comments, Fig. 3 shows the entropy distributions of the fraud and normal items' comments, Fig. 4 shows the length distributions of the fraud and normal items' comments, and Fig. 5 shows the distributions of the unique word ratio in the fraud and normal items' comments.

From Fig. 2 – Fig. 5, we can observe that fraud and normal items have different distributions concerning the punctuation number, the comment entropy, the comment length and the number of unique words in their comments. Based on this observation, we introduce five features: *averagePunctuationRatio*, *sumPunctuationNumber*, *averageCommentEntropy*, *averageCommentLength* and *sumCommentLength*, to differentiate fraud items from normal ones. Formally, given an item I_i, the *averagePunctuationRatio* measures the average ratio of punctuations in I_i's comments, the *sumPunctuationNumber* measures the number of punctuations in I_i's comments, the *averageCommentEntropy* measures the average entropy of I_i's comments, the *averageCommentLength* measures the average length of I_i's comments, the *sumCommentLength* measures the sum of comment length of I_i's comments, and the *uniqueWordRatio* measures the ratio between the number of unique words and the overall words in I_i's comments.

The comment for a fraud item:
之前在别家买了一个，用了不到一年就坏了，所以这次看了很多家，最后买了这个，是因为相信这个品牌，而且这个价格实惠！扫码枪做工挺好，拿来试用了一下识别很准很快，精度高，质量还好，希望耐用些。
(I had bought the same kind of barcode scanner from another online shop before. However, it took less than a year to go bad. Therefore, in this time, I shopped around and finally bought this barcode scanner. I chose to buy this item since I believed its brand and the price of this brand was affordable to me. Moreover, this barcode scanner has a high performance-price ratio. Specifically, this barcode scanner is well-made and qualified, it can quickly recognize the bar code, and it achieves high precision in recognizing bar code. I hope the scanner can be used for a long time.)

The comment for a normal item:
书很好，涨知识。(The book is good and widens my knowledge.)

Listing 1. the fraud item's comment v.s. the normal item's comment

In summary, we have identified 11 features from the word level, the semantic level and the structure level, as shown in Table II.

B. Design of CATS

CATS includes four components: *a data collector*, *a semantic analyzer*, *a feature extractor*, and *a detector*. Fig. 6 illustrates the architecture of CATS. The data collector can obtain data directly from e-commerce platforms (e.g., through APIs) as well as collecting data from the public domain of

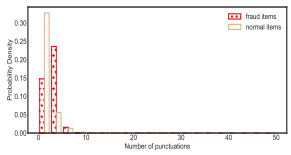


Fig. 2. Distribution of the number of punctuations in items' comments.

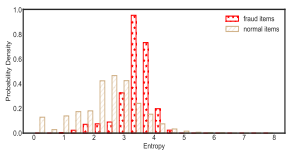


Fig. 3. Entropy distribution of the comments entropy.

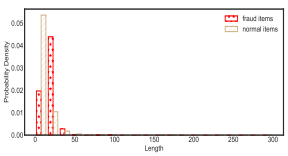


Fig. 4. Length distribution of the comments.

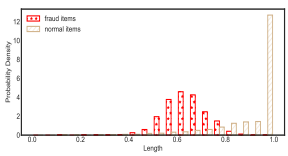


Fig. 5. The distribution of unique word ratio.

TABLE II
DESCRIPTIONS OF FEATURES.

Name	Description
averagePositiveNumber	the average number of positive words in an item's comments
averagePositive/NegativeNumber	the difference between the number of positive words and the number of negative words in an item's comments
uniqueWordRatio	the ratio between the number of unique words and the overall words in an item's comments
averageSentiment	the average sentiment of items' comments
averageCommentEntropy	the average entropy of an item's comments
averageCommentLength	the average length of an item's comment
sumCommentLength	the sum of comment length of an item's comment
sumPunctuationNumber	the number of punctuations in an item's comments
averagePunctuationRatio	the average ratio of punctuations in an items' comments
averageNgramNumber	the average number of positive n-grams in an item's comments
averageNgramRatio	the average ratio of positive n-grams in an item's comments

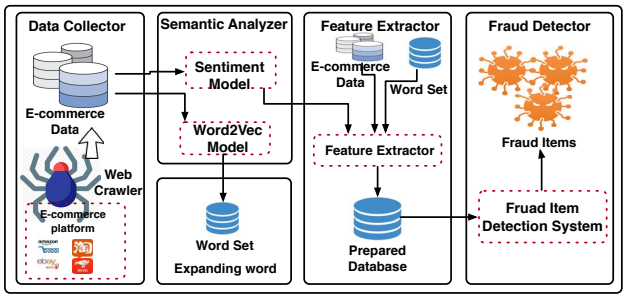


Fig. 6. Architecture of CATS.

e-commerce platforms. Also, the data collector can filter the noisy data (e.g., duplicated data records). In Subsection IV-A, we will give an example of the data collector. The semantic analyzer is mainly responsible for analyzing the semantic relationships within e-commerce data. To be specific, the semantic analyzer trains a *word2vec* model by using a large-scale corpus of e-commerce comments, which is further utilized to search the positive set and the negative set of words. In addition to the word2vec model, the semantic analyzer provides a sentiment analysis model, which calculates the sentiment for each comment.

Given an item $I_i \in I$, the feature extractor prepare its features as follows. It first segments each of its comments into a word set C_i^j. Then, based on all the word sets of I_i's comments, the feature extractor calculates the *averageSentiment* with the help of the semantic analyzer, prepares the *averagePositiveNumber* and the *averagePositive/NegativeNumber* by using the positive set P and the negative set N, and prepares the *averageNgramNumber* and *averageNgramRatio* by using the 2-gram set G. After that, the feature extractor calculates the *uniqueWordRatio*, *averageCommentEntropy*, *averageCommentLength*, *sumCommentLength*, *averagePunctuationRatio* and *sumPunctuationNumber*, for I_i in a statistical way.

TABLE III
PERFORMANCE COMPARISON UNDER FIVE-FOLD CROSS VALIDATION.

Classifier	Precision	Recall
Xgboost	0.93	0.90
SVM	0.99	0.62
AdaBoost	0.90	0.90
Neural Network	0.83	0.65
Decision Tree	0.86	0.90
Naive Bayes	0.91	0.65

The detector detects fraud items through two stages. First, it filters part of the items according to some rules, e.g., filtering the e-commerce items, of which the sales volumes are less than 5, and filtering the e-commerce items which contain no positive n-grams or words. Then, it trains a binary classifier based on the extracted features to detect the fraud items. Since different features have different power in differentiating fraud items and normal items, the detector needs a binary classifier with a model for weighting the features. In the following, we experimentally show the selection of the binary classifier in our research.

Selection of the Classifier. CATS utilizes an Xgboost [12] model as the classifier in its detector. Below, we describe how and why we select the Xgboost model as the binary classifier. In our research, we do a performance comparison experiment to pick up the best one from the six commonly used models: Xgboost, SVM, AdaBoost, Neural Network, Decision Tree and Naive Bayes. The dataset used in this performance comparison experiment is a small ground-truth dataset with 5000 fraud items and 5000 normal items, provided by `Taobao`. We first use CATS' feature extractor to prepare the numerical features for this dataset. Then, we test the effectiveness of the six candidate classifiers all through the standard five-cross validation: that is, $4/5$ of the data is used for training the classifier, and $1/5$ of the data is used for testing the classifier.

TABLE IV
THE LABELED DATASET FROM TAOBAO. FI = FRAUD ITEMS, NI = NORMAL ITEMS.

Dataset	#FI	#NI	#comments
D_0	14,000	20,000	474,000

Table III shows the performance of the six candidate classifiers. From Table III, we can see that Xgboost shows a relatively better performance than the other five. Thus, we choose the Xgboost [12] model as the classifier. Note that, in practice, it is not necessary to choose the Xgboost model, and any classifier that shows satisfactory performance can be employed. Fig. 7 illustrates the feature importance of the Xgboost model, which is measured by the times this feature is split during the construction process of the Xgboost model. From Fig. 7, we observe that all of the extracted features are important to our classifier, of which *sumCommentLength*, *averageCommentEntropy* and *averageSentiment* are the three most important features.

Implementation. According to the design, we implement CATS as a prototype system. CATS' data collector is built upon data APIs provided by e-commerce platforms and the open source collaborative framework, Scrapy [13]. For the semantic analyzer, its major functions are written in Python, its word2vec model is provided by the machine learning system library, TensorFlow [14], and its sentiment model is provided by the simplified Chinese text processing library, SnowNLP [11]. CATS' feature extractor is implemented in a parallelized style for fast processing. In CATS' detector, we incorporate Xgboost [12] for determining whether an item is fraudulent or normal. The Xgboost model is pre-trained on a labeled dataset provided by `Taobao`, denoted by D_0. D_0 contains $14,000$ fraud items, $20,000$ normal items, and $474,000$ comments. We summarize the statistics of D_0 in Table IV.

III. EVALUATION ON TAOBAO

In this section, we evaluate the performance of CATS on the popular e-commerce platform of China, `Taobao`. As described in Subsection II-B, we pre-train CATS' detector on the labeled dataset D_0 before using it.

A. Dataset

TABLE V
THE DATASET FROM TAOBAO. FI = FRAUD ITEMS, AND NI = NORMAL ITEMS.

Dataset	#FI	#NI	#comments
D_1	18,682	$1,461,452$	$72,340,999$

To evaluate CATS' performance on `Taobao`, we obtain a large-scale labeled e-commerce dataset from the `Alibaba` Group, which belongs to the same data repository as the dataset used by our recently published paper [15]. This dataset, denoted by D_1, was generated in 2017. D_1 contains $1,480,134$ online items collected from $15,992$ shops, and $72,340,999$ comments for these items. Within D_1, there are $18,682$ fraud items and $1,461,452$ normal items. Among the $18,682$ fraud items, $16,782$ items are labeled as fraud since there exist sufficient evidence (e.g., the evidence of financial transactions between the malicious merchants and users) for them, and the remaining $1,900$ items are labeled as fraud through the manual analysis from `Alibaba`'s anti-fraud experts. We summarize the statistics of D_1 in Table V. Note that, (1) D_1 is not overlapped with D_0; and (2) D_1 is used for examining the performance of CATS, and any derived information (e.g., the fraud item-benign item ratio) does not represent the real scenario of `Alibaba`.

B. Performance

Now, we test CATS on D_1. This experiment is conducted on a server equipped with 40 Intel Xeon E5-2640V4 vCPUs and 96 GB memory. The precision, recall and F-score of CATS on D_1 are shown in Table VI.

From Table VI, we can see that CATS detects 92% of the overall fraud items with a precision of 83%. For the fraud items labeled with sufficient evidence, CATS detects 92% of them with a precision of 83%. In summary, CATS has high precision, recall and F-score on D_1, which suggests the effectiveness of CATS.

IV. APPLICATION AND ANALYSIS

Based on the success of CATS on `Taobao`, we now employ CATS to detect fraud items on another e-commerce platform. We select `E-platform` in our experimental evaluation since it is a large-scale B2C retailer in China and has tremendous e-commerce data. In this evaluation, the binary classifier contained in CATS' detector is also pre-trained on the labeled dataset D_0 provided by `Taobao`.

We then make a comprehensive measurement study to demonstrate that the fraud items reported by CATS on `E-platform` are fraudulent with a high confidence level.

A. Data Collection

The dataset we used in this experiment is collected from the publicly available website of `E-platform` and is then subsampled and anonymized for the privacy commitments and ethical concerns. Moreover, this collected dataset is only used for the e-commerce fraud detection research. Moreover, this collected dataset is only used for the purpose of the e-commerce fraud detection research. Our data collector is built upon Python Scrapy [13] and collects three types of data: the shop data, the item data and the comment data. The following shows how our web collector works.

1) **Shop Data.** Our data collector first fetches all the homepages of third-party shops from `E-platform`. Then, for each scanned homepage of the shop, our data collector extracts its basic information, including *shop id*, *shop url* and *shop name*.
2) **Item Data.** After preparing the basic information for third-party shops, our web collector scrapes all the online items from these shops. Then, it extracts the *item id*, *item name*, *price*, and *sales volume* of the collected items.
3) **Comment Data.** Since a single item may contain more than one comment, our data collector continues to collect all the comments of a single item. Listing 2 gives an

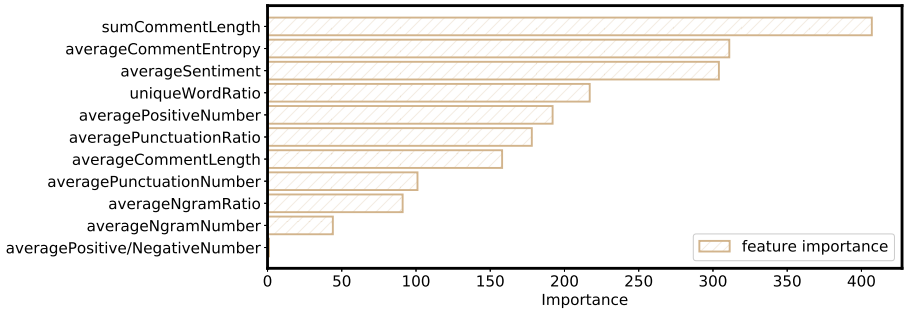


Fig. 7. Feature importance.

TABLE VI
THE PERFORMANCE OF CATS ON D_1.

Category	Precision	Recall	F-score
fraud items labeled with sufficient evidences	0.83	0.92	0.87
the overall fraud items	0.91	0.90	0.90

example of the comment record, which contains many features, including *item id*, *comment id*, *comment content*, *user nickname* and *userExpValue*.

```
{"item id": "545470505476",
"comment id": "40805023517",
"comment content":
"这个商品很好(This item is very,
good) ..."
"nickname": " 0***莉",
"userExpValue": "100",
"client information": "Android",
"date":"2017-09-10 12:10:00",...}
```

Listing 2. An example of comment record.

We deploy our web crawler on three servers to collect data from E-platform. These three servers are equipped with a total of 60 vCPUs and 260 GB memory. To obtain sufficient data, the data collector continuously runs for about one week (from 2017-12-24 to 2017-12-31). In summary, we collect $\sim$ 4.5 million online items and over 100 million comments for these items.

B. Evaluation

To evaluate the performance of CATS on E-platform, we use CATS to detect fraud items leveraging the data collected from E-platform. This experiment is conducted on a server equipped with 40 Intel Xeon E5-2640V4 vCPUs and 96 GB memory. After running CATS, it reports a total of $10,720$ fraud items.

To validate our evaluation results, we employ a methodology that combines manual labeling and statistical analysis. Specifically, we first randomly sample $1,000$ items from the $10,720$ fraud items reported by CATS. These $1,000$ items are then manually examined by anti-fraud experts through multiple aspects, e.g., inspecting their comments, the emotion conveyed by their comments, the contents of their online pages, etc. This validation confirms 960 fraud items with a precision of 0.96. Then, we statistically analyze the fraud items to find their fraud characteristics. Specifically, we compare the characteristics of the reported fraud items of E-platform with the fraud characteristics of the labeled fraud items of Taobao. Our measurement study demonstrates that the reported fraud items on E-platform are truly fraudulent with a high confident level. Below, we show the measurement study in detail.

(a) E-platform's words in English (b) Taobao's words in English

Fig. 8. The word clouds of fraud items on E-platform and Taobao.

V. MEASUREMENT AND ANALYSIS OF E-COMMERCE FRAUDS

Based on the detection results on two different platforms, Taobao and E-platform, we further conduct a comprehensive study to (1) statistically validate CATS' detection results on E-platform and (2) to better understand the common characteristics of e-commerce frauds.

In addition to the comments, we collect various additional features of items, including the basic information, the detailed trade information, and the user information of each comment. Therefore, we statistically analyze the reported frauds from three aspects: the item aspect, the user aspect, and the order aspect.

Item Aspect. We first inspect the word distributions in fraud and normal items' comments. For simplicity, we refer the reported fraud items on Taobao or E-platform as fraud items and the remaining unreported items as normal items.

Fig. 8(a) and Fig. 8(b) show the word clouds of the fraud items on E-platform and Taobao respectively, where the large size of a word represents a high frequency of this word while the small size of a word represents a low frequency of

(a) E-platform's words in English (b) Taobao's words in English

Fig. 9. The word clouds of normal items on E-platform and Taobao.

this word. From Fig. 8(a), we observe that the comments of the fraud items on E-platform are filled up with positive words, e.g., 很好 (very good), 好用 (easy to use), and 实惠 (high performance-price ratio), etc. From Fig. 8(b), we can see that the comments of Taobao's fraud items are filled up with positive words as well. It is obvious that the word distribution of the fraud items on E-platform is almost the same as that of the fraud items on Taobao (see more details in Appendix A). For example, both E-platform's fraud items and Taobao's fraud items have the same highest frequency words, including 不错 (look good), 喜欢 (like), 很好 (very good) and 满意 (satisfaction). Furthermore, we find that on both E-platform and Taobao, the top 50 words with the highest frequency in fraud items' comments are positive words (as shown in Appendix A), which occupy $\sim 28\%$ of a total.

The above analysis suggests that (1) compared with the normal items, the fraud items are filled with more positive words, and their comments are seemingly more deceptive, and (2) the fraud items reported on E-platfom are accurate based on the comparison of the word distribution with Taobao.

Next, we examine the word distributions of normal items' comments. Fig. 9(a) and Fig. 9(b) show the word clouds of normal items on E-platform and Taobao, respectively. From Fig. 9(a) and Fig. 9(b), we can observe that on both two e-commerce platforms, the frequent words in normal items' comments contain several negative words, e.g., 没用 (useless) and 不好 (bad), etc.

In summary, we conclude that (1) normal item comments' are less deceptive than those of fraud items, and (2) the detection results of CATS on E-plartform are highly confident through the comparison with Taobao.

We further investigate the comment sentiments of the fraud and normal items. Fig. 10 illustrates the comment sentiment distributions of E-platform's fraud and normal items, and Taobao's fraud and normal items. As we can see from Fig. 10, the fraud items on E-platform tend to have more positive comments: more than 99.8% of the comments from the reported fraud items are positive. Whereas, compared with these fraud items, the comment sentiments of E-platform's normal items are less positive and tend to be neutral. Moreover, the comment sentiment distributions of E-platform's fraud and normal items generally agree with those of Taobao's fraud and normal items, as illustrated in Fig. 10. The above sentiment distribution analysis shows that the fraud and normal items reported by CATS on E-platform have a high

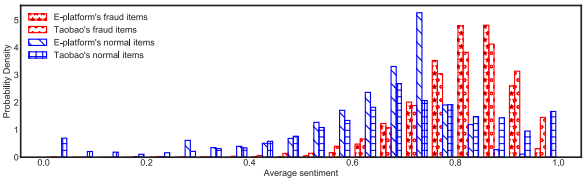


Fig. 10. The comment sentiment distributions of the reported fraud and normal items on E-platform v.s. the labeled fraud and normal items on Taobao.

TABLE VII
EXAMPLES OF USER INFORMATION.

nickname	userExpValue
0***莉	100
0***莓	100
深***壳	3958
三***鱼	55911

accuracy.

User Aspect. Now, we examine the fraud and normal items from the perspective of the e-commerce users. Since the user information of Taobao is unavailable to us, we only analyze the e-commerce users on E-platform.

To study the fraud items from the user perspective, we first need to identify unique users who have ever purchased those fraud items. Table VII shows some examples of e-commerce users collected by our data collector. In Table VII, the *nickname* is the anonymous name of a user, and the *userExpValue* is the user rating score calculated by E-platform based on various factors, including the user credit, the user consumption history, and the user activation. In our research, we employ *userExpValue* and *nickname* to approximately identify unique users.

The value of *userExpValue* can to some extent reveals the reliability of an e-commerce user of E-platform, with the the minimum value 100 and the maximum value $27,158,720$. The lower value of the *userExpValue* of a user, the lower reliability of this user. Table VII also shows several examples of *userExpValue*s, where the user with a nickname "三***鱼" and the user with a nickname "0***莓" are two of the least reliable users among the four listed users. Due to the interpretability of *userExpValue*, we then examine the *userExpValue* of e-commerce users after identifying them.

We dive deep into *userExpValue* to answer two questions: (1) *how is userExpValue distributed among the users*, and (2) *what are the shopping behaviors of those who purchased fraud items.*

First, we measure the *userExpValue* distributions of the users who have bought the fraud and normal items respectively, as shown in Fig. 11. From Fig. 11, we can observe that those fraud items are purchased by a large portion of low reliable users: 45% users have their *userExpValue* below $2,000$, 39% users have their *userExpValue* below $1,000$, and 15% users have the smallest *userExpValue* 100. Compared with the normal items, those who purchased the fraud items tend to have much smaller *userExpValue*s. Moreover, as for the overall users, we find that only $\sim 20\%$ of them have their *userExpValue* smaller than $2,000$.

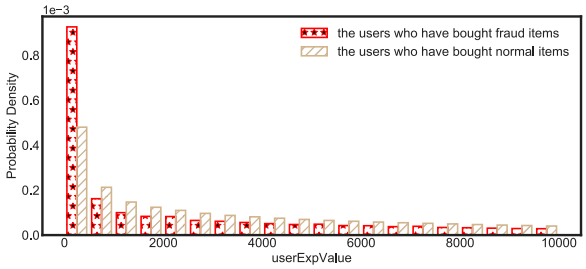
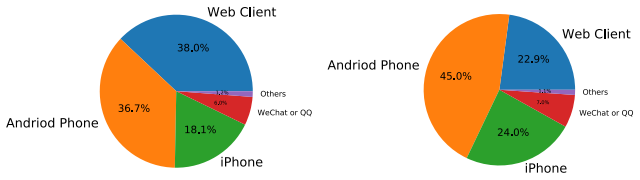


Fig. 11. The distributions of *userExpValue* of the users who have bought fraud and normal items.



(a) Client distribution of fraud items. (b) Client distribution of normal items.

Fig. 12. Client distribution.

Based on the above analysis, we can conclude that, from the perspective of the users, the fraud items are likely not as good as shown by their comments, which are filled up with positive words, since they are purchased and commented by a large number of less reliable users.

Next, for each fraud item, we calculate the average *userExpValue* of the users purchased this item, denoted by *avgUserExpValue*. We find that, 70% of the fraud items have their *avgUserExpValue*s are less than the expectation value of *userExpValue*. This is to say that most of the fraud items are purchased by less reliable users, which are probably hired by malicious merchants to promote their targeted items.

Finally, we examine the shopping behavior of those users who purchased fraud items. For simplicity, we name those users as *risky user*s. We find that, 20% of the risky users have purchased the fraud items for more than once, among which there exist some extreme cases that some risky users have purchased the fraud items for $400+$ times. Then, we analyze the shopping behavior among pairs of users. We find $83,745$ pairs of risky users that have purchased $2+$ same fraud items. After detailed analysis, we find that these $83,745$ pairs of risky users belong to a set of $1,056$. Due to the huge amount of items on `E-platform`, the probability for users to purchase the same item is comparatively small, and it is even smaller to purchase same items for many times. Therefore, we conjecture that these $1,056$ users might be hired by malicious merchants to promote their targeted items.

Order Aspect. As we know, users can use various clients to purchase items on an e-commerce platform, including the iPhone app, the Android app, and the Wechat client. To this end, our data collector also collects the client information from `E-platform`, as shown in Listing 2. For an item on `E-platform`, only the user who has purchased this item can comment on it. Hence, the client information contained in the comment record can be roughly seen as the order source of items. Note that similar to the user aspect, we only examine the orders of `E-platform`'s items for any order information of `Taobao` is unavailable.

Now, we analyze the order source distributions of the fraud and normal items on `E-platform`. Fig. 12(a) shows the client distribution of fraud items' orders, and Fig. 12(b) shows the client distribution of normal items' orders. As we can see from these two figures, the largest portion of fraud items' orders are purchased through the web client while the largest portion of normal items' orders are purchased through the Android client. This client distribution difference is relatively large. We conjecture that it is faster and more convenient for malicious users to promote fraud items through `E-platform`'s web client. In summary, the client distribution demonstrates that, from the perspective of orders, the reported fraud items on `E-platform` tend to be true.

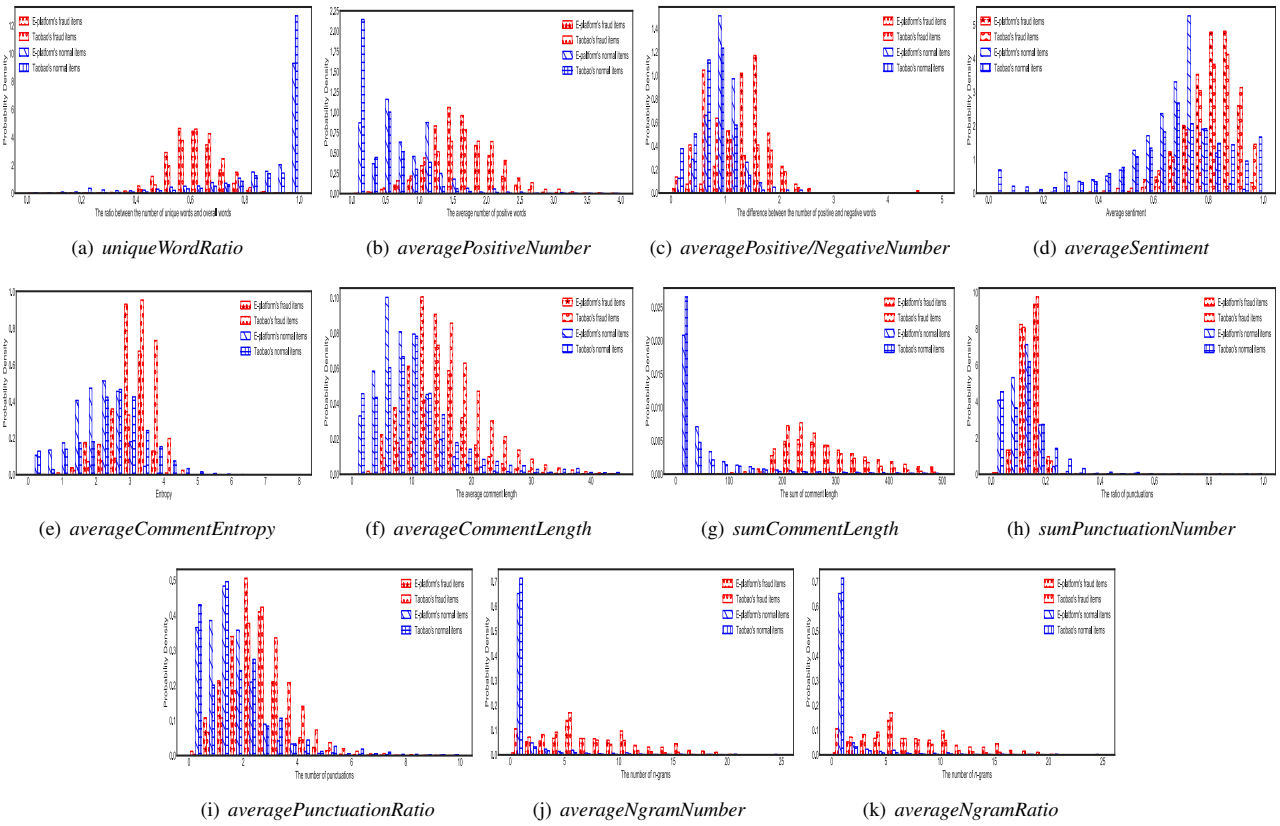
In addition to the analysis from the item aspect, the user aspect and the order aspect, we examine the distribution differences of the extracted features between `E-platform` and `Taobao`. Recall that CATS' detector uses 11 features: *averagePositive/NegativeNumber*, *uniqueWordRatio*, *averageSentiment*, *averageCommentEntropy*, *averageCommentLength*, *sumCommentLength*, *averagePunctuationRatio*, *sumPunctuationNumber*, *averageNgramNumber* and *averageNgramRatio*. Figs. 13(a)–(k) show the feature distributions of the fraud and normal items on `E-platform`, and the fraud and normal items on `Taobao`. From Figs. 13(a)–(k), we observe that (1) the feature distributions of the fraud items reported on `E-platform` roughly agree with those of the fraud items on `Taobao`, and (2) the differences in feature distributions between the reported fraud and normal items reported on `E-platform` are similar to those distribution differences on `Taobao`. Based this observation, we conclude that, the reported fraud items on `E-platform` are highly confident, since they have the same fraud characteristics as the labeled fraud items on `Taobao`.

VI. System Deployment

CATS is an efficient and platform-independent defending system against e-commerce frauds. The design of CATS has been implemented into a prototype system.

We have deeply discussed our ideas and our system with `Alibaba`. `Alibaba` is very interested in our system, and it has partially incorporated CATS into its e-commerce platform, `Taobao`, to detect fraud items in eight categories: *men's clothing*, *women's clothing*, *men's shoes*, *women's shoes*, *computer & office*, *phone & accessories*, *food & grocery* and *sports & outdoors*. On `Taobao`, CATS detects fraud items with a high accuracy from millions of e-commerce items belonging to third-party shops.

For `E-platform`, we evaluate the performance of CATS through collecting the public e-commerce data of `E-platform`. In addition to `Taobao`, we also expect the success of CATS in helping `E-platform` and other platforms migrate the threat posed by e-commerce frauds, and building a more healthy e-commerce environment.



(a) *uniqueWordRatio* (b) *averagePositiveNumber* (c) *averagePositive/NegativeNumber* (d) *averageSentiment*

(e) *averageCommentEntropy* (f) *averageCommentLength* (g) *sumCommentLength* (h) *sumPunctuationNumber*

(i) *averagePunctuationRatio* (j) *averageNgramNumber* (k) *averageNgramRatio*

Fig. 13. Comparison of the feature distributions.

VII. Discussion

Ethical Issues. In our experiments, we employ the data from two real e-commerce platforms, `Taobao` and `E-platform`, which might contain sensitive information from e-commerce users. For the experimental evaluation on `Taobao`, all the used data records have been anonymized, and all the experiments are supervised by `Taobao`'s experts. For the experimental evaluation on `E-platform`, all the data used in our research is collected from the public domain of `E-platform`'s website. Furthermore, considering the ethical issues, our data collector was designed to minimize server impact, and all the experiments were done with the full attention to protect both `E-platform` users' privacy and any information that may involve possible trade secrets. For the measurement analysis of the reported frauds, we only study and report the statistical results.

Further Thoughts. In addition to the detection effort made by CATS, more can be done to migrate the threats posed by e-commerce frauds, from both of the e-commerce service providers and the cyber police. (1) The e-commerce service providers could work harder to perform fraud detection to remove the fraudulent transactions made on their platforms, making sure that the information present to users is reliable. Most importantly, the e-commerce service providers do have the responsibility to move more aggressively on detecting, warning and removal of e-commerce frauds from their systems. This, however, is non-trivial, given the privacy concern and the fact that some e-commerce frauds can only be considered to be malicious by looking at the malicious activities the merchants involved in, such as hiring a group of malicious users. (2) The cyber police could scan and monitor the cyber network to detect malicious websites that provide malicious e-commerce promotion services. Further research is needed to combine the efforts made by the e-commerce service providers and the cyber police to address this issue.

Limitations and Future Work. Our evaluation mainly leverages the data from `Taobao` and `E-platform`. Though the users of `Taobao` and `E-platform` spread all over the world, the majority of these users are Chinese speakers. Moreover, as reported in [2], [3], [4], frauds commonly exist across many large-scale e-commerce platforms worldwide, including `Amazon`, `Taobao`, `eBay`, etc. Therefore, one of the feature research directions is to extend CATS and apply it to other e-commerce platforms, e.g., `Amazon` and `eBay`.

CATS' detector chooses the Xgboost model as the classifier and pre-trains the Xgboost model on a labeled dataset using 11 features. CATS might be more powerful if we design a better classifier dedicated to e-commerce fraud detection, and/or identify more features that can discriminate whether an item is fraudulent or normal. Hence, another feature research direction is to identify more useful features and optimize CATS' detector.

Finally, our findings in this paper also reveal that there might exist a dedicated underground economy targeted e-commerce. Hence, in the future, CATS is expected to work together with e-commerce platforms, to mine and understand the underground ecosystem of e-commerce frauds, including reporting fraud e-commerce items, monitoring malicious promotion platforms, and mining the underground forums of e-commerce frauds.

VIII. Related Work

Fraud detection is a topic applicable to many industries including finance, content service, cloud service, and online service.

In finance, many researchers have devoted themselves to detecting the insurance frauds and tax frauds [6], [16], [17], [18]. In content service, frauds can be malicious advertisements injected into online systems by adversaries [19], [20]. For Wikidata vandalism detection as an example, Heindorf *et al.* proposed a set of features that utilize both content and context information and then trained a classifier to defend against Wikidata vandalism [6]. In cloud service, frauds can be bad repositories that provide malicious online services. For this type of frauds, Liao *et al.* developed a prototype system, *BarFinder*, to automatically detect bad repositories through analyzing the topological features of online repositories. In online advertising, frauds can be fake clicks automatically generated by advertisers. A large number of researchers have committed to detect this kind of frauds [21], [22], [23], [24], [25], [26]. In the online service area, frauds can be fake accounts manipulated for reputation-enhancement services [7], [8], [27], [28], [29], [30]. For example, Wang *et al.* proposed to use the clickstream model to detect fake accounts in social networks [7].

In e-commerce, many fraud detection works have focused on credit card frauds in online transactions [31], [32], [33], [34], [35], [36]. For example, Santiago *et al.* proposed a comprehensive approach to address the problem of fraud detection in the emerging market of online payment services [35]. Raj *et al.* presented a survey of fraud detection techniques of the credit card and evaluated them according to their design criterion [36].

Some work has focused on fraudulent transaction detection [37], [4]. For example, Zhao *et al.* designed a novel detection framework that can reason about implicit online user behaviors for detecting collusive fraudulent transactions [37]. Most recently, Wang *et al.* presented a novel deep learning based transaction fraud detection system and designed and deployed this system on the `Jindong` platform [4].

Remark. Different from most of the existing fraud detection techniques, CATS' application domain is for the online e-commerce platforms, and it aims to detect fraud items only by analyzing the publicly available e-commerce data. More importantly, CATS is designed as a cross-platform fraud detection system, and therefore, it can be used to detect frauds on various platforms. In our work, we have applied CATS on two of the world's popular online e-commerce platforms.

IX. Conclusion

In this paper, we study the fraud detection problem in e-commerce. First, we present and implement an efficient, platform-independent, and robust e-commerce fraud detection system, CATS. Second, we evaluate the performance of CATS on `Taobao`, which is one of the world's popular e-commerce platforms. The evaluation results indicate that CATS achieves both high precision and recall. Then, we employ CATS to detect fraud items on another e-commerce platform, named `E-platform`, which is also a popular B2C online retailer worldwide. Through manual labeling and statistical analysis, we demonstrate that the fraud items detected on `E-platform` are fraudulent with a high confidence level. Finally, we make a discussion on the limitations and possible future research directions of this work. Our study in this paper is expected to be helpful for defending against frauds for various e-commerce platforms.

X. Acknowledgement

This work was partly supported by NSFC under No. 61772466, the Zhejiang Provincial Natural Science Foundation for Distinguished Young Scholars under No. LR19F020003, the Provincial Key Research and Development Program of Zhejiang, China under No. 2017C01055, and the Alibaba-ZJU Joint Research Institute of Frontier Technologies.

References

[1] C.-H. Park and Y.-G. Kim, "Identifying key factors affecting consumer purchase behavior in an online shopping context," *International journal of retail & distribution management*, 2003.

[2] W. Shepard, https://www.forbes.com/sites/wadeshepard/2017/01/02/amazon-scams-on-the-rise-in-2017-as-fraudulent-sellers-run-amok-and-profit-big/.

[3] J. Lim, https://www.forbes.com/sites/jlim/2015/04/11/jd-com-brushing-fake-orders-to-inflate-sales.

[4] S. Wang, C. Liu, X. Gao, H. Qu, and W. Xu, "Session-based fraud detection in online e-commerce transactions using recurrent neural networks," in *Proceedings of European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD)*, 2017.

[5] X. Li, M. Zhang, Y. Liu, S. Ma, Y. Jin, and L. Ru, "Search engine click spam detection based on bipartite graph propagation," in *International Conference on Web Search and Data Mining (WSDM)*, 2014.

[6] S. Heindorf, M. Potthast, B. Stein, and G. Engels, "Vandalism detection in wikidata," in *Proceedings of the 25th ACM International Conference on Information and Knowledge Management (CIKM)*, 2016.

[7] G. Wang, T. Konolige, C. Wilson, X. Wang, H. Zheng, and B. Y. Zhao, "You are how you click: Clickstream analysis for sybil detection." in *USENIX Security Symposium*, 2013.

[8] G. Stringhini, G. Wang, M. Egele, C. Kruegel, G. Vigna, H. Zheng, and B. Y. Zhao, "Follow the green: growth and dynamics in twitter follower markets," in *Proceedings of the 2013 conference on Internet measurement conference*, 2013.

[9] *CATS System*, https://gitlab.com/HaiQW/cats.

[10] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *CoRR*, 2013.

[11] *SnowNLP: Simplified Chinese Text Processing*. [Online]. Available: https://github.com/isnowfy/snownlp

[12] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2016.

[13] *Scrapy: A Fast and Powerful Scraping and Web Crawling Framework*, https://github.com/scrapy/scrapy.

[14] *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems*, 2015, http://tensorflow.org/.

[15] H. Weng, Z. Li, S. Ji, C. Chu, H. Lu, T. Du, and Q. He, "Online e-commerce fraud: A large-scale detection and analysis," in *ICDE*, 2018.
[16] P. Picard, "Economic analysis of insurance fraud," 2000.
[17] D. Yue, X. Wu, Y. Wang, Y. Li, and C.-H. Chu, "A review of data mining-based financial fraud detection research," in *International Conference on Wireless Communications, Networking and Mobile Computing (WiCom)*, 2007.
[18] A. Abbasi, C. Albrecht, A. Vance, and J. Hansen, "Metafraud: a meta-learning framework for detecting financial fraud," *Mis Quarterly*, 2012.
[19] X. Liao, K. Yuan, X. Wang, Z. Pei, H. Yang, J. Chen, H. Duan, K. Du, E. Alowaisheq, S. Alrwais *et al.*, "Seeking nonsense, looking for trouble: Efficient promotional-infection detection through semantic inconsistency search," in *Security and Privacy (SP), 2016 IEEE Symposium on*, 2016.
[20] M. Marciel, R. Cuevas, A. Banchs, R. González, S. Traverso, M. Ahmed, and A. Azcorra, "Understanding the detection of view fraud in video content portals," in *Proceedings of the 25th International Conference on World Wide Web (WWW)*, 2016.
[21] S. Mittal, R. Gupta, M. K. Mohania, S. K. Gupta, M. Iwaihara, and T. S. Dillon, "Detecting frauds in online advertising systems," in *International Conference on E-Commerce and Web Technologies (EC-Web)*, 2006.
[22] C. Kim, H. Miao, and K. Shim, "CATCH: A detecting algorithm for coalition attacks of hit inflation in internet advertising," *Inf. Syst.*, 2011.
[23] V. Dave, S. Guha, and Y. Zhang, "Measuring and fingerprinting click-spam in ad networks," in *ACM SIGCOMM*, 2012.
[24] R. J. Oentaryo, E. Lim, M. Finegold, D. Lo, F. Zhu, C. Phua, E. Cheu, G. Yap, K. Sim, M. N. Nguyen, K. S. Perera, B. Neupane, M. A. Faisal, Z. Aung, W. L. Woon, W. Chen, D. Patel, and D. Berrar, "Detecting click fraud in online advertising: a data mining approach," *Journal of Machine Learning Research*, 2014.
[25] L. Zhang and Y. Guan, "Detecting click fraud in pay-per-click streams of online advertising networks," in *IEEE International Conference on Distributed Computing Systems (ICDCS)*, 2008.
[26] A. Metwally, D. Agrawal, and A. El Abbadi, "Detectives: detecting coalition hit inflation attacks in advertising networks streams," in *Proceedings of the 16th international conference on World Wide Web*. ACM, 2007.
[27] E. De Cristofaro, A. Friedman, G. Jourjon, M. A. Kaafar, and M. Z. Shafiq, "Paying for likes?: Understanding facebook like fraud using honeypots," in *Proceedings of the 2014 Conference on Internet Measurement Conference*, 2014.
[28] Y. Zhao, Y. Xie, F. Yu, Q. Ke, Y. Yu, Y. Chen, and E. Gillum, "Botgraph: Large scale spamming botnet detection." in *NSDI*, 2009.
[29] G. Stringhini, P. Mourlanne, G. Jacob, M. Egele, C. Kruegel, and G. Vigna, "Evilcohort: detecting communities of malicious accounts on online services." USENIX, 2015.
[30] Y. Li, O. Martinez, X. Chen, Y. Li, and J. E. Hopcroft, "In a world that counts: Clustering and detecting fake social engagement at scale," in *Proceedings of the 25th International Conference on World Wide Web*, 2016.
[31] T. P. Bhatla, V. Prabhu, and A. Dua, "Understanding credit card frauds," *Cards business review*, 2003.
[32] N. Carneiro, G. Figueira, and M. Costa, "A data mining based system for credit-card fraud detection in e-tail," *Decision Support Systems*, 2017.
[33] A. D. Pozzolo, O. Caelen, Y. L. Borgne, S. Waterschoot, and G. Bontempi, "Learned lessons in credit card fraud detection from a practitioner perspective," *Expert Syst. Appl.*, 2014.
[34] E. W. T. Ngai, Y. Hu, Y. H. Wong, Y. Chen, and X. Sun, "The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature," *Decision Support Systems*, 2011.
[35] G. P. Santiago, A. C. M. Pereira, and R. H. Jr., "A modeling approach for credit card fraud detection in electronic payment services," in *Proceedings of the 30th Annual ACM Symposium on Applied Computing (SIGCHI)*, 2015.
[36] S. B. E. Raj and A. A. Portia, "Analysis on credit card fraud detection methods," in *International Conference on Computer, Communication and Electrical Technology (ICCCET)*, 2011.
[37] J. Zhao, R. Y. K. Lau, W. Zhang, K. Zhang, X. Chen, and D. Tang, "Extracting and reasoning about implicit behavioral evidences for detecting fraudulent online transactions in e-commerce," *Decision Support Systems*, 2016.

APPENDIX

Table VIII shows the details of the top 50 words with the highest frequency in the comments for the reported fraud items on `E-platform`. Table IX shows the details of the top 50 words with the highest frequency in the comments for the labeled fraud items on `Taobao`. From Table. VIII and Table IX, we observe that (1) the top 50 frequent words of `E-platform`'s reported fraud items are filled up with positive words, and (2) are very similar to those of `Taobao`'s fraud items.

TABLE VIII
TOP 50 WORDS OF E-PLATFORM'S FRAUD ITEMS.

Top 50 Words
不错(well), 很好(very good), 质量(quality), 喜欢(like) , 满意(satisfy), 收到(received), 好好(good), 好看(good look), 物流(material flow), 东西(goods), 包装(packaging), 宝贝(goods), 很快(fast), 好评(good reputation), 舒服(comfort) 挺好(good), 卖家(seller) , 值得(deserve), 挺(very) , 很漂亮(beautiful), 价格(price), 发货(shipments), 购买(purchase), 合适(suitable), 款式(style), 非常好(good), 颜色(color), 漂亮(beautiful), 下次(next time), 买的(buy), 也很(very), 还不错(good), 感觉(feel), 效果(effect), 喜欢(like), 衣服(cloth), 实惠(high performance-price ratio) , 特别(specially), 穿着(dress), 精致(delicacy), 大小(size), 还会(still) 速度(speed), 赞(like), 还可以(all right), 购物(shopping), 推荐(recommend), 服务(service), 正品(qulitified goods) 做工(workmanship)

TABLE IX
TOP 50 WORDS OF TAOBAO'S FRAUD ITEMS.

Top 50 Words
不错(well), 买(purchase), 质量(quality), 收到(received), 喜欢(like), 舒服(comfort), 挺(very), 宝贝(goods), 物流(material flow), 很快(fast), 好评(good reputation), 穿(dress), 包装(packaging), 好看(good look), 做工(workmanship), 卖家(seller), 效果(effect), 穿着(dress), 发货(shipments), 东西(goods), 价格(price), 款式(style), 客服(customer service), 特别(specially), 值得(deserve), 购买(purchase), 合适(suitable), 感觉(feel), 真的(true), 服务(service), 颜色(color), 推荐(recommend), 速度(speed), 服务态度(service attitude), 安装(install), 正品(qualitified goods), 朋友(friends), 实惠(high performance-price ratio), 很漂亮(very beautiful), 高(high), 购物(shopping), 性价比(cost performance), 店家(merchant), 大小(size), 漂亮(beautiful), 还会(still), 态度(attitude), 精细(delicacy), 便宜(cheap), 舒适(comfort)